Pakistan Journal of Social Sciences 13 (3): 25-31, 2016
ISSN: 1683-8831

Assessment of Outlier Detection Procedures in Analysis of Regression Model

Azeez Adeboye, Ndege James and Odeyemi Akinwumi
Department of Biostatistics and Epidemiology, University of Fort Hare,
Alice, South Africa, PMB X1314, 5700 Alice, South Africa

Abstract: Five detection of outliers procedures in Multiple regression model are looked into, compared and investigated with a simulated data. The researchers reviewed five outlier detection methods in multiple linear regression model and then compares theirs results by using two criteria of robust diagnostics called the Median Absolute Deviation (MAD) and the Standard Deviation (SD) parameter estimate. Data were generated with 10, 20 and 30% of outliers on X_1's, X_2's and both X_1's and X_2's, respectively with different sample sizes (20, 50 and 100) to check and compare outliers in the residual space of CovRatio which will flag observations that are influential because of large residual, outliers in the X-space of Hat Diagonal which flags observations that is influential because they are outliers in the X-space, the Dffits shows the influence on fitted values and measures the impact on the regression coefficients. Cook's D measures the overall impact that a single observation has on the regression coefficient estimates and Mahalanobis Distance measures the hat leverage through the means of Md_i.

Key words: Coefficient, diagonal, measures, P-P plot, residual, simulation, unbiased estimator

INTRODUCTION

Regression is a technique that is used for functional relationships analysis among variables. Regression analysis is a statistical tool that utilizes the relationship between two or more quantitative variables. Relationship that exists between the variables can be functional or statistical: a functional relationship between mean y_i denoted by E(yi) and x_i is the equation of a straight line $E(y_i) = \beta_0+\beta_1 x_i$ while a statistical relationship is the deviation of an observation of y_i from population mean by adding a random error to a given statistical model $y_i = \beta_0+\beta_1 x_i+\varepsilon_i$ where $\varepsilon_i \sim N(0, \sigma^2)$, $E(\varepsilon_i) = 0$, var $(\varepsilon_i) = \sigma^2$. Linear regression model involves any one independent variable and state that true mean of dependent variable changes at a constant rate as the value of the independent variable increases or decreases. In matrix term, the regression model becomes $\underset{\sim}{Y} = \beta\underset{\sim}{X} + \underset{\sim}{\varepsilon}$:

$$\begin{bmatrix} y_1 \\ y_2 \\ : \\ y_n \end{bmatrix}_{(nx1)} = \begin{bmatrix} 1\, x_{11}\, x_{21} \ldots\ldots x_{p1} \\ 1\, x_{12}\, x_{22} \ldots\ldots x_{p2} \\ : \; : \quad : \; \ldots\ldots : \\ 1\, x_{1n}\, x_{2n} \ldots\ldots x_{pn} \end{bmatrix}_{n\times(p+1)} \begin{bmatrix} \beta_0 \\ \beta_1 \\ : \\ \beta_p \end{bmatrix}_{(p+1)\times 1} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ : \\ \varepsilon_n \end{bmatrix}_{(n\times 1)}$$

$E(\varepsilon_i) = 0$, var $(\varepsilon_i) = \sigma^2$ implies that the errors have zero mean constant variance and are independent with an assumption of Gauss-Markov theorem is $\varepsilon_i \sim N(0, \sigma^2)$ (Amphanthong, 2012). The two most used approaches to estimate Best Linear Unbiased Estimator (BLUE) of $\underset{\sim}{\beta} = X^1Y\left(X^1X\right)^{-1}$ are the ordinary least squares method which minimizes the sum of squares and Maximum likelihood estimator (Adnan *et al*., 2003; Ampanthong and Prachoom, 2009). However, when the error terms are normally and independently distributed with mean = 0 and without a constant variance (homogeneity of variance or homoscedasticity), the regression model will not be adequately fit the model and the parameter estimates and inferences will be arbitrarily bad. This shows that there is a presence of one or more outliers that could influence the appropriateness of the fitted regression model. At times, the data might contains outlying observations in independent variables (X's values), dependent variables (Y's values) or both Y's and X's values (Ampanthong and Prachoom, 2009; Oyeyemi *et al*., 2015).

In statistics, an outlier is a point that lies far away from the fitted line and thus produces a large residual. Many authors have studied and analyzed regression model diagnostics of outliers in multiple data. Oyeyemi *et al*. (2015) and Gurunlu *et al*., 2011) explained an outlier as a point which deviates from other member of the sample data in which it occurs. They are observations that are not consistence with the rest of the data. Such outliers can a have damageable influences on

Corresponding Author: Azeez Adeboye, Department of Biostatistics and Epidemiology, University of Fort Hare, Alice, South Africa, PMB X1314, 5700 Alice, South Africa

the result of the analysis (Adnan *et al*., 2003; Alma, 2011; Filzmoser, 2004; Kumar and Nasser, 2012; Newton *et al*., 2010; Pereira and Pires, 2002).

Although, outlying observations are usually considered as an error or noise which may have important information (Oyeyemi *et al*., 2015). Outlier can be seen as a value which falls more or less than one and a half times the interquartile range, i.e., is a point below Q_1-1.5 (IQR) or above Q_3-1.5 (IQR) which is seen as been too far from the middle values. Outliers can make regression model fails to capture one or more important characteristics of the dataset and have a strong influence on the model but deleting these outliers from the regression model can give a different result from the dataset. A regression model with outliers will have a large residual and can be an observation with an unusual value of the response variable Y, conditionally on its value of the independent variable X but neither the variable X nor Y is necessary unusual on its own. Outliers in the response variable are known as 'model failure' while in the predictors are called 'leverage point' and can affect regression model. These error can easily be seen on Box and Whisker plots and as well as histogram.

There are numerous outliers' detection methods explained by different authors. Most of these authors try to establish algorithms to detect outliers that are based on distance. Some of these approaches to detect outliers in regression model are clustering-based methods, distance based methods, density based method, subspace based methods and statistical approaches (Alma, 2011) and robust regression approaches such as Least Median of Squares (LMS), Least Trimmed Squares (LTS), Least Absolute Value Method (LAVM) of robust method (Alma, 2011). Detection approaches have been designed to find solution to the problem that is not easily affected by the influence of the outliers. It is absolutely important to understand both the analytical and computational point of view of these approaches (Adnan *et al*., 2003).

In this study, five approaches of outlier detection in regression model are considered and the results are used to know which of them perform best.

MATERIALS AND METHODS

Analytical methodology: In this study, there are numbers of methods use in detecting outlying observations in multiple regression model. These are categorized as graphical method and analytical method. Graphical methods include normal P-P plot, scatter plot, residual plot and box and whisker plot. Analytical method is flagged into three types of influence on the regression. Firstly, outliers in the residual space: they are studentized residual, RStudent and CovRatio which will flag observations that are influential because of large residual. Secondly, outliers in the X-space such as Hat Diagonal which flags observations that is influential because they are outliers in the X-space. Lastly, the authors reviewed five outlier detection methods in multiple linear regression model and then compares theirs results by using two criteria of robust diagnostics called the Median Absolute Deviation (MAD) and the Standard Deviation (SD) parameter estimate and fit: the Dffits shows the influence on fitted values and measures the impact on the regression coefficients. Cook's D measures the overall impact that a single observation has on the regression coefficient estimates. However, the five analytical methods of detecting outliers that were discussed in this study are: Hat Diagonal, Cook's D, Dffits, CovRatio and Mahalanobis Distance. The computational procedures are as follows.

The Hat Diagonal (h_{jj}): This is also called hat matrix leverage, it is a measure of leverage in the X-space. h_{jj} captures observation's remoteness in the X-space that is it measures the distance between the x-space of the jth observation and the mean of the X-spaces for all n cases. h_{jj} greater than 2 times the number of coefficients in the model divided by the number of observations are said to have high leverage (i.e., $h_{jj}>2p/n$) which indicate that the jth case is distance from the center of X-space. It can be compute as:

$$h_{jj} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$h_{jj} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}, 0 \le h_{jj} \le 1, \sum h_{jj} = p + 1$$

$$\text{where } S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

for n = 20, p = 2; /hat/>0.2000, for n = 50, p = 2; /hat/>0.0800, for n = 100, p = 2; /hat/>0.0400.

The Cook's distance: The Cook's D (Cook, 1977), measures the influence on each observation on all n-fitted values. It measures the distance between estimates of the regression coefficients with ith observation $\hat{\beta}$ and without $\hat{\beta}_{-i}$ for a metric $\frac{1}{p\hat{\sigma}^2}(X^TX)$ case on all fitted values:

$$D_i = \frac{\left(\hat{\beta} - \hat{\beta}_{-i}\right)^T \left(X^TX\right)\left(\hat{\beta} - \hat{\beta}_{-i}\right)}{\hat{\sigma}^2 p}$$

The Cook's is compared to a central F distribution with p and n-p degree of freedom. This gives however

exaggeratedly high cutoff values. Practically a cutoff value of 4/n-p seems more reasonable. Cook's distance can be expressed as follows in terms of the studentized residual and the classical leverage indicator:

$$D_i = \frac{r_i^2}{k}\frac{h_{ii}}{1-h_{ii}}$$

where, r_i^2 is ith internally studentized residual and h_{ij} the ith diagonal element of the hat matrix. High D_i requires large values of both r_i^2 and h_{ij}. Thus, like other classical diagnostic measures, it becomes unreliable in the case of multiple data. It is considered that an observation is an influential observation when D_i exceeds the cut-off point of 4/n-p (Cook, 1977; Wichn and Wahba, 1998). The interpretation is identical to the single-row version, therefore, it is assumed that the subset with $D_i > F_{0.5, p, n-p} \approx 1$ is an outlier.

DFFITS (the welsch-kuh distance): DFFITS is the standardized difference between the predicted value with and without that observation. It represents the number of estimated standard errors that the fitted value changes if the ith observation is omitted from the dataset. The impact of the ith observation on the ith predicted value can be measured by scaling the change in prediction at x_i when the ith observation is omitted. This can be mathematical generated as:

$$DFFITS_i = \frac{\left|\hat{y}_i - \hat{y}_{i,-i}\right|}{\hat{\sigma}_{-i}\sqrt{h_{ii}}} = \frac{\left|x_i^T\left(\hat{\beta}-\hat{\beta}_{-i}\right)\right|}{\hat{\sigma}_{-i}\sqrt{h_{ii}}} = \left|r_i^*\right|\sqrt{\frac{h_{ii}}{1-h_{ii}}}$$

Where:

r_i^* = The ith externally studentized residual

h_{ij} = The h_{ij} diagonal elements of the matrix

Belsley *et al.* (1980), Filzmoser (2004) recommend using $/DFFITS/ > 2\sqrt{p/n}$ as a cut-off point for DFFITS (Filzmoser, 2004; Hadi, 1992). For n = 20, p = 2; /dffits/>0.6325, for n = 50, p = 2; /dffits/>0.4000, for n = 100, p = 2; /dffits/>0.2828

CovRatio: CovRatio predicts the observations that have a major impact on the generalized variance of the regression coefficients. A value >1, implies that ith observation provides a reduction in the generalized variance of the coefficients and a value of CovRatio below 1, shows an observation that increase the estimated generalized variance of the coefficients. The general form of CovRatio is:

$$\text{CovRatio}_j = \frac{\det\left[s_{(j)}^2\left(X_{(j)}^T X_{(j)}\right)^{-1}\right]}{\det\left[s^2\left(X^T X\right)^{-1}\right]}$$

$$\text{CovRatio}_j = \frac{1}{1-h_{jj}}\left[\frac{s_{(j)}^2}{s^2}\right]^p$$

The itt point of observation is considered influential if $\text{CovRatio}_i > 1+3p/n$, then omitting the observation improves the precision of the regression estimates and if $\text{CovRatio}_i < 1+3p/n$, omitting the observation reduces the precision of at least some of the estimates. where j denotes the quantity computed without the group observations in question, p is the number of parameters in the model including the intercept, s^2 is the estimated mean squared error and β is the vector of least squares estimates of the parameters. Belsley *et al.* (1980), Cook (1977), Albert and Chib (1995) and Welsch (1980) demonstrate that the multiple-case diagnostics successfully assess the joint influences exerted by the outliers. For n = 20, p = 2; /covratio/>0.3500, for n = 50, p = 2; /covratio/>0.1400, for n = 100, p = 2; /covratio/ >0.0700.

Mehalanobis Distance (MD): Mehalanobis Distance measures the hat leverage through the means of Md_i and it can be computed as:

$$Md_i^2 = \left(x-\bar{x}\right)\Sigma^{-1}\left(x-\bar{x}\right)^T$$

$$= \left(n-1\right)\left[h_{jj}-\frac{1}{n}\right]$$

where i = 1, 2, 3, ..., n;

Σ^{-1} = inverse of covariance matrix; $\bar{x} = \frac{\sum x_i}{n}$

With the following properties:

- Its variances in each direction are differ
- It shows covariance among the variables
- The Euclidean distance is reduced for uncorrelated variables with one variance

Md_i can be determined to be too large if the distance is compared with 95th percentile of Chi-square distribution with p-1 degree of freedom. 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, then it is an outlier.

Robust diagnostic criteria: Many statistical values can be computed from the data sets that are used to identify

the presence of outliers. Most used robust diagnostics statistical criterion of the Standard Deviation (SD) of the residuals, e_1, e_2, ..., e_n. The measure which is based on the Mean Squared Error (MSE) is not robust, since this may be highly influenced by events of small probability. For this study, researchers use the Median Absolute Deviation (MAD) (Filzmoser, 2004; Reading, 2011) and the Robustness Standard Deviation (RSD) (Rasheed *et al*., 2014) and are computed as: mean absolute deviation:

$$MAD_{(e_i)} = \frac{med|e_i - med(e_i)|}{0.6745}$$

Robust standard deviation:

$$RSD_{(e_i)} = 2.1med\{|e_i|\}$$

Analysis: In this study, a simulation study is conducted to compare outlier detection procedures in analysis of regression model and all the data used were simulated. Simulation is used because most times it hard to get secondary data that contain the required number of outliers. Simulation technique is a dependable tool in situations where mathematical or statistical analysis is either too complex or costly (Oyeyemi *et al*., 2015). It is a very useful tool which allows experimentation without exposure to risk. It is an abstraction of reality which specifies application of models to arrive at some outcomes (Hadi, 1992).

However, replicas of dataset were generated from regression model with independent variables: $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_2 + \beta_3 x_3 + \varepsilon_i$, $i = 1, 2, 3, ..., n, i$ indicates the particular unit of observation. x_i are the n observation on the independent variable and are assumed to be measured without error. The random error-ε_i has zero mean and is assumed to have common variance and independent which are assumed to be normally and independently distributed, i.e., $\varepsilon_i \sim NID(0, \sigma^2)$. In comparing the five outlier detection, the following steps were carried out in simulating the required dataset for the analysis:

- Data were generated with 10, 20 and 30% of outliers on X_1's, X_2's and both X 's and X 's, respectively with different sample sizes (20, 50 and 100)
- Explanatory variables X_1, X_2 and X_3 were simulated from normal distribution N(2, 5) using R statistical software
- The simulations were replicated 1000 times
- The simulated data were analyzed with the outlier detection procedures and the number of times the outliers were detected
- Each procedure was computed for each sample size with the replicates and the probabilities of time of outliers were identified
- Comparisons of outlier detection were made by identifying the probabilities of each of the procedure that is above below and accurate
- The best method (s) were recommended for the three sample sizes, provided the comparison variation indicates sensitivity of procedure used

RESULTS AND DISCUSSION

The result of outlier procedures: The computations of r-codes for the detection of outliers give the correct probabilities of each of the procedures for three different sample sizes and percentages on each of the sample sizes of outliers from 1000 replications. The probabilities of the results of statistics of five outlier detection procedures are as follows shown in Table 1.

The result from Table 1 indicates that the best outlier detection is CovRatio in each of the sample sizes and for all the percentages of outliers except in 20% of outlier follow by the Dffits, hat leverage and Mahalanobis distance. It is also seen that the Cooks performed better in 30% of sample 20 and 10% of sample 50, respectively but does not indicate best in the rest of the sample sizes.

The result from Table 2 indicates that the best outlier detection is CovRatio in each of the sample sizes and

Table 1: Probability of comparisons of outcome values of outlier detection procedures by sample size and percentages of X_1's outliers

		X_1				
Sample size (n)	n-outlier (%)	Hat	Cooks	Dffits	Covratio	Md
20	10	0.4713	0.0000	0.5263	1.0000	0.4314
	20	0.4504	0.0000	0.6489	0.9899	0.4312
	30	0.5287	0.7062	0.4301	1.0000	0.4756
50	10	0.4642	0.6393	0.2674	1.0000	0.4513
	20	0.3989	0.0000	0.4692	1.0000	0.5802
	30	0.4549	0.0000	0.3418	1.0000	0.3333
100	10	0.4100	0.0000	0.2452	1.0000	0.3284
	20	0.4290	0.0000	0.2876	1.0000	0.4339
	30	0.3700	0.0000	0.3525	1.0000	0.4692

Table 2: Probability of comparisons of outcome values of outlier detection procedures by sample size and percentages of X_2's outliers

		X_2				
Sample size (n)	n-outlier (%)	Hat	Cooks	Dffits	Covratio	Md
20	10	0.3550	0.0000	0.3582	1.0000	0.4725
	20	0.5364	0.0000	0.4953	1.0000	0.4404
	30	0.4145	0.0000	0.4239	1.0000	0.3243
50	10	0.3885	0.0000	0.3001	1.0000	0.3975
	20	0.5063	0.0000	0.3625	1.0000	0.4082
	30	0.4607	0.0000	0.2365	1.0000	0.5050
100	10	0.3983	0.0000	0.2803	1.0000	0.5108
	20	0.4402	0.0000	0.2916	1.0000	0.4327
	30	0.5084	0.0000	0.3616	1.0000	0.5187

for all the percentages of outliers follow by the hat leverage, Mahalanobis distance, Dffits. With small sizes, Mahalanobis distance performed better than other procedures, and its value of the outlier detection in low percentage of X's is 0.4725. With large sample sizes of performance, Md still performs better than the rest procedures with large percentages of X's.

The result from Table 3 indicates that the best outlier detection is CovRatio in each of the sample sizes and for all the percentages of outliers follow by the hat leverage, Mahalanobis distance, Dffits. With small sizes, Mahalanobis distance performed better than other procedures and its value of the outlier detection in low percentage of X's is 0.6914. With large sample sizes of performance, the hat leverage performs better than the rest of the procedure follows by Md's and the Cooks distance.

The result from Table 4 shows that the best criterion is Median Absolute Deviation (MAD) and the best of X_1's outlier detection approaches are covratio and Md. Their performances are highest outcome values of outlier detection for all sample sizes and percentage of outliers. Furthermore, the performance of Cooks and dffits are high for large sample size.

Table 5 shows the probability of outcome values of outlier detection procedures by robust diagnostic criteria

Table 3: Probability of comparisons of values of outlier detection procedures by sample sizes and percentages of X_1 and X_2 outliers

		X_1 and X_2				
Sample size (n)	n-outlier (%)	Hat	Cooks	Dffits	Covratio	Md
20	10	0.5433	0.8559	0.5876	1.0000	0.6914
	20	0.4249	0.5391	0.3381	1.0000	0.5809
	30	0.3914	0.0000	0.4161	1.0000	0.4681
50	10	0.4807	0.0000	0.2514	1.0000	0.3215
	20	0.4271	0.8644	0.5821	1.0000	0.3737
	30	0.4590	0.0000	0.3387	1.0000	0.4433
100	10	0.4156	0.0000	0.2393	1.0000	0.4142
	20	0.4726	0.0000	0.2847	1.0000	0.4613
	30	0.4634	0.0000	0.3543	1.0000	0.4536

Table 4: Comparisons of parameter estimate outcome values of outlier detection procedures by robust diagnostic criteria on X_1's outliers

	Sample sizes								
	20			50			100		
Outliers (%)	10	20	30	10	20	30	10	20	30
Hat									
MAD	0.4621	0.4472	0.6354	0.3982	0.6342	0.6691	0.7844	0.7739	0.8119
RSD	0.4127	0.4076	0.6009	0.5387	0.5983	0.7002	0.7374	0.5738	0.8003
Cooks									
MAD	0.7232	1.0000	0.9781	0.7464	0.5398	0.8193	1.0000	0.9005	1.0000
RSD	0.8701	0.8976	0.8873	0.6990	0.5502	0.7463	0.9376	0.8775	1.0000
Dffits									
MAD	0.7623	0.5892	0.8971	0.6743	0.8476	0.7800	0.6988	0.9844	1.0000
RSD	0.7533	0.6354	0.8827	0.5467	0.7588	0.7367	0.6212	0.8992	0.9564
Covratio									
MAD	1.0000	1.0000	1.0000	0.9782	0.9032	1.0000	1.0000	1.0000	1.0000
RSD	0.9901	0.8977	1.0000	0.9433	0.8874	0.9892	0.9996	1.0000	1.0000
Md									
MAD	1.0000	0.9962	1.0000	0.8099	1.0000	0.7837	1.0000	1.0000	0.8976
RSD	1.0000	0.9709	0.7925	0.9001	0.9764	0.7143	0.9987	0.9178	0.7621

Table 5: Comparisons of parameter estimate outcome values of outlier detection procedures by robust diagnostic criteria on X_2's outliers

	Sample sizes								
	20			50			100		
Outliers (%)	10	20	30	10	20	30	10	20	30
Hat									
MAD	0.6782	0.9826	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
RSD	0.7839	0.6354	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Cooks									
MAD	0.0000	0.0000	0.0000	0.5673	0.8722	0.0000	0.0000	0.0000	0.0000
RSD	0.0000	0.0000	0.0000	0.3874	0.6789	0.2873	0.0000	0.0000	0.0000
Dffits									
MAD	0.1323	0.0278	0.0007	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
RSD	0.1109	0.0034	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Covratio									
MAD	0.4673	0.5547	0.8732	0.3874	0.8739	0.9981	1.0000	1.0000	1.0000
RSD	0.1190	0.2690	0.5632	0.1187	0.5632	0.6994	0.8664	0.7698	0.9376
Md									
MAD	0.0899	0.1988	0.3894	0.5763	0.4897	0.6453	0.4762	0.7221	0.9981
RSD	0.0000	0.0000	0.0000	0.1143	0.3023	0.4887	0.2652	0.5572	0.8643

Table 6: Comparisons of parameter estimate outcome values of outlier detection procedures by robust diagnostic criteria on X_1's and X_2's outliers

	Sample sizes								
	20			50			100		
Outliers (%)	10	20	30	10	20	30	10	20	30
Hat									
MAD	0.6777	0.7898	0.7123	1.0000	0.8976	1.0000	0.8990	1.0000	1.0000
RSD	0.5678	0.6899	0.5674	0.8997	0.7765	0.8970	0.6789	0.8879	0.9665
Cooks									
MAD	0.5437	0.4675	0.0000	0.0000	0.5438	0.0000	0.0000	0.0000	0.0000
RSD	0.4894	0.3216	0.0000	0.0000	0.4755	0.0000	0.0000	0.0000	0.0000
Dffits									
MAD	0.3098	0.4789	0.6732	0.3877	0.6454	0.7254	0.6574	0.6700	0.7833
RSD	0.2887	0.3762	0.4776	0.2998	0.4998	0.5004	0.5578	0.6282	0.5830
Covratio									
MAD	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
RSD	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Md									
MAD	0.6764	0.5783	0.4387	0.7253	0.6902	0.8800	0.4873	1.0000	1.0000
RSD	0.5927	0.4011	0.5887	0.4762	0.6874	0.8376	0.4009	0.7221	1.0000

on X_2's outliers. The best criterion is Median Absolute Deviation (MAD). The best of X_2's outliers detection is Hat diagonal, its performance are good for all sample sizes and percentage of outliers which indicate that Hat diagonal measures the good and high leverage in X-space. Furthermore, the performance of covratio is better than the Md for large sample sizes and percentage of outliers.

Result from Table 6 indicates that the best criterion is Median Absolute Deviation (MAD). The best outlier detection approaches in both X_1's and X_2's outliers are covratio and Md. The performance of covratio and Md are highest values of detection outlier for all sample sizes and percentage of outliers. Furthermore, the performance of hat was better than the dffits and the Cooks for large sample sizes and percentage of outliers, the performance of dffits is better than the Cooks for all sizes excepts in sample size 20 and 10% of outliers.

CONCLUSION

According to analysis simulated data, the results indicate that the detection of outlier procedures follow the same pattern irrespective of the kind of variables contained the outliers, i.e., first, second variables or both first and second variable). CovRatio detects outliers more accurately in all set of sample sizes and in all percentages. Hat leverage detects outliers more in large sample sizes (n = 50 and 100) and when the percentages of the outliers are large. Dffits performs better with medium sample sizes (n = 50) with all percentages of outliers (10, 20 and 30%). Mahalanobis distance shows more outlier in small sample sizes at 10 and 30% outliers while the Cooks distance is more liberal among the all the outlier procedures.

All the procedures above are only use to identify points that are far from the rest because it could cause us to misinterpret patterns in a plot and can affect visual resolution of the rest of the data in the plots which force all observations into clusters. This does not mean that these points should be removed automatically. Removing them might have negative impact by destroying some important information. However, unless there are strong indications to remove outliers, alternative analysis models to ordinary least square (e.g., robust regression) should be used which weight down outlying observations.

However, the performance of robust diagnostics criteria approaches used, Mean Absolute Deviation (MAD) and Robust Standard Deviation (RSD) shows that MAD performs better than RSD for all situation and Cov-ratio identifiers the presence of outliers more often than the others for small, medium and large sample sizes with different % of outliers in the X_1, X_2-outliers and in both the X's outliers. The next best statistics for the detection are Mahalanobis distance and diffits distance.

AKNOWLEDGEMENTS

My sincere appreciation goes to Professor Y. Qin for his constructive criticisms and helpful discussions on the drafted copy of this research. I particularly thank Dr. A.A Lukman for valuable support and encouragement. I will always remain grateful to God, the Sustainer.

REFERENCES

Adnan, R., M.N. Mohamad and H. Setan, 2003. Multiple outliers detection procedures in linear regression. Math., 19: 29-45.

Albert, J. and S. Chib, 1995. Bayesian residual analysis for binary response regression models. Biom., 82: 747-769.

Alma, O.G., 2011. Comparison of robust regression methods in linear regression. Int. J. Contemp. Math. Sci., 6: 409-421.

Ampanthong, P. and S. Prachoom, 2009. A comparative study of outlier detection procedures in multiple linear regression. Proceedings of the International Multi Conference on Engineers and Computer Scientists, March 18-20, 2009, News wood Academic Publishing, Hong Kong, China, pp: 978-988.

Belsley, D.A., E. Kuh and R.E. Welsch, 1980. Identifying Influential Data and Sources of Collinearity. Wiley, New York, USA.,.

Cook, R.D., 1977. Detection of influential observation in linear regression. Technometrics, 19: 15-18.

Filzmoser, P., 2004. A multivariate outlier detection method. Proceedings of the 7th International Conference on Computer Data Analysis and Modeling, December 13-15, 2004, IEEE, Belarusian State University, Minsk, Belarus, pp: 18-22.

Gurunlu, A.O., S. Kurt and A. U-ur, 2011. Genetic algorithms for outlier detection in multiple regression with different information criteria. J. Stat. Comput. Simul., 81: 29-47.

Hadi, A.S., 1992. A new measure of overall potential influence in linear regression. Comput. Stat. Data Anal., 14: 1-27.

Kumar, N. and M. Nasser, 2012. A new graphical multivariate outlier detection technique using singular value decomposition. Int. J. Eng. Res. Technol., 1: 1-6.

Oyeyemi, G.M., A. Bukoye and I. Akeyede, 2015. Comparison of outlier detection procedures in multiple linear regressions. Am. J. Math. Stat., 5: 37-41.

Pereira, S.C. and A. Pires, 2002. Detection of Outliers in Multivariate Data: A Method Based on Clustering and Robust Estimators. In: Compstat. Wolfgang, H. and R. Bernd (Eds.). Springer, Berlin, Germany, ISBN: 978-3-7908-1517-7, pp: 291-296.

Rasheed, B.A., R. Adnan, S.E. Saffari and D.K. Pati, 2014. Robust weighted least squares estimation of regression parameter in the presence of outliers and heteroscedastic errors. J. Technol., 71: 11-17.

Welsch, R.E., 1980. Regression Sensitivity Analysis and Bounded-Influence Estimation. In: Evaluation of Econometric Models. Kmenta, J. and J.B. Ramsey (Eds.). Academic Press, Cambridge, Massachusetts, ISBN: 978-0-12-416550-2, pp: 153-167.

Wichn, D. and Wahba, 1988. Applied Regression Analysis: A Research Tool. Wordsworth and Brooks, Pacific Gross, California.