http://dx.doi.org/10.5455/sf.XXXXX

Science Forum (Journal of Pure and Applied Sciences)

journal homepage: www.atbuscienceforum.com

# Evaluation of outlier detection procedures in multiple linear regressions

Rufai Iliyasu[1*], Abbas Umar Farouk[2], Abdulazeez Lasisi Kazeem[2], Yusuf Sani Muhammad[1], Salihu Gambo[3], Ismail Manzo[1], Zurki Ibrahim[4], Sabo Muhammad[1], Hassan Adamu[1], Auwal Saleh[1]

[1] *Department of Statistics, Binyaminu Usman Polytechnic, Jigawa, Nigeria*

[2] *Department of Mathematical Sciences, Abubakar Tafawa Balewa University, Bauchi, Nigeria*

[3] *Department of Statistics, Kano State Polytechnic, Kano, Nigeria*

[4] *Department of Biological Science Sule Lamido University Kafin Hausa, Jigawa, Nigeria*

## ABSTRACT

Regression analysis is conceptually the simplest method used for investigating the functional relationship between dependent and independents variables. In this paper, the problems of over and under detection of outliers in datasets are put to test by applying the various methods to the dataset without outlier injection at various sample sizes.

This study reviews methods of outlier detection in multiple linear regressions using DFFITS, Cook's distance, DFBETAS, R-students, and Mahalanobis distance. It was seen from the result analyzed that the methods of outlier detection had different performances when detecting outliers in a dataset at various sample sizes. Data simulation was carried out without injection of outliers to independent and dependent variables.

The R-code simulation shows the performance of five outlier detection methods in multiple linear regressions. From the five techniques compared, DFBETAS performed better than all the other methods for all the sample sizes except at sample size 10. The next best method was Cook's distance, specifically for the higher sample sizes of 30, 50, and 100. Mahalanobis and DEFFITS were more liberal among the all other outlier procedures.

## ARTICLE INFO

***Article history:***
Received 15 January 2021
Received in revised form
31 July 2021
Accepted 01 August 2021
Published 27 April 2022
Available online 27 April 2022

## KEYWORDS

Outliers
Outlier detection
Multiple linear regressions
Simulation

## 1. Introduction

Regression analysis is conceptually the simplest method used for investigating the functional relationship between dependent and independents variables. Several works have been carried out on outlier detection and how to tackle it if present in a data analysis. An outlier is an observation that lies outside the overall pattern of a distribution (Moore and Mccabe, 1999). A convenient definition of an outlier is a point which falls more than 1.5 times the interquartile range above the third quartile or below the first quartile. It can also occur when comparing the relationship between two sets of data. According to Oxford's Dictionary of Statistics (2008), an outlier is an observation that is very different to other observations in a set of data. It is a data value which is unusual with respect to the group of data in which it is found. It may be a single isolated value far away from all others or a value which does not follow the general pattern of the rest. Usually, the presence of outliers indicates some sort of problem.

For example, the university management may want to relate the performance of students with the age of the students or the number of hours spent by the

* **Corresponding author** Rufai Iliyasu ✉ iliyas.rufai@gmail.com ✉ Department of Statistics, Binyaminu Usman Polytechnic, Jigawa, Nigeria.

© 2022 Faculty of Science, ATBU Bauchi. All rights reserved

students using Internets. Also, one may want to examine whether the consumption of cigarettes has any relation with some socioeconomic and demographic variables such as sex, age, education, income, and price of cigarettes. Another example is in engineering real estate appraisal. The price of a home is the response variable, while may be the taxes paid for the building and the characteristics of the building are the explanatory variables.

The relationship is normally expressed in an equation form or model that connects the dependent variable with one or more predictors or explanatory variables. For example, in the consumption of cigarettes, the dependent variable is cigarette consumption (measured by the number of packs of cigarettes sold in a given state on a per capita basis during a given year) and the predictors or the explanatory variables are the various demographic and socioeconomic variables.

Regression models are the most commonly used methods to investigate or study the functional relationship existing between a response or dependent variable (*Y*) and explanatory or independent variable(s) (*Xs*). 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 *y* on the regressors $X_1, X_2, \ldots, X_P$, $p \geq 1$.

Sometimes, however, there might be outliers contained in the dependent variables, i.e., the *Y* values, the independent variables, i.e., the *X* values, or in both *Y* and *X* values. In this situation, many procedures of estimation in either linear or multiple regressions model may hardly be precise.

An outlier is a data point which is significantly different from the remaining data. Hawkins formally defined the concept of an outlier as follows: an outlier is an observation which deviates so much from the other observations in order to arouse suspicions that it was generated by different mechanisms. Outliers are also referred to as abnormalities, discordant, deviants, or anomalies in the data mining and statistics literature. In most applications, the data are created by one or more generating processes, which could either reflect activity in the system or observation collected about entities. When the generating process behaves in an unusual way, it sometimes results in the creation of outliers.

Several works have been carried out on outlier detection and how to tackle it if present in a data analysis, for example, see detecting multiple outliers in multiple linear regressions by Adnan et al. (2003). Adnan et al. (2003) describe a procedure for identifying multiple outliers in linear regre2ssion. Tiwari et al. (2007) evaluate five different methods and treatment techniques for outlier detection. The Bayesian approach to detect linear regression with heteroscedastic noise under linear regression model (Pimpan and Suwattee, 2009; Ting et al., 2007) compared some outlier detection procedures in multiple linear regressions using eight statistics. Chrominsk and Tkacz (2010) compared outlier detection method in biomedical data. She and Owen (2010) studied the outlier detection procedure using a non-convex penalized regression. Pimpam (2012) compared four outlier detection procedures in multiple linear regressions using five regressors. Turkan et al. (2012) used regression diagnostics based on robust parameter estimates to detect the outlier. Yuen and Mu (2012) proposed a novel probabilistic method for robust parametric identification and outlier detection in linear regression problems. Rahman et al. (2012) used Cook's distance and DEFFITS in multiple linear regression models to detect the outliers. Kuppusamy and Kaliyaperumal (2012) compared some methods for outlier detection. Tripathy et al. (2013) compared some statistical methods for detecting outliers in proficiency testing data on lead in aqueous solution analysis. Zakaria et al. (2014) proposed a measure for detecting influential outliers in linear regression analysis. Nacy et al. (2015) proposed a new method for detecting and estimating outliers in multiple linear regression models. Oyeyemi et al. (2015) compared some methods for outlier detection. Zakaria et al. (2014) proposed a method called coefficient of determination ratio for the detection of influential outliers in linear regression analysis. Choi et al. (2018) and Cui et al. (2019) compared the proposed five outlier detection methods under some conditions which consisted of the combination of the outlier situations, several error distributions, and the number of predictor variables. On the basis of the linear model, this paper starts with a sum of squares of residuals based on the data deletion model, mean shift model, and the sample quantile method. The population parameters are estimated and a special statistic is used to detect outliers.

## 2. Methodology

### *2.1. Introduction*

There are many methods for the detection of outliers in multiple linear regressions. They are classified into two: graphical and analytical methods. However, five different methods of detecting outliers are considered in this study. They are Cook's square distance, DEFFITSdistance, Mahalanobis distance, DFBETAS,

and *R*-student. These methods are both graphical and analytical and their procedures are explained below.

#### 2.1.1. Cook's square distance

Cook's square distance (1977) of unit $i$ is a measure based 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 (1989) suggest examining cases with $CD_i^2 > 0.5$, and that case where $CD_i^2 > 1$ should always be studied. This distance measure can be expressed in the following general form:

$$CD_i^2 = (\hat{\beta}_i - \beta)'(X'X)(\hat{\beta}_i - \beta) / P\hat{\sigma}$$

where $i = 1,2...n$. However, substituting $CD_i^2$ statistic may also be rewritten as follows:

$$CD_i^2 = [(e_i^2 / p)(h_{ii} / (1 - h_{ii})]$$

All are related to the full data.

#### 2.1.2. DEFFITS distance

For each observation $i$ compute, $(\hat{y}_i - y_{i(i)})$ or $(h_{ii}e_i)/(1-h_{ii})$, which determines how much the predicted value $\hat{y}_i$ at the design point *xi* would be affected if the $i$th case was deleted. The standardized version of DEFFITS is

$$DEFFIT = [(h_{ii}^{\frac{1}{2}}e_i) / (\sigma_i(1 - h_{ii})]$$

where $i = 1,2...n$. Belsley et al. (1980) suggested that any observation for which $/DEFFIT/ > 2/\sqrt{p/n}$ warrants attention for outliers.

#### 2.1.3. Mahalanobis distance

The measure of leverage by means of $MD_i$ (Mahalanobis distance) is as follows:

$$MD^2 = (\mu_i - \bar{\mu})\sigma^{2-1}(\mu_i - \bar{\mu})' = (n-1)(h_{ii} - \frac{1}{n})$$

$$\text{where } i = 1,2,3.....n, \bar{\mu} = \Sigma\mu_i / n, \text{and } \sigma^2 = \frac{1}{n-1}\sum(\mu_i - \bar{\mu})'(\mu_i - \bar{\mu})$$

If $MD^2 > X^2_{P-1,0.9}$ where $X^2_{P-1,0.9}$ is the 9th percentile of a chi-square distribution with $p$ −1 degrees of freedom, then there is an outlier.

#### 2.1.4. *R*-student

A common way to model an outlier is by using the mean shift outlier model. However, the *R*-student statistic will be more sensitive at this point. A formal testing procedure for outlier detection based on *R*-student is given as follows:

$$ti = ei / / \sqrt{\hat{o}^2_{(i)}(1 - hii)},$$

where $i = 1, 2..., n$, $/ti/ > t(\alpha/2n)$, and $n - (p - 1) >$ indicate the existence of outliers. This is referred to as an estimate of Mean Square Error (MSE) based on a dataset with the $i$th observation removed. The estimate of MSE so obtained from the $i$th observation is as follows:

$$\sigma^{\wedge 2}_{(i)} = [(n - p)\,MSE - ei / (1 - hii)] / [n - (p + 1)].$$

In this work, data will be simulated from multiple linear models with five independent variables using the *R* statistical package.

#### 2.1.5. DFBETAS

It is a rule of thumb that drop observations have too much influence on the regression line, where "too much" is defined by the 2/sqrt ($n$) or 1.00 cut-offs. Choices about outliers are controversial. Some statisticians would suggest that dropping valid observations (i.e., ones that conform to the design of the survey instrument) is never a good idea: the sample is what it is. This may result in a bias or the range over which the regression is applicable may change. For instance, in the gender dataset, many of the high leverage, high-influence observations are for people who are very old or very young. One strategy may be to restrict the nature of the study—to those individuals between 18 and 65 years. Yet, we lose something in the process.

### *2.2. Method of simulation and analysis*

A set of replication of datasets were generated from the multiple linear regression models with three independent variables stated as follows:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e_i + i = 1,2,\ldots,n$$

where all regression coefficients $\beta_i$ were fixed to be $\beta_i = 1, i = 0,1,2$ and the errors are assumed to be independent.

The *R*-statistic simulation would be used to show the best method for detecting the absence of outliers among the five methods under study.

This work reviews and compares five different methods of outlier detection.

The comparison of five detection statistics has been carried out by the following steps:

1) Generation of the data with certain percentage of *X* outliers, *Y* outliers, and both *X* and *Y* outliers at different sample sizes.

2) Each statistic is computed from each of the 1,000 replications.

3) Comparison of the detection of outliers by counting the number of times each statistic can detect the

**Table 1.** Probability values obtained from 1,000 simulated datasets from normal distribution without the injection of outliers.

| Sample size | Outlier detection techniques | | | | |
|---|---|---|---|---|---|
| | DEFFITS | Mahalanobis | Cook's distance | *R*-students | DFBETAS |
| 10 | 0.025 | 0.979 | 0.732 | 0.892 | 0.969 |
| 30 | 0.451 | 0.377 | 0.980 | 0.926 | 1.000 |
| 50 | 0.754 | 0.136 | 0.999 | 0.887 | 1.000 |
| 100 | 0.952 | 0.013 | 1.000 | 0.841 | 1.000 |

**Table 2.** Wrong detection of one outlier from the five methods adopted. Simulated datasets from normal distribution without the injection of outliers gave different probability values in 1,000 simulations.

| Sample size | Outlier detection techniques | | | | |
|---|---|---|---|---|---|
| | DEFFITS | Mahalanobis | Cook's distance | *R*-students | DFBETAS |
| 10 | 0.197 | 0.021 | 0.247 | 0.108 | 0.030 |
| 30 | 0.405 | 0.456 | 0.016 | 0.074 | 0.000 |
| 50 | 0.217 | 0.386 | 0.001 | 0.113 | 0.000 |
| 100 | 0.048 | 0.064 | 0.000 | 0.146 | 0.000 |

absence of outliers. The variation in comparison of the five outlier detection methods provides an indication of the sensitivity of the methods.

## 3. Results and Discussion

The R-code computation for the outlier detection gives the correct probability of each of the methods or techniques for four different sample sizes of $n$ = 10, 30, 50, and 100 computed from 1,000 replications. Moreover, five different methods were used to analyze the simulated data. (Cook's square, DFBETAS, *R*-students, Mahalanobis, and DEFFITS).

The below computation gives the best outlier detection procedure for different sample sizes from 1,000 replications. Table 1 gives the results of five outlier detection approaches.

The five methods of outlier indicate different degrees of performance and consistency when detecting outliers at various sample sizes of 10, 30, 50, and 100, respectively. For a small sample size of 10, Mahalanobis had the best performance of reporting no outliers in the absence of outliers in the datasets with the probability value of 0.979, followed by the performances of DFBETAS, *R*-students, Cook's distance,

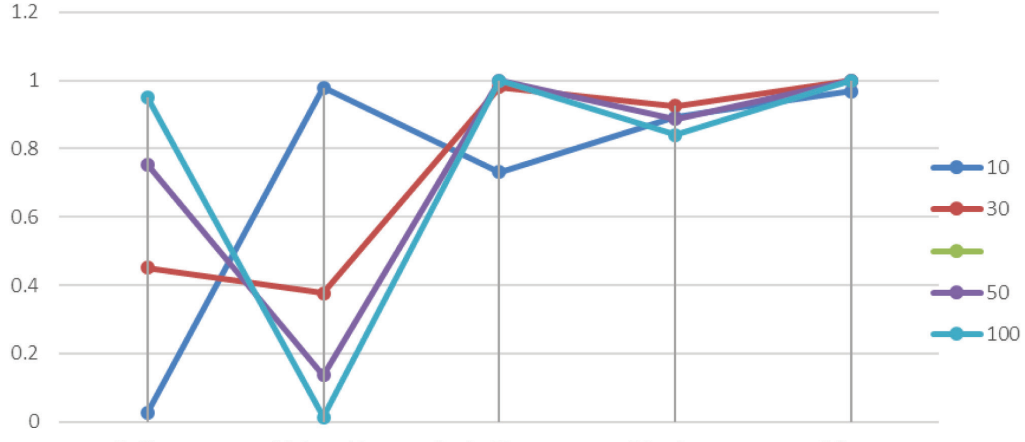


FIGURE 1. Power performance and consistency when detecting the absence of outliers at various sample sizes.

**Figure 1.** Power performance and consistency when detecting the absence of outliers at various sample sizes.

**Table 3.** Wrong detection of two outliers from the five methods adopted. Simulated datasets from normal distribution without the injection of outliers gave different probability values in 1,000 simulations.

| Sample size | Outlier detection techniques | | | | |
|---|---|---|---|---|---|
| | DEFFITS | Mahalanobis | Cook's distance | *R*-students | DFBETAS |
| 10 | 0.442 | 0.000 | 0.000 | 0.000 | 0.001 |
| 30 | 0.129 | 0.149 | 0.000 | 0.000 | 0.000 |
| 50 | 0.028 | 0.343 | 0.000 | 0.000 | 0.000 |
| 100 | 0.000 | 0.219 | | 0.013 | 0.000 |

**Table 4.** Wrong detection of three outliers from the five methods adopted. Simulated datasets from normal distribution without the injection of outliers gave different probability values in 1,000 simulations.

| Sample size | Outlier detection techniques | | | | |
|---|---|---|---|---|---|
| | DEFFITS | Mahalanobis | Cook's distance | *R*-students | DFBETAS |
| 10 | 0.295 | 0.000 | 0.000 | 0.000 | 0.000 |
| 30 | 0.015 | 0.017 | 0.000 | 0.000 | 0.000 |
| 50 | 0.001 | 0.121 | 0.000 | 0.000 | 0.000 |
| 100 | 0.000 | 0.261 | 0.000 | 0.000 | 0.000 |

and DEFFITS, with probability values of 0.969, 0.892, 0.732, and 0.025, respectively.

For a sample size of 30, DFBETAS correctly evaluated the absence of outliers in 1,000 replicates. However, the performance of *R*-students, DEFFITS, and Cook's distance improved as compared to probability values obtained in the sample size of 10. As the sample sizes increases, Mahalanobis performance tends to poorly detect the absence of outliers correctly.

DFBETAS and Cook's distance performances tend to improve as the sample sizes increase with probability values 1.000 and 0.999 for sizes 50 and 1.000 and 1.000 for size 100, respectively. Mahalanobis' wrong detection of outliers increased as the sample sizes increased with probabilities of 0.136 and 0.013 for sample sizes 50 and 100, respectively. However, the performance of *R*-students and DEFFITS also improved.

The result from Table 2 shows that Cook's distance wrongly detected the presence of one outlier in the small sample size of 10 with a probability value of 0.247 in 1,000 simulations, followed by DEFFITS, *R*-students, and DFBETAS with probability values 0.197, 0.108, 0.030, and 0.021 respectively. As the sample size increased, Mahalanobis had more false identifications of the presence of one outlier than the remaining procedures with probability values of 0.456, 0.386, and 0.064 at sample sizes of 30, 50, and 100, respectively. Cook's distance tends to improve in identifying the absence of outlier as the sample sizes increase. However, DFBETAS had the best

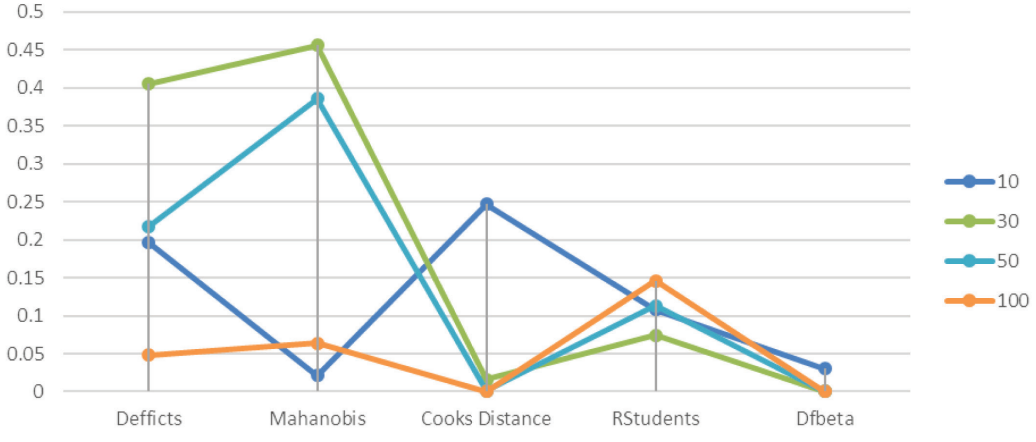


**Figure 2.** Wrong detection of one outlier from the five methods adopted.

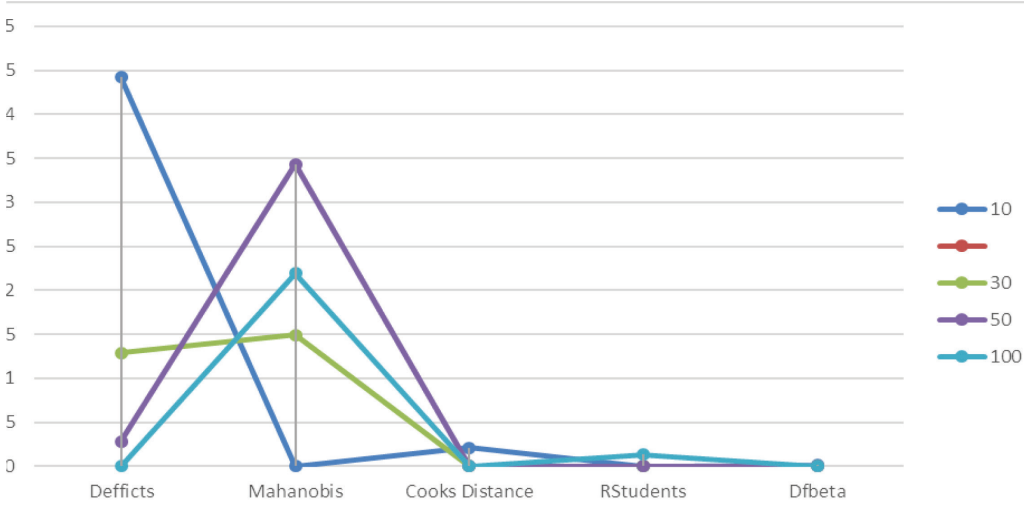


**Figure 3.** Wrong detection of two outliers from the five methods adopted.

identification of the absence of outliers with probability values of 0.000, 0.000, and 0.000 at sample sizes of $n$ = 30, 50, and 100, respectively.

Mahalanobis and *R*-students correctly identified the absence of outliers at a sample size of 10 with 0.000 probability. At sample sizes $n$ = 30 and 50, Cook's distance, *R*-students, and DFBETAS had 0.000, 0.000, and 0.000 probabilities, which indicate they all correctly detected the absence of two outliers in the datasets. At sample size 100 DFBETAS, Cook's distance, and DEFFITS had the best observation of the absence of outliers. However, Mahalanobis' wrong detection of outliers continued as the sample size increased.

*R*-students, Cook's distance, and DFBETAS had the best performance showing no presence of three outliers in 1,000 simulations at all the sample sizes tested. DEFFITS' performance tended to improve as the sample sizes increased. However, Mahalanobis' poor performance increased as the sample increased.

## 4. Conclusion

It is seen from the results analyzed that the methods of outlier detection had different performances when detecting outliers in a dataset at various sample sizes. Data simulation was carried out without the injection of outliers to independent and dependent variables.

The R-code simulation showed the performance of five outlier detection methods in multiple linear regressions. From the five techniques compared, DFBETAS performed better than all the other methods for all the sample size except at sample size 10. The next best method was Cook's distance, specifically for the higher sample sizes of 30, 50, and 100.

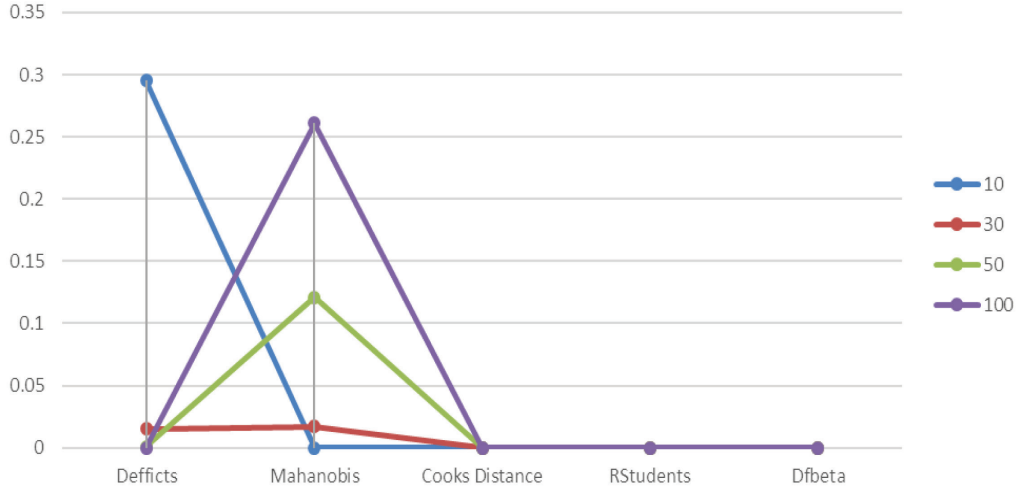


**Figure 4.** Wrong detection of three outliers from the five methods adopted.

Mahalanobis and DEFFITS were more liberal among the all other outlier procedures.

## Acknowledgments

The author sincerely acknowledge the contribution of our supervisor Dr. Abbas Umar Farouk and his friend and colleague Abdulwasiu Bukoye whose contributions to the success of the study cannot be quantified.

## REFERENCES

Adnan R, Mohamad MN, Setan H. Multiple outlier detection procedure in linear regression. Mat 2003; 19:29–45.

Belsley DA, Kuh E, Welsch RE. Regression diagnostics: identifying influential data and source of collinearity. John Wiley & Sons, New York, NY, 1980.

Cook RD. Detection of influential observations in regression. Technometrics 1977; 19:15–8.

Chrominsk K, Tkacz M. Comparison of outlier detection method in biomedical data. J Med Inform Technol 2010; 16:89–94.

Cook RD, Weisberg S. Residuals and influence in regression. Chapman & Hall, New York, NY, 1989.

Choi Y, Park CG, Lee KE. Evaluation of outlier detection methods for multiple linear regression model. J Korean Data Inform Sci Soc 2018; 29(6):1663–77.

Cui L, Cheng L, Jiang X, Chen Z, Albarka. Robust estimation and outlier detection based on linear regression model. J Intell Fuzzy Syst 2019; 37(4):4657–64; doi:10.3233/JIFS-179300

Kuppusamy M, Kaliyaperumal K. Comparison of method for detecting outlier. Int J Sci Eng Res 2012; 4(9):707–14.

Moore DS, McCabe GP. Introduction to the practice of statistics. Freeman & Company, New York, NY, 1999.

Nacy NG, Raho GI, Ahmad ZS. A new method for detection and estimating outliers in multiple linear regression model. Int J Comput Technol 2015; 3(29):120–8.

Oyeyemi GM, Bukoye A, Akeyede I. Comparison of outlier detection procedure in multiple linear regression. Am J Math Stat 2015; 5(1):37–41.

Pimpam A. Comparison of outlier detection in multiple linear regressions proceeding of the world congress on engineering. 2012, vol 3(1), pp. 978–88. ISBN: 978-988-19251-3-8 ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Pimpan A, Suwattee P. A comparative study of outlier detection procedure in multiple linear regression. Proceeding of the multi conference of engineers and computer scientist, Hong Kong, China, 2009, vol. 2(2), pp. 978–88.

She Y, Owen AB. Outlier detection using nonconvex penalized regression. J Am Stat Assoc. 2012; 106(494): 626–39.

Ting J-A, Theodorou E, Schaal S. A Kalman filter for robust outlier detection. Computational Learning & Motor Control Lab University of Southern California IROS, Log Angels, CA, 2007.

Tiwari K, Mehta K, Jain N, Tiwari R, Kanda G . Selecting the appropriate outlier treatment for common industry application. In NESUG Conference Proceedings on Statistics and Data Analysis, Baltimore, Maryland, 2007.

Tripathy SS, Saxena RK, Gupta PK. Comparison of statistical methods for outlier detection in proficiency testing data on analysis of lead in aqueous solution. Am J Theo App Stat 2013; 2(6):233–42.

Turkan S, Cetin MC, Toktamis ¨O. Outlier detection by regression diagnostics based on robust parameter estimates. Hacettepe J Mathematics Stat 2012; 41:147–55.

Yuen K, Mu H. A method probabilistics method for Robust parametric identification and outlier detection. Probabilistic Eng Mech 2012; 30(2012):48–59.

Zakaria A, Howard N, Nkansah BK. On the detection of influential outlier in linear regression analysis. Am J Theo Appl Stat 2014; 3(4):100–6.