Open Access

SCHOLARENA

Journal of Mathematical and Statistical Analysis

RESEARCH ARTICLE

ISSN: 2642-018X

Alternative Robust Methods of Multivariate Outlier Detection

Md Doulah SU* and Md Islam H

Department of Statistics, Begum Rokeya University, Rangpur, Bangladesh

***Corresponding author:** Md Doulah SU, Department of Statistics, Begum Rokeya University, Rangpur, Bangladesh, Tel: +8801737087133, E-mail: sdoulah_brur@yahoo.com

Citation: Md Doulah SU, Md Islam H (2018) Alternative Robust Methods of Multivariate Outlier Detection. J Math Stat 1: 108

Article history: Received: 17 September 2018, **Accepted:** 13 November 2018, **Published:** 15 November 2018

Abstract

A multivariate outlying observation is an amalgamation of infrequent marks on as minimum two or more variables. The purpose of this study is recognition of unusual observations in multivariate dataset engaging several techniques, mainly using Mahalanobis method, Cook's distance method, Leverage point, DFFITS, Standardized residual, Studentized residual, DFBETAS. Numerous traditional unusual observations detection techniques are established on the sample mean and covariance matrix in common. Nonetheless they do not continuously show outperforms, because they theirself are affected by the outlying observations. Occasionally one outlying observation has veiled the other unusual observations. This is the masking effect of the measures and which have to detect. A suitable technique is accepted to detect the unmasking outlying observation and also to compare the several methods. That's why we have proposed two robust outlier detection methods: (i) $median|\varepsilon|-MAD|\varepsilon|$ And (ii) $(HM(|\varepsilon|)-SD_{1})$. Finally we found that our proposed methods gave the better results than any other methods in versatile aspects.

Keywords: Multivariate Outliers; Outlier Detection Methods; $median|\varepsilon|-MAD|\varepsilon|$; $(HM(|\varepsilon|)-SD_{1})$; Simulation Study

Introduction

Multivariate outlying observation recognition is the vital mission of statistical analysis of multivariate data. Many approaches have been suggested for univariate outlying observation recognition. Identifying outliers in multivariate data pose challenges that univariate data do not discussed in [1]. A multivariate outlying observation need not be an extreme in any of its workings the knowledge of extremeness arises unavoidably form some form of 'ordering' of the data introduced in [2,3].They are grounded on (robust) estimation of mean and covariance matrix of the data. A key drawback is that these procedures are free from the sample size. The foundation for multivariate outlying observation recognition is the Mahalanobis distance. The usual process for multivariate outlying observation recognition is robust estimation of the parameters in the Mahalanobis distance and the contrast with a critical value of the χ^2distribution discussed in [4,5]. Nevertheless, also observations greater than this critical value are not essentially outlying observation; they could quiet fit to the data distribution. The elementary values and difficulties of 'the ordering of multivariate data' discussed in [6,7].Attention in outlying observation recognition in multivariate data continued the identical as for the univariate instance. Extreme observations could again deliver indeed interpretable procedures of ecofriendly alarm in their particular correct and if they were not only unusual, but 'amazingly' unusual or misleading, they might again recommend that some unexpected impact is current in the data source described in [8,9].So once more we faced the concept of 'maverick' or 'rogue' unusualness which we would term outlying observation. Thus, as in a univariate sample, an unusual observation may 'stick out' so far from the others that it must be affirmed an unusual and a suitable methods of discordance may reveal that it is statistically awkward even when watched as an 'unusual'. Such an outlying observation is supposed to be conflicting and may lead us to conclude that some anomalies present in the data described in [10,11].The approaches were applied to a set of data to clarify the many outlying observation recognition procedure in multivariate linear regression models. Outliers can misinform the regression results described in [12,13].When an outlying observation is complicated in the study, it drew the regression line towards itself. This could outcome in a explanation that is more exact for the outlying observations, but less accurate for all of the other cases in the data set described in [14,15]. The outlying observation recognition challenge is one of the initial of statistical interests, and since nearly all data sets contain outlying observation of fluctuating ratios, it remains to be one of the most significant. Occasionally outlying observation can totally mislead the statistical analysis, at other times their impact may not be as obvious. Statisticians have consequently established many procedures for the identification and handling of outliers, however most of these approaches were established for univariate data sets. This paper focuses on multivariate outlying observation recognition. Mainly when using some of the general summary statistics such as the location and scale, outlying observation can cause the analyst to influence a

decision entirely contrary to the case if outlying observation weren't present. For instance, a assumption might or might not be avowed significant as a result of a handful of outlying observation. Classical outlying observation recognition is influential when the data hold a single outlier. However, the efficiency of outlier detection methods is decreasing dramatically when the dataset hold more than one outlier. This defeat of accuracy is frequently as a result of what are recognized as the masking difficulties. Furthermore, these approaches do not always flourish in identifying outlying observations. Therefore, a technique which escapes these difficulties is required. How do the outliers occur in the data sets? Outliers can also come in different flavors, depending on the environment: point outliers, contextual outliers, or collective outliers. Most common causes of outliers on data sets are: Data entry error, Appliance errors, planned errors, Data handling errors, Sampling errors and Normal. What are the problems when outlier occurs in the data sets? Outliers have been a major problem in the area of statistics, including modeling, analysis and forecasting. Lots of methods has been portrayed as a means of detecting outliers in multivariate data but not many works has been done on which of this methods is best for detecting outliers in multivariate models. This work addresses that problem by using nine multivariate outlier detection methods and checking which of them is best or more efficient in detecting outliers [16].

Proposed methods

$median|\varepsilon| - MAD|\varepsilon|$

Step 1: Run the regression model $\left(y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_{k-1} x_{(k-1)i} + \varepsilon_i\right)$
Step 2: Calculate the absolute error of the regression model
Step 3: Calculate median of the absolute error
Step 4: Calculate median absolute deviation (MAD) of absolute error
Step 5: Compute the formula $\frac{|\varepsilon| - median|\varepsilon|}{MAD_{|\varepsilon|}/0.6745}$
Step 6: Any data value that is greater than $\sqrt{\chi^2_{0.975,1}}$ is considering being an outlier(s).

(ii) $HM(|\varepsilon|) - SD\frac{1}{|\varepsilon|}$

Step 1: Run the regression model $\left(y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_{k-1} x_{(k-1)i} + \varepsilon_i\right)$
Step 2: Calculate the absolute error of the regression model
Step 3: Calculate harmonic mean (HM) of the absolute error
Step 4: Calculate standard deviation of HM is $SD_{\frac{1}{|\varepsilon|}} = \frac{1}{n-1}\sum(\frac{1}{|\varepsilon|} - HM)^2$
Step 5: Compute the formula $\frac{|\varepsilon| - HM|\varepsilon|}{SD_{\frac{1}{|\varepsilon|}}} \sim t(n-1)\ (\alpha\%)$
Step 6: Any data value that is greater than $t_{n-1}(\alpha\%)$ is considering being an outlier.

Methods and Materials

Sl No.	Methods	Formula	Cut-off value
1.	Mahalanobis Distance (Md)	$D_i^2 = (x_i - \bar{x})^T S^{-1}(x_i - \bar{x})$	$\chi^2_{(p,\ \alpha\%)}$
2.	Leverage Point (hi)	$H = X(X^T X)^{-1} X^T$.	$2p/n,\ 3p/n$
3.	DFFITS	$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\hat{\sigma}_i \sqrt{h_{ii}}}$	$2\sqrt{\frac{p}{n}}$
4.	Standardized Residuals	$t_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}}$	If $t_i\ > 3$ then we can say that it is outlier.
5.	Studentized Residuals	$t_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$	If $t_i\ > 3$ then we can say that it is outlier.
6.	DFBETAS	DFBETA $S_j = \frac{c_{ji}}{\left[\sum_{k=1}^{n} c_{jk}^2\right]^{1/2}} \frac{r_i}{s(i)(1-h_{ii})}$	$2\ /\sqrt{n}$
7.	Cook's Distance (Di)	$Di = \frac{\sum_{j=1}^{1}(\hat{Y}_i - \hat{Y}_{j(i)})2}{p^{MSE}}$	$D_i > 1$
8.	Proposed method (median|ε|-MAD|ε|)	$\frac{|\varepsilon| - median|\varepsilon|}{MAD_{|\varepsilon|}/0.6745}$	$\sqrt{\chi^2_{0.975,1}}$
9.	Proposed method $(HM(\|\varepsilon\|) - SD_{\frac{1}{|\varepsilon|}})$	$\frac{|\varepsilon| - HM|\varepsilon|}{SD_{\frac{1}{|\varepsilon|}}} \sim t(n-1)\ (\alpha\%)$	$t_{n-1}(\alpha\%)$

Table 1: Outlier Detection Methods

Here, we have taken two types of outlier detection methods: one is graphical test and another is analytical test. It is very difficult to identify the unusual observations from the graphical methods but we can suspect from the depiction. Most of the researcher used various diagram for outlier detection but we have consider three types of graphs such as scatter diagram, normal QQ plot, and box-plot diagram. Finally, we have tried to detect the actual number of outliers based on analytical methods; the commonly used analytical methods are shown in the Table 1.

Results and Discussions

(i) Example 1

In this section, datasets are extracted from [1] which has 20 observations and 6 variables.

Graphical Representation

Visualization of dataset is very important part of any kinds of analysis. Researchers firstly want to show the data pattern and based on the data pattern they take the decision what they will want to do further. But it is very difficult task sometimes. One of the difficult tasks to detect the outliers in the dataset is very critical. Here, we have considered the graphical methods to suspect the presence and absence of unusual observations in the datasets. This is the preliminary process to progress the work and finally we will apply the analytical methods for outlier detection. The results of the graphical representation are following below-

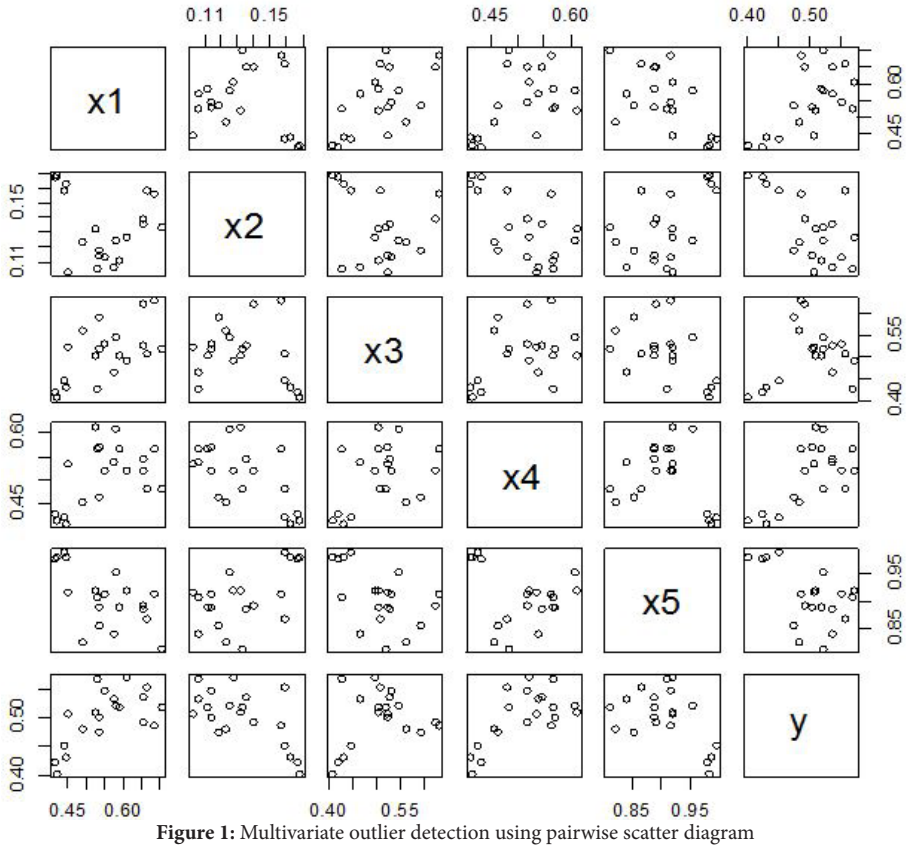


Figure 1: Multivariate outlier detection using pairwise scatter diagram

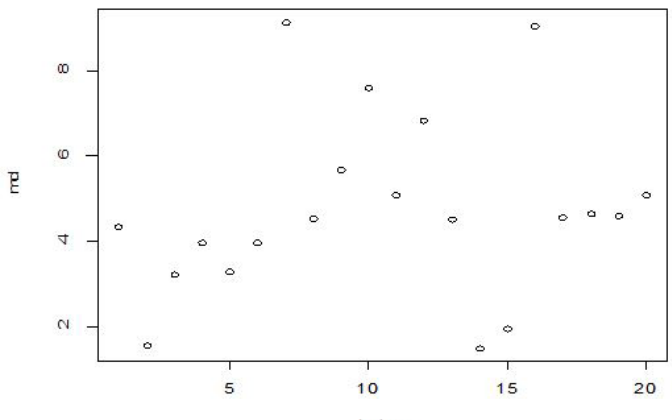


Figure 2: Mahalanobis Distance plot

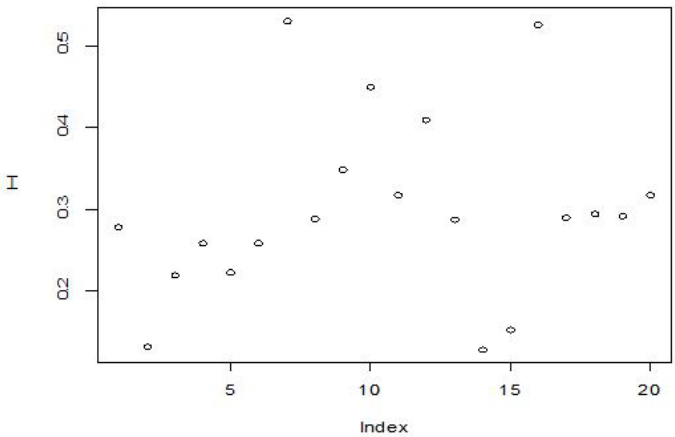


Figure 3: leverage values plot

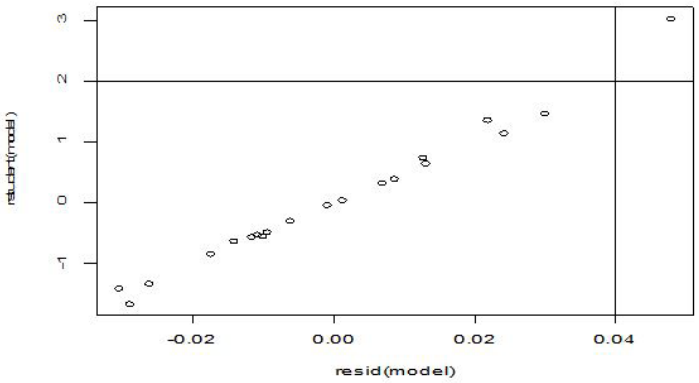


Figure 4: Regression Model Residuals plot

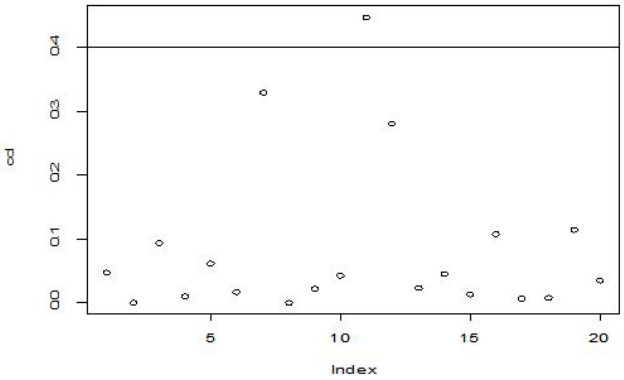
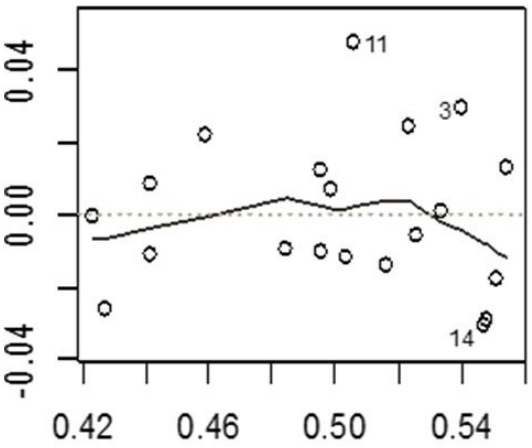
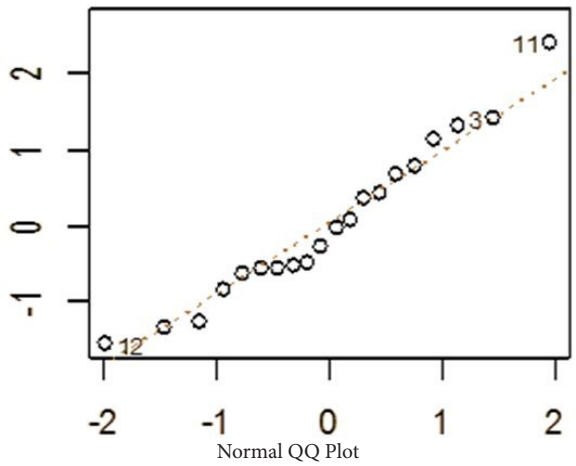


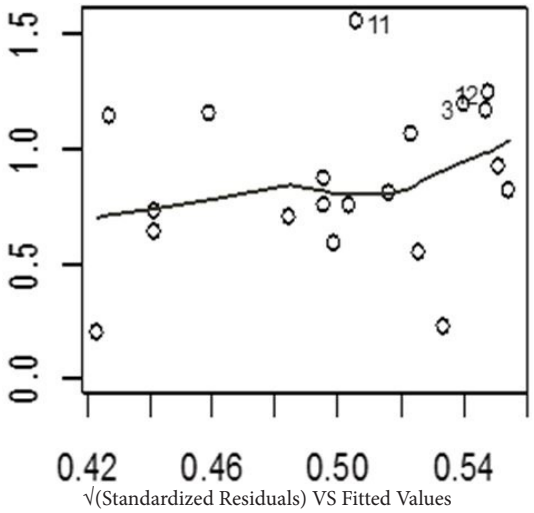
Figure 5: Cook's Distance plot



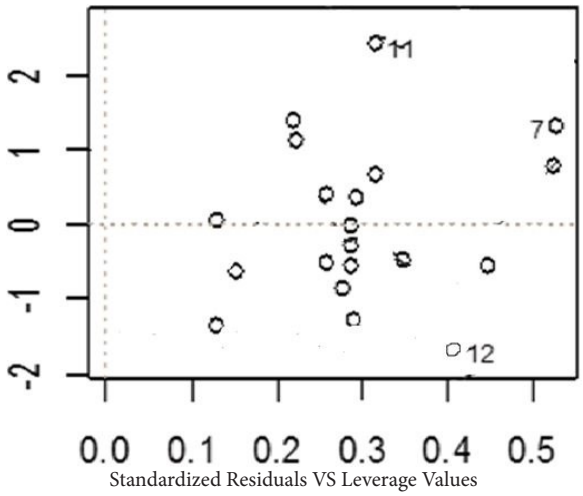
Residual VS Fitted Values



Normal QQ Plot



√(Standardized Residuals) VS Fitted Values



Standardized Residuals VS Leverage Values

Figure 6: Standard diagnostic plots

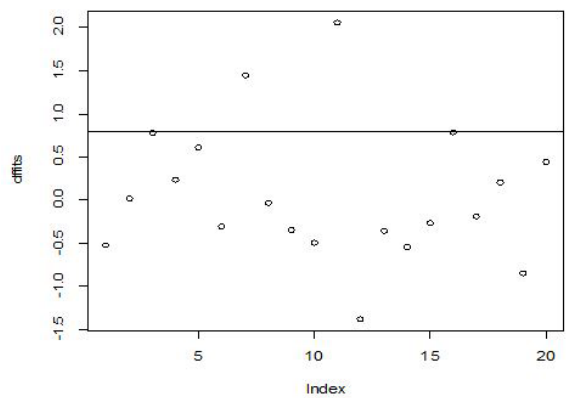


Figure 7: DFFITS plot

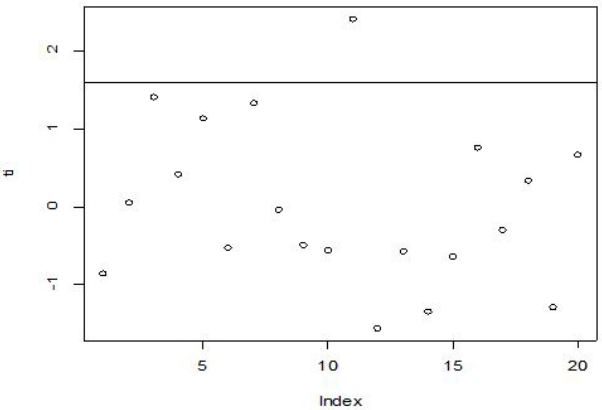


Figure 8: Standardized residuals plot

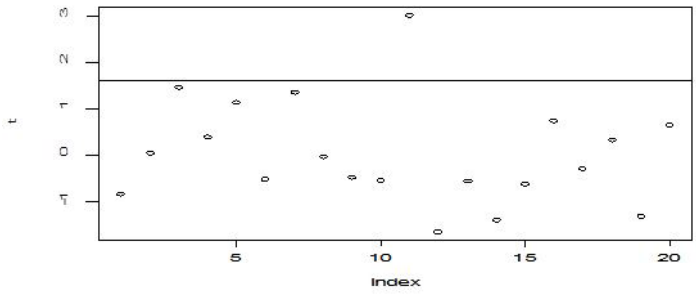


Figure 9: Studentized residuals plot

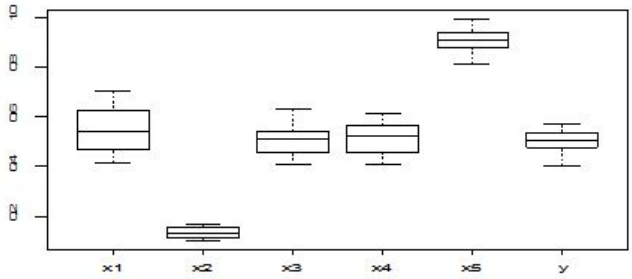


Figure 10: Box plot

A scatter diagram focuses the data dispersion that means how the data spread from one point to others. However, it is the very simple process to visualize the data pattern.. From Figure 1 we observed that there may have outlier in the dataset. From Figures 2 & 3, it is very clear that there were three outliers in the dataset. From Figure 4, 5, 6, 8 & 9 we observed that there is a single outlier in the dataset. The DEFFITS values in Figure 7 exhibited two outliers. The box-plot diagram is most popular plot to detect the outliers; sometimes it's called a five number summary. It exhibits the highest value, lowest value, 1st quartile, 2nd quartile and 3rd quartile. Therefore, it is very easy to detect outliers in this way. However, from Figure 10 we showed that there was no outlier in the dataset.

Analytically Outlier Detection

After a primary assessment of outlier revealing is showed with graphical approaches, the ultimate decision on outliers is completed using analytical techniques .The results of the analytical methods are shown in the Table 2.

Methods		Case Number	No. of outliers
Mahalanobis Distance(MDi)		-	0
Leverage Point (hii)		-	0
DFFITS		7,11,12	3
Standardized residual		-	0
Studentized residual		11	1
DFBETAS	Intercept	7,11,12	3
	x1	3,7,12,16	4
	x2	3,5,11,16	4
	x3	7,11,12,16	4
	x4	16	1
	x5	3,7,11	3
Cook's distance (*CDi*)		-	0
Proposed method (median|ε|-MAD|ε|)		11	1
Proposed method $\frac{(HM(\lvert\varepsilon\rvert)-SD_1)}{\lvert\varepsilon\rvert}$		-	0

Table 2: Number of Outliers detected by various measures

The analytical results in Table 1 deals with the procedure for computing the presence of outliers using various measures such as Mahalanobis Distance (*MDi*), Cook's Distance(*Di*), Leverage point(*hii*), DFFITS, Standardize residual, Studentized residual and DFBETAS, Proposed method (median|ε|-MAD|ε|) and Proposed method $\frac{(HM(\lvert\varepsilon\rvert)-SD_1)}{\lvert\varepsilon\rvert}$. From the dataset, the outlier identification level of Mahalanobis Distance (*MDi*), Leverage Point (*hi*), Standardized residual, Cook's distance and Proposed $\frac{(HM(\lvert\varepsilon\rvert)-SD_1)}{\lvert\varepsilon\rvert}$ methods are approximately the same, but DFFITS and DFBETAS outlier detection sensitivity is higher than others methods, since maximum number of outlier points are identified. This result clearly reveals that DFFITS and DFBETAS identify the maximum number of outliers.

(ii) Example 2

The dataset has taken from [2] which contain 7 independent variables and the response variable y. The result of the dataset is shown in the following below-

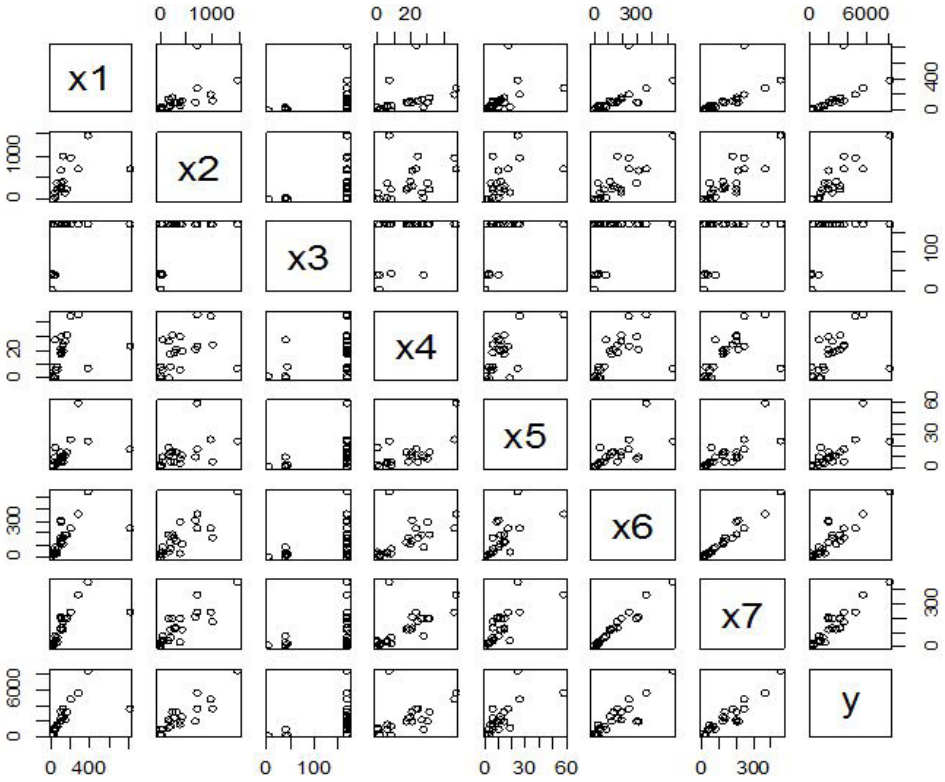


Figure 11: Multivariate outlier detection using Scatter plot

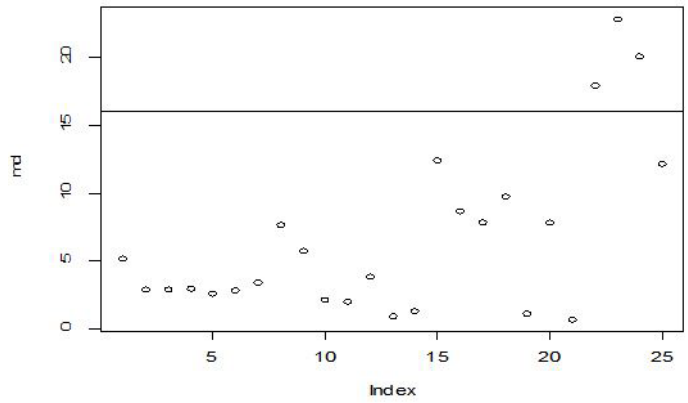


Figure 12: Mahalanobis Distance plot

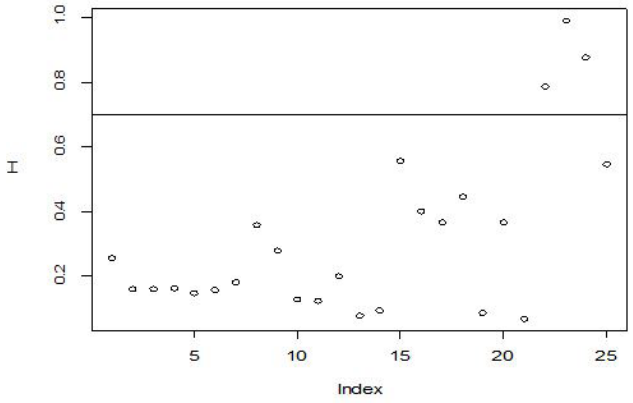


Figure 13: Leverage values plot

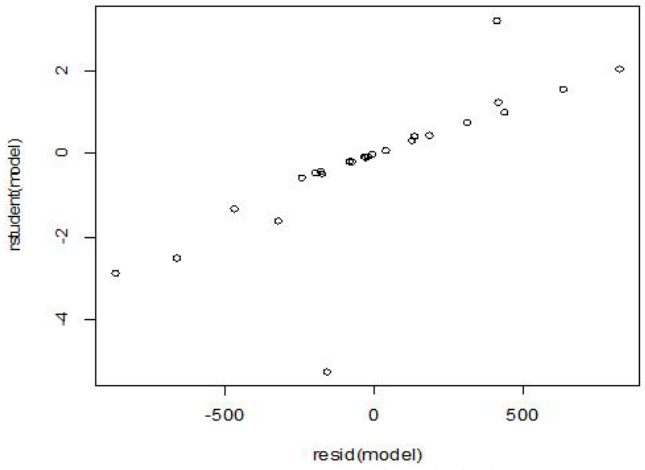


Figure 14: Regression Model Residuals plot

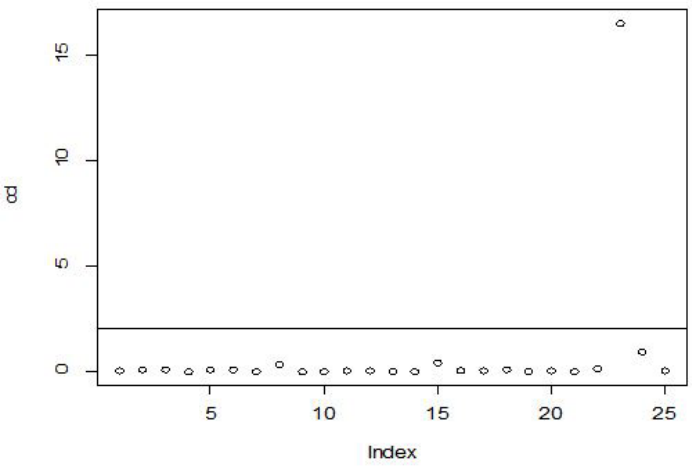


Figure 15: Cook's Distance plot

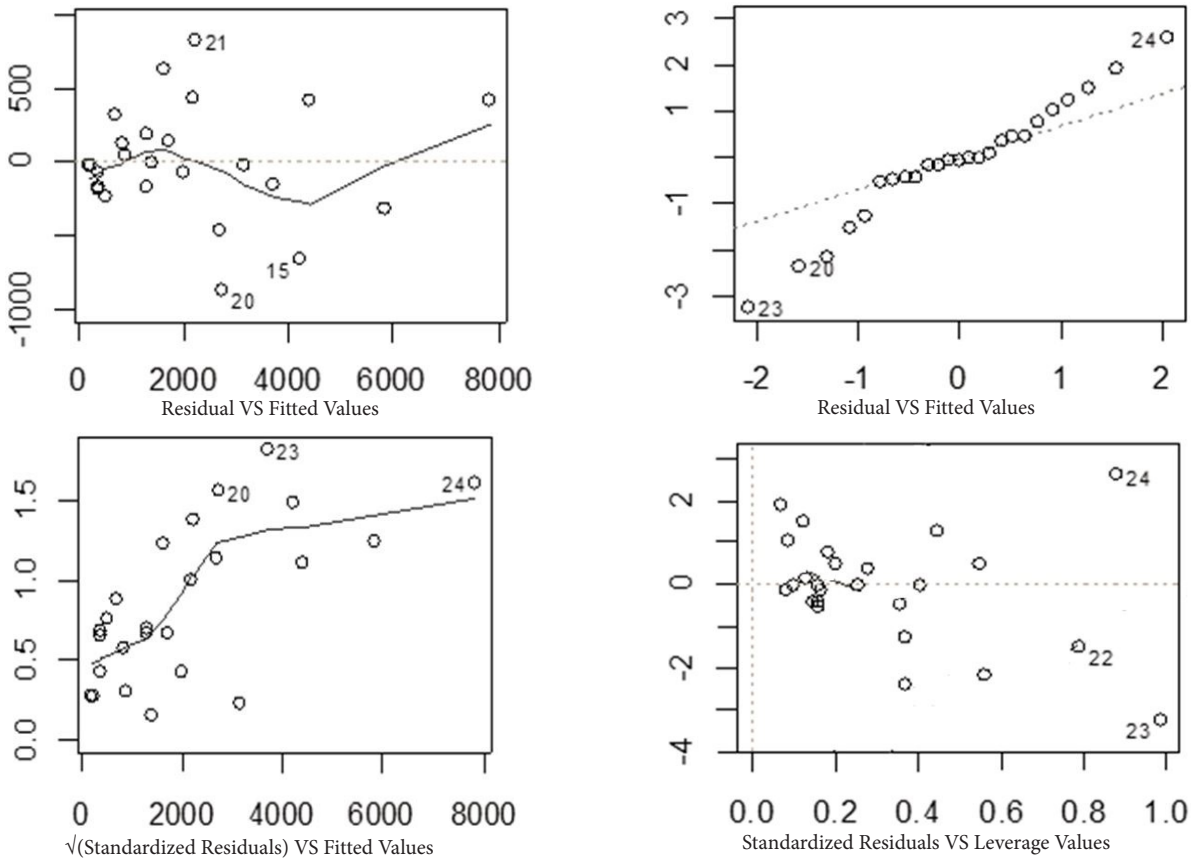


Figure 16: Standard diagnostic plots

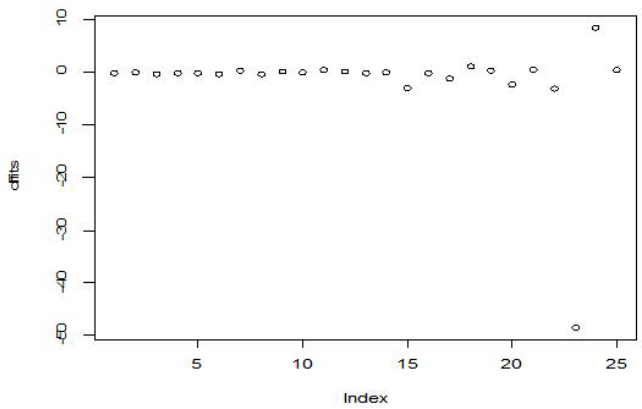


Figure 17: DFFITS plot

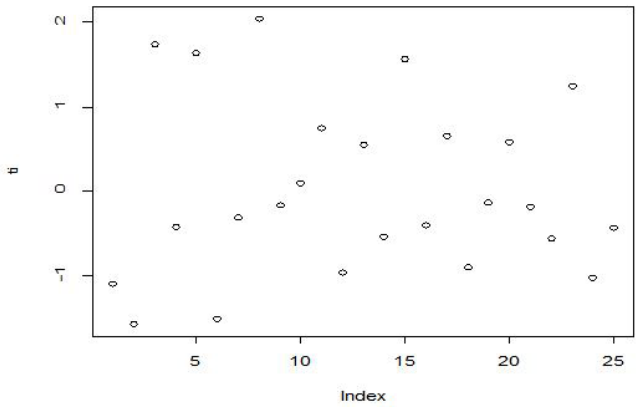


Figure 18: Standardized residuals plot

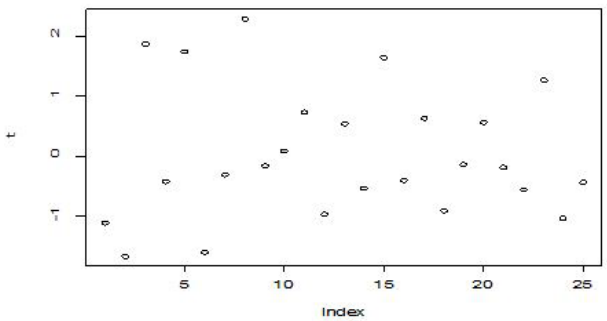


Figure 19: Studentized residuals plot

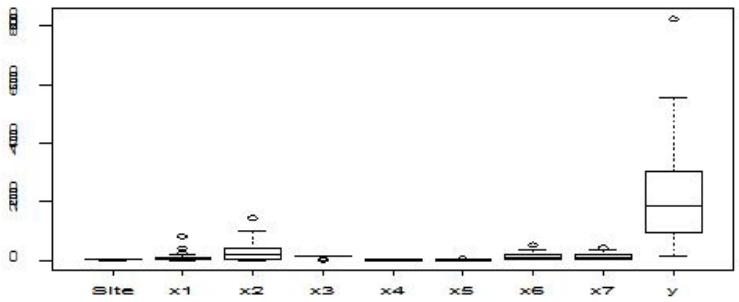


Figure 20: Box plot

In Figure 11 we showed that several pairwise scatter plots indicating the observations with outlying additive outlier value. Figures 12 & 13 depicted three outliers based on Mahalanobis distance values and leverage values. In Figure 14, 15, 16 & 17 displayed single outlier. Alternatively Figures 18, 19 & 20 revealed more than one outlier.

Methods		Case Number	No. of outliers
Mahalanobis Distance(*MDi*)		22,23,24	3
Leverage Point (*hii*)		22,23,24	3
DFFITS		17,18,20,22,23,24	6
Standardized residual		23	1
Studentized residual		23,24	2
DFBETAS	Intercept	24	1
	x1	15,23,24	3
	x2	15,17,18,29,23,24	6
	x3	23,24	2
	x4	15,18,20,23,24	5
	x5	15,17,22,23,24	5
	x6	15,17,20,23,24	5
	x7	15,17,20,23,24	5
Cook's distance (*CDi*)		22,23,24	3
Proposed method (median|ε|-MAD|ε|)		15,20,21	3
Proposed method $(HM(\lvert\varepsilon\rvert) - SD_{\lvert\varepsilon\rvert})$		20,21	2

Table 3: Number of Outliers detected by various measures

From Table 3 we observed that DFFITS and DFBETAS detected highest number of outliers than all others methods.

Simulation Study

Here, we have compared the ability of outlier detection in different methods based on simulation process. We have generated data from a multivariate normal distribution. Firstly, we generate the dataset which is absolutely free from outliers. The results of the study are shown in the Table 4. Secondly we generate the data set which contains outliers. The certain percentage of outlier was casted in the data set at random. In our study, we have taken 5% outliers in data sets of different sizes such as n = 100, 500 and 1000. The results of the simulation study are shown in the following Table 5.

Methods	n=100 (in percentage)	n=500(in percentage)	n=1000(in percentage)
Mahalanobis Distance(*MDi*)	**67.88**	**72.74**	**83.67**
Leverage Point (*hii*)	68.07	78.57	86.06
DFFITS	71.05	88.46	92.85
Standardized residual	73.45	84.17	89.01
Studentized residual	65.13	73.48	84.06
DFBETAS	82.57	89.99	96.57
Cook's distance (*CDi*)	68.44	75.82	89.67
Proposed method $(median\lvert\varepsilon\rvert - MAD\lvert\varepsilon\rvert)$	4.97	7.87	10.52
Proposed method $(HM(\lvert\varepsilon\rvert) - SD_{\lvert\varepsilon\rvert})$	3.20	6.80	9.02

Table 4: Outlier detection percentage when dataset free from outlier

Methods	n=100 (in percentage)	n=500(in percentage)	n=1000(in percentage)
Mahalanobis Distance(*MDi*)	**72.11**	**79.54**	**83.09**
Leverage Point (*hii*)	73.28	78.56	81.88
DFFITS	87.89	92.87	96.35
Standardized residual	76.17	86.39	90.09
Studentized residual	75.65	87.36	93.14
DFBETAS	81.89	91.87	96.35
Cook's distance (*CDi*)	78.99	87.05	92.06
Proposed method $(median\lvert\varepsilon\rvert - MAD\lvert\varepsilon\rvert)$	**93.05**	**96.88**	**99.87**
Proposed method $(HM(\lvert\varepsilon\rvert) - SD_{\lvert\varepsilon\rvert})$	**91.01**	**94.78**	**99.07**

Table 5: Outlier detection percentage when dataset contain 5% outliers

We have taken two categories: one is free from outliers, as presented in Table 4 and for comparison; another is based on 5% contaminations, as presented in Table 5. However, as we have noticed in Table 4 and in our simulations, the methods always yield a massive observations that are (wrongly) specified as outliers except from our proposed two methods but the actual situation is the simulation study was free from contamination. In Table 5 we have displayed the simulation study results of nine methods most of the methods were unable to identify as outliers except only detected our proposed two methods. It is clear that our proposed two methods outperforms considerably with respect to the detection of the accurate outliers. The performance becomes even more obvious as the sample size increases.

Conclusions

To sum up the whole discussion, we have compared the several outlier detection methods such as Mahalanobis Distance *(MDi)*, Cook's Distance(*Di*), Leverage point(*hii*), DFFITS, Standardize residual, Studentized residual, DFBETAS, Proposed method $(median|\varepsilon| - MAD|\varepsilon|)$ and Proposed method $(HM(|\varepsilon|) - SD_{|\varepsilon|})$. In this work, the outlier detection way of Mahalanobis Distance (*MDi*), Leverage Point (*hi*) and Proposed methods (median|ε|-MAD|ε are approximately the same, this result clearly reveals that DFFITS identify the maximum number of outliers. In this paper we have used several methods for outlier detection. It is very difficult task to recommend one method to detect outlier, because most of the methods have masking and swamping problems in different contexts. Here we can recommend that our proposed two methods may give you a better yield than any other methods.

References

1. Davies L, Gather U (1993) The identification of multiple outliers. J Am Stat Assoc 88: 782-92.

2. Marsh GM (2002) Standard protocol for outlier analysis of dissatisfaction Rate, Technical Report, University of Pittsburgh. University of Udine.

3. Ben-Gal I (2005) Outlier detection, Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers, Kluwer Academic Publishers.

4. Rubinstein RY, Kroese DP (2008) Simulation and the Monte Carlo Method, (2nd edn). New-York: Wiley.

5. Rousseeuw P, Hubert M (2017) Anomaly Detection by Robust Statistics. Journal of the Interdisciplinary Reviews: Data Mining and Knowledge Discovery.

6. Barnett, Lewis (1994) Outliers in statistical data. (3rd edn) Wiley.

7. Bendre SM, Kale BK (1987) Masking effect on test for outliers in normal sample. Biometrika 74: 891-6.

8. Draper N, Smith H (1998) Applied Regression Analysis, 3rd (edn) Wiley, USA.

9. Kannan KS, Manoj K (2015) Outlier Detection in Multivariate Data. Appl Math Sci 9: 2317-24.

10. Iglewicz, Hoaglin (1993) How to detect and handle outliers (E BOOK). ASQC Quality Press.

11. Nkechinyere EM, Andrew I, Idochi O (2015) Comparison of Different Methods of Outlier Detection in Univariate Time Series Data. Intl J Res Math Stat 208-2662.

12. Rousseeuw P, Leroy A (1987) Robust Regression and Outlier Detection. Wile, USA.

13. Yan X, Su GX (2009) Linear regression analysis. Theory and computing.

14. Carling K (2000) Resistant outlier rules and the non-Gaussian case. Comput stat data anal 33: 249-58. Wiley.

15. Hoaglin D, Tukey JW (1986) Performance of some resistant rules for outlier labeling. J Am Stat Assoc 81: 991-9.

16. Michael J, Crawley (2007) The R Book, version 3.2. John Wiley & Sons, England.