Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

Effect of outliers on the coefficient of determination in multiple regression analysis with the application on the GPA for student

Afrah Yahya AL Rezami [1, 2, *]

[1]*Department of Mathematics, Al-Aflaj College of Science and Humanities Studies, Prince Sattam Bin Abdulaziz University, Al-Kharj, Saudi Arabia*
[2]*Department of Statistics and Information, College of Commerce and Economics, Sana'a University, Sana'a, Yemen*

ARTICLE INFO

Article history:
Received 14 December 2019
Received in revised form
10 May 2020
Accepted 11 June 2020

Keywords:
Regression
Determination coefficient
Outliers
Cumulative average

ABSTRACT

This study aims to solve the problem of contradiction between the statistical significance and real significance of regression parameters when using multiple linear regression analysis. In this regard, an algorithm was presented based on the simple and multiple of determination coefficient, and the sum of averages to estimate multiple outliers when outliers are real. Regression analysis was applied to a phenomenon, whose results are known in advance (The relationship between Semester average and Cumulative average). The results were misleading, and we cannot firmly stand on analysis results. Also, the regression model did not improve much when an increased sample size more than doubled, so the study presents an algorithm for finding a solution to this contradiction. After checking Ordinary Least Squares (OLS) assumptions, outliers were identified, based on Cook's distance because it was the best. The proposed algorithm was compared with some robust regression methods, [Weighted Least Squares, Fully Modified Least Squares, and Least Median of Squares]. The results proved that the proposed method is a robust solution for outliers' estimation. Therefore, it is recommended to use the proposed algorithm to estimate multiple outliers on other similar phenomena (e.g., The algorithm can be applied to a credit card transaction control system in a bank), and also software Packages statistical for the proposed algorithm. Also, the novelty of this study can be observed by investigating testing the significance of outliers as most of the previous researchers were interested in diagnosing the outliers without checking its significance.

© 2020 The Authors. Published by IASE. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

The Ordinary Least Squares (OLS) method is the most common way to fit the regression model, but this method fails in dealing with data that contains outliers. Therefore, one cannot firmly stand on regression analysis results because OLS is said to be not robust to violations of its assumptions. All major software packages (SAS, SPSS, R, MINITAB, and STATA) provide both the model estimates and the diagnostic of the model fit. However, the wide popularity and routine use of linear regression create some problems. The problems of multiple linear regression models arise when there is an outlier in the data. Identifying and estimating outliers is an important step in building the regression model. If outliers are identified and estimated, they will lead to a different model (Rahman et al., 2012). Sometimes, when natural phenomena are studied, an effect of one or more independent variables is insignificant, although it is known that these variables only effect on the dependent variable. For example, the balance of any person in the bank depends on only two variables (Addition and withdrawal), so the relationship between them strong and significant will be expected. Any behavior other than this expectation is due to one or several outliers. One should be worried about outliers because it can distort estimates of regression coefficients, can produce misleading results, and the interpretation of the results may be in doubt. It is possible that another researcher could analyze these data and question these results showing an improved analysis that may contradict these results and undermine the conclusions (Gad and Qura, 2016). In this regard, a new algorithm is

[*] Corresponding Author.
Email Address: a.alrezamee@psau.edu.sa

https://doi.org/10.21833/ijaas.2020.10.004
Corresponding author's ORCID profile:
https://orcid.org/0000-0003-1176-0286
2313-626X/© 2020 The Authors. Published by IASE.

This is an open access article under the CC BY-NC-ND license
(http://creativecommons.org/licenses/by-nc-nd/4.0/)

presented based on the partial and multiple correlation coefficient, coefficient of determination, and the sum of averages for predictors to estimate multiple outliers in the multiple linear regression model. One of the conditions for estimating multiple outliers is the true presence of outliers, which cannot be presented in the form of errors. The novelty of this study can be observed by investigating testing the significance of outliers as most of the previous researchers were interested in detecting and addressing the outliers, without checking its significance. The importance of research is to present a new idea for estimating outliers in independent variables and dependent variables, using an easy algorithm, to obtain the reliable model of prediction when only these variables affect the dependent variable.

1.1. Multiple linear regression models

Multiple linear regression helps to predict the values of a dependent variable by knowing the values of independent variables with statistical significance. It can be expressed in the following form (Salleh et al., 2015; Park et al., 2012);

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_n + e_n \tag{1}$$

Fit multiple linear regression model (Neter et al., 1996);

$$\hat{y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_n, \tag{2}$$

where, ŷ is Fitted response, x_{n} is independent variables; n is Number of observations; p is Number of model parameters; β_{n} is a regression coefficient; e_n is i th residual.

Estimation of Parameters with Ordinary Last Square OLS (Freund et al., 2006):

$$\beta=(x\backslash x)^{-1}x\backslash y \tag{3}$$

1.2. The goodness-of-fit (OLS) regression

R-Sq→R2 is known as the coefficient of determination. A commonly used measure of goodness of fit of a linear model can be measured as (Altland, 1999);

$$\text{Formula} \rightarrow R^2 = 1 - \frac{\text{SS Error}}{SS\,Total} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2} \tag{4}$$

$$\text{Formula} \rightarrow Adj.R^2 = 1 - \frac{\text{MS Error}}{MS\,Total} = 1 - \left(\frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2}\right)\left(\frac{n-1}{n-p-1}\right) \tag{5}$$

1.3. Unusual and Influential observations

1.3.1. Outliers

Which are extreme values in the y-direction relative to the fitted regression line, or as an observation that has a large residual (Kim, 2000; Adikaram et al., 2014; Weisberg, 2013). Rousseeuw (1987) explained how the single outlier changed from the direction of the lower squares. Huber and Ronchetti (1981) explained the effect of outliers on the OLS estimates by destroying the least squares.

1.3.2. Leverage

Extremes values are in the x-direction, which will pull the regression line towards it and can have a large effect on regression coefficients (Cerioli et al., 2013).

1.3.3. Influential observations

Influential observations can change the slope of the line, and it has a large influence on the fit of the model. On the other hand, an observation is said to be influential if removing the observation substantially changes the estimate of regression coefficients (Alguraibawi et al., 2015).

1.4. Identification of unusual observations

To identify unusual observations, the study has used diagnostic measures, which include Residuals, standardized residual, Studentized Deleted Residuals, leverage values, and Cook's D. Formulas of diagnostic measures are as following (Cook, 1977; Turkan et al., 2012);

1.4.1. Residuals

Residuals are the distance between observed values and the predicted values (Rahmatullah Imon and Ali, 2005; Richard et al., 2019). The residual is defined as

$$e_i = y_i - \hat{y}_i.$$

Studentized Deleted Residuals (Greenwell et al., 2018; Cook and Weisberg, 1982):

$$t_i = e_i\left(\frac{n-p-1}{S(1-h_i)-e_i^2}\right)^{\frac{1}{2}} \tag{6}$$

1.4.2. Cook's distance

It combines information on residual and leverage (Judd et al., 2017; Belsley et al., 1980). It identifies influential cases as it considers changes in all residuals when a case is omitted. It is calculated from the following relationship:

$$D_i = \frac{\sum_{j=1}^{n}\left(\hat{y}_j - \hat{y}_j(i)\right)^2}{(k+1)S^2} = \frac{e_i^2}{ps^2}\left[\frac{h_i}{(1-h_i)^2}\right] = \frac{\left(b-b_{(i)}\right)' x' x\left(b-b_{(i)}\right)}{ps^2} \tag{7}$$

where, $b_{(i)}$ is a coefficient vector calculated after deleting the ith observation.

DFFITS is as follows:

$$DFFITS = \frac{\hat{y}_i - \hat{y}_{i(i)}}{\sqrt{S^2{}_{(i)}h_{ii}}} = e_i\left(\frac{n-p-1}{S^2(1-h_i)-e_i^2}\right)^{1/2}\left(\frac{h_i}{1-h_i}\right)^{1/2} = t_i\left(\frac{h_i}{1-h_i}\right)^{1/2} \tag{8}$$

where, $\hat{y}_{i(i)}$ is fitted value calculated without the i^{th} observation (Srivastava and Lee, 1984).

COVRATIO is as follows:

$$\text{COVRATIO}_i = \frac{\det\left[\left(X'_{(i)}X_{(i)}\right)^{-1} S^2_{(i)}\right]}{\det\left[(X'X)^{-1}S^2\right]} = \left(\frac{1}{1-h_{ii}}\right)\left(\frac{S^2_{(i)}}{S^2}\right)^p \qquad (9)$$

where, $\det\left[\left(X'_{(i)}X_{(i)}\right)^{-1} S^2_{(i)}\right]$ is determinant of the coefficient covariance matrix with observation i. $\det\left[(X'X)^{-1}S^2\right]$ is determinant of the covariance matrix for the full model (Valliant, 2012).

2. Proposed work

Influential observations should be examined carefully both in the dependent variable and independent variables, before applying the proposed algorithm in x_i and y.

2.1. Estimating the outliers in the independent variables

If x_i is an independent variable and it is regression coefficient is not statistical significant, then independent variable contains one or multiple outliers, the algorithm will be as follows; The coefficient of determination (R^2_{yx}) is calculated in the simple linear regression, and calculating the sum of the averages of the independent variables for the same observation $(\sum_{i=1}^{p}\bar{x}_{i_m})$, using the following formula:

$$x^*_{i_m} = \sum_{i=1}^{p}\bar{x}_{i_m}\left(R^2_{yx}\ \right) \qquad (10)$$

where, $x^*_{i_m}$ is The outlier estimation; $\bar{x}_{i_m}$ is Average independent variables for the outlier (m); R^2_{yx} is the coefficient of determination in the simple linear regression.

2.2. Estimating the outliers in the dependent variable

If $\mathrm{y_i}$ is a dependent variable and this variable contains one or multiple outliers, then the algorithm will be as follows:

$$y^*_{\mathrm{j}} = \sum_{i=1}^{p}\bar{x}_{ij}\left(R^2_{yx_i}\right) \qquad (11)$$

where, y^*_{j} is Outlier estimation; $R^2_{yx_i}$ is multiple determinant coefficient; $\bar{x}_{ij}$ is average independent variables for the outlier (j).

3. Empirical results

3.1. Overview of data

The data was obtained from the academic record of the student from the Prince Sattam Bin Abdulaziz university website. Independent variables used in this study are represented as (x_i); from the semester average for the first level to the semester average for the sixth level. Dependent Variable(y): Cumulative Grade Point Average (GPA).

3.2. Fitting the regression model using (OLS) before regression diagnosis

Table 1 shows that the parameter for the third level has a probability value of less than 0.05. This result indicates that this variable has a statistically significant effect on the cumulative average. But a probability value for the other parameters indicates that there is not a statistically significant effect on the cumulative average (This contradicts reality). Although the cumulative average of the student is affected only by the semesters average, these results are misleading, so the study makes efforts to find a solution to this contradiction.

Table 1: Fitting the regression model using (OLS) before regression diagnosis

Model Summary and Coefficients				
S.E. of regression	Adjusted R^2	F-statistic	Prob. (F-statistic)	Durbin-Watson
0.42193	0.733188	27.56363	0.000	2.889955
$\boldsymbol{\beta_i}$	Coefficient	Std. Error	t-Statistic	Prob.
$\boldsymbol{\beta_0}$	0.550095	0.238249	2.308905	0.0250
$\boldsymbol{\beta_1}$	0.088475	0.070860	1.248589	0.2174
$\boldsymbol{\beta_2}$	0.109720	0.101698	1.078886	0.2856
$\boldsymbol{\beta_3}$	0.315284	0.095071	3.316299	0.0017
$\boldsymbol{\beta_4}$	0.105943	0.110328	0.960255	0.3414
$\boldsymbol{\beta_5}$	0.152190	0.097423	1.562159	0.1243
$\boldsymbol{\beta_6}$	0.060783	0.083192	0.730638	0.4683

3.3. Assumptions of the OLS estimator

Many graphical methods and numerical tests have been developed over the years for regression diagnostics (Abuzaid et al., 2011). Statistical Software makes many of these methods easy to access and use. Consider the following assumptions.

3.3.1. Linearity and multicollinearity

Checking the linearity assumption is not so straightforward in the case of multiple regression. The study has fitted the best fit line, known as the Loess Curve through the scatterplot to see if any nonlinear relationship can be detected. And to verify the absence of multiple linearities between the predictors was used Variance Inflation Factor (VIF). The values of the inflation factor should be less than 10 (Müller, 1992; Ibrahim and Yahya, 2017). From the Loess curve, it appears that the relationship of fitted value against residuals is roughly linear. And this indicates that the linearity assumption is satisfied. And also, it has been observed that the Variance Inflation Factor (VIF) is less than 10. This is evidence of the absence of multicollinearity between predictors. This is confirmed by the matrix plot. See (Fig. 1).

3.3.2. Normality and heteroscedasticity for residuals

Fig. 2 shows for probability plot that the points do not cluster around the line; this indicates that the residues are not normality. This is confirmed by the Kolmogorov-Smirnov test. Also, Levene's test clearly indicates that the residuals have a constant variance.

3.3.3. Independence residuals and stability of the regression model

Table 2 shows that the Durbin-Watson statistic (D.W=2.89955) is far from a tabulated value (D.W=1.639). We will use the Breusch-Godfrey Serial Correlation LM Test to make sure there is no autocorrelation, and this test indicates there is autocorrelation between the residues and also the model is not stability.

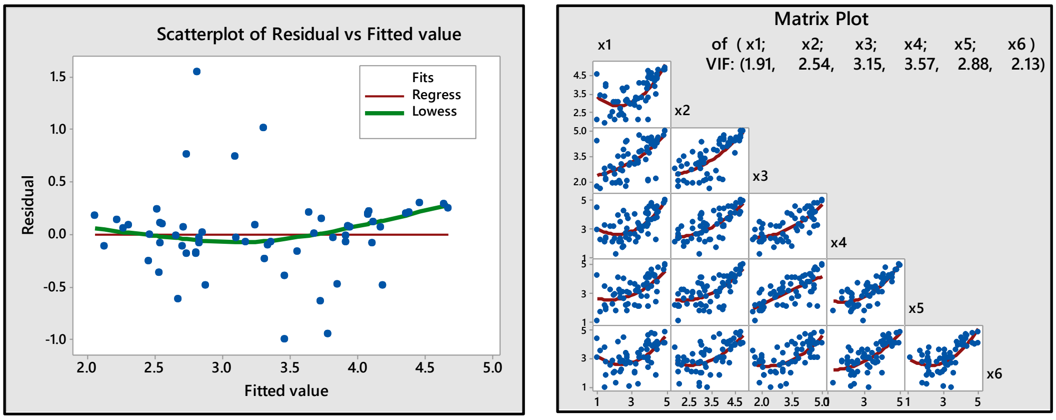


Fig. 1: Checking linearity and multicollinearity

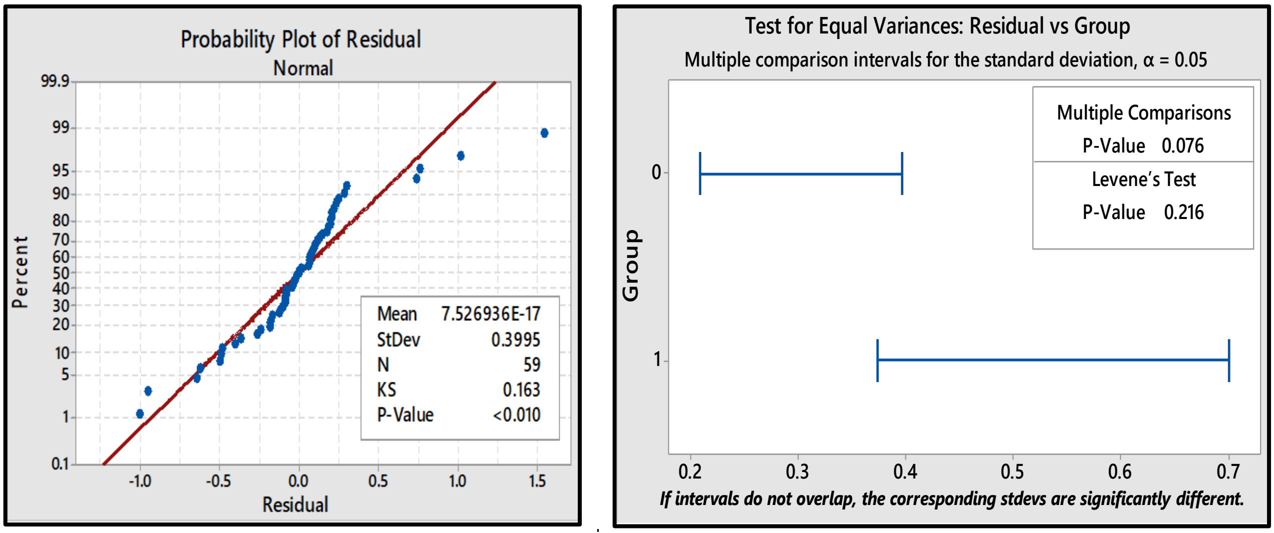
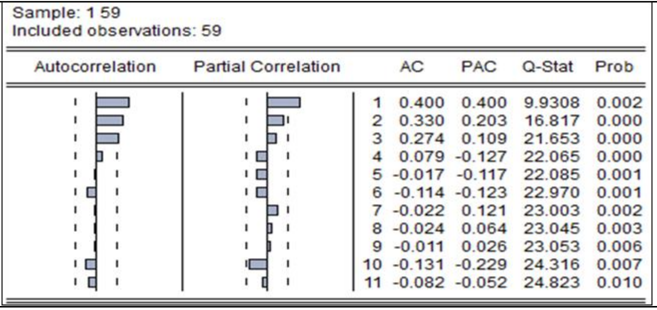


Fig. 2: Checking normality and homogeneity of variance for residuals

Table 2: Check independence residuals and the stability of the regression model

Breusch-Godfrey Serial Correlation LM Test	
F-statistic	= 11.43413
ProF(2,50)	= 0.0001
Durbin-Watson	= 2.889955



```
Sample: 1 59
Included observations: 59

Autocorrelation   Partial Correlation      AC     PAC    Q-Stat  Prob

                                       1  0.400  0.400  9.9308  0.002
                                       2  0.330  0.203  16.817  0.000
                                       3  0.274  0.109  21.653  0.000
                                       4  0.079 -0.127  22.065  0.000
                                       5 -0.017 -0.117  22.085  0.001
                                       6 -0.114 -0.123  22.970  0.001
                                       7 -0.022  0.121  23.003  0.002
                                       8 -0.024  0.064  23.045  0.003
                                       9 -0.011  0.026  23.053  0.006
                                      10 -0.131 -0.229  24.316  0.007
                                      11 -0.082 -0.052  24.823  0.010
```

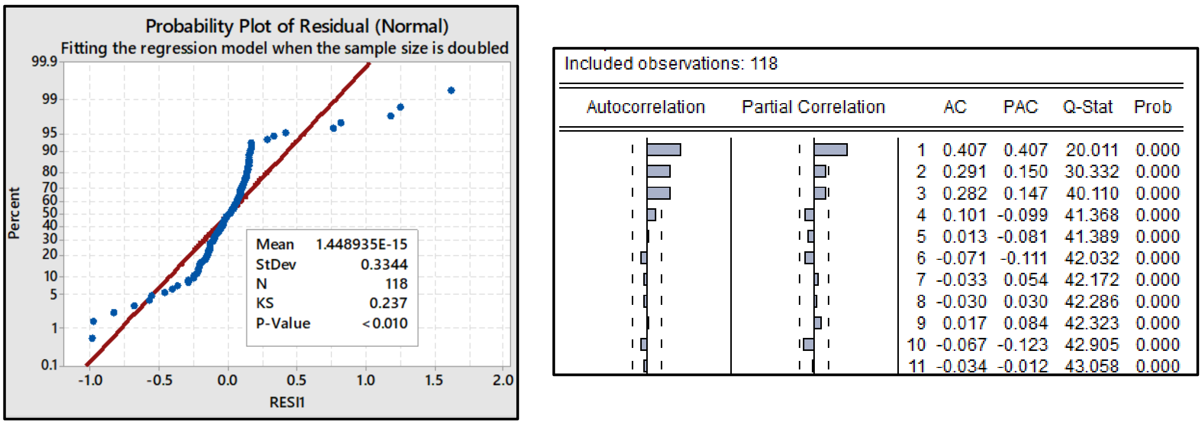
From Tables 1-2 and Figs. 1-2 the above, we conclude that four assumptions have not achieved (Statistical significance for five regression coefficients, normality, the independence of residuals, and the stability of the model).

3.4. Fitting the regression model when the sample size is doubled

Table 3 shows that the regression model did not improve much when increasing the size of the sample, where still a parameter 4 and a parameter 6 are not statistically significant. Also, there is autocorrelation between the residues. The resulting probability plot shows in Fig. 3 that the points no cluster around the line, this indicates that the residues are not distributed according to normal distribution. This is confirmed by Jarque-Bera and Kolmogorov-Smirnov test, and also the regression model hasn't stability. After fitting the regression model, when the sample size is doubled.

Table 3: Fitting the regression model when the sample size is doubled

Model Summary and Coefficients				
S.E. of regression	Adjusted R^2	F-statistic	Prob. (F-statistic)	Durbin-Watson
0.34327	0.815929	82.00458	0.0000	2.879763
β_i	Coefficient	Std. Error	t-Statistic	Prob.
β_0	0.328262	0.148356	2.212660	0.0290
β_1	0.155687	0.046806	3.326223	0.0012
β_2	0.131489	0.065369	2.011500	0.0467
β_3	0.228933	0.059941	3.819301	0.0002
β_4	0.125195	0.063184	1.981433	0.0500
β_5	0.188672	0.059169	3.188693	0.0019
β_6	0.077556	0.052567	1.475382	0.1429



Included observations: 118

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.407	0.407	20.011	0.000
		2	0.291	0.150	30.332	0.000
		3	0.282	0.147	40.110	0.000
		4	0.101	-0.099	41.368	0.000
		5	0.013	-0.081	41.389	0.000
		6	-0.071	-0.111	42.032	0.000
		7	-0.033	0.054	42.172	0.000
		8	-0.030	0.030	42.286	0.000
		9	0.017	0.084	42.323	0.000
		10	-0.067	-0.123	42.905	0.000
		11	-0.034	-0.012	43.058	0.000

Breusch-Godfrey Serial Correlation LM Test=Prob. F(2.109)=0.0000; Normality Test: Jarque-bera =302.0571 (Prob.=0.0000

Fig. 3: Checking assumptions (OLS)

3.5. Diagnosis of outliers

In commencing, one should get familiar with the data file and looking for errors to collect and input data using the Moses test (Nussbaum, 2014). Table 4 shows that there are no errors in data collection.

Table 4: Identifying outliers using Moses test

Test Statistics (Moses Test)		
	Observed Control (Sig)	Trimmed Control (Sig)
y	1.000	1.000
x_1	1.000	0.940
x_2	1.000	1.000
x_3	1.000	0.940
x_4	1.000	0.813
x_5	0.746	0.940
x_6	0.293	0.648

3.6. Identifying outliers using the residuals

The goal is to detect the cases which have large residuals (outliers) and the cases that, if they are removed, lead to a different result. The distinction between these two kinds of cases is not always obvious. Both types of points are of great concern. (Choonpradub and McNeil, 2005). Not necessarily that all outliers are influential. In this regard, a box plot will be used by overall measures of influence (DFFITS, COVRATIO, and Cook's D) to discover influential cases.

Fig. 4 shows that Star-shaped states are influential, while circle cases are not influential (For example, Cook's Distance indicates that case number 32 is one that has a large residual, which suggests that it may be influential, and the observations (37, 34, 33,36,6,41, 29) are outliers, but cases (32, 37,36,34,33,6) is are an influential case. And these cases require more attention as they stand out from all other points.

3.7. Significance test of outliers

Fig. 5 shows that the cases diagnosed as outliers through the Grubbs' test had a significant effect on the regression coefficients. But cases that have been diagnosed as outliers through the Dixon's test had not any effect.

4. Application proposed work

4.1. Application of the proposed algorithm

Application of the proposed algorithm according to overall measures of influence. Cook distance was relied on to identify influential outliers because it is the best (Table 5).

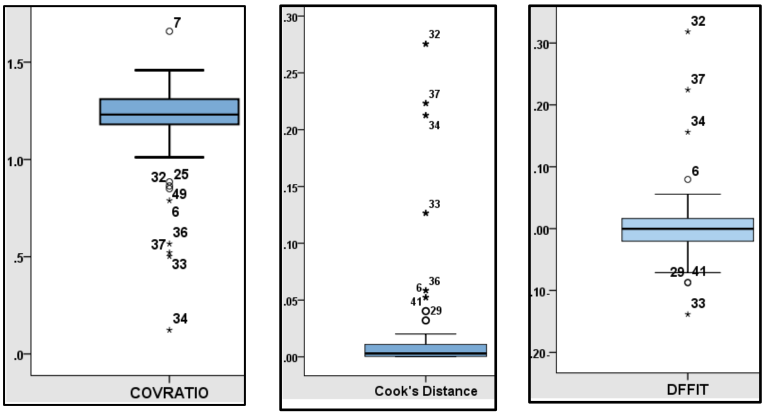


Fig. 4: Box plot for overall measures to identify an influential case in y

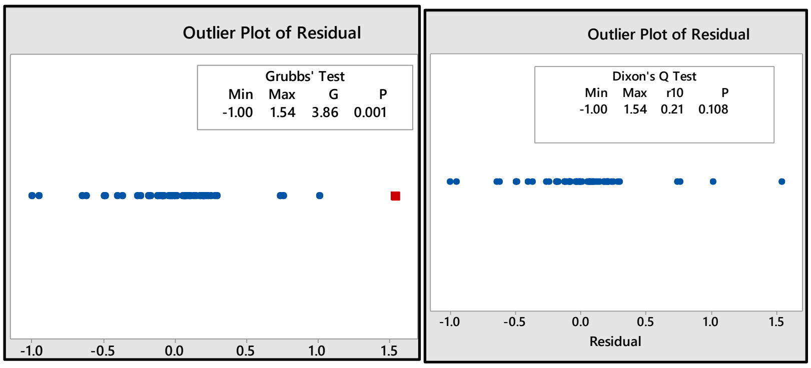


Fig. 5: Significance test of outliers

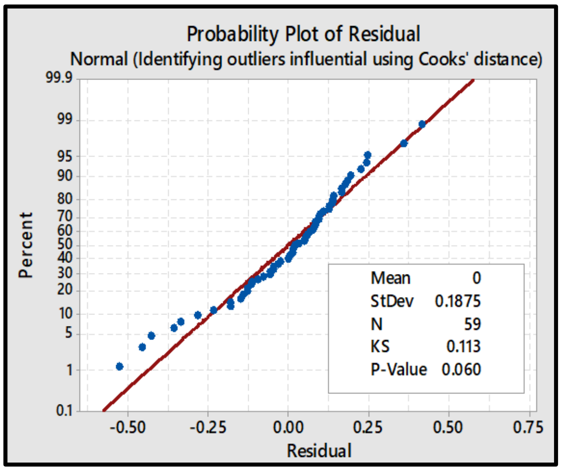
Table 5: Comparison between overall measures of influence

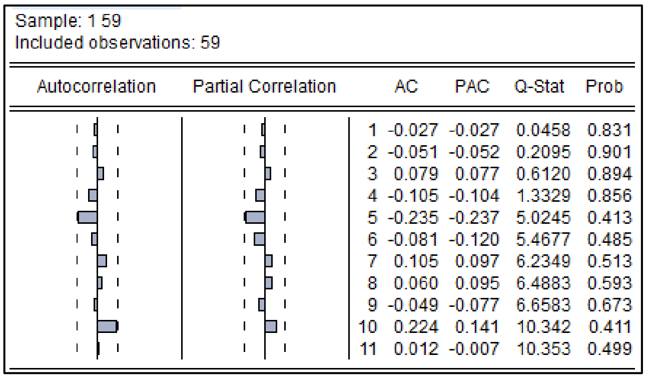
OVRATIO				Cook's distance				DFFITS			
S.E	Norm.	R^2	D-W	S.E	Norm.	R^2	D-W	S.E	Norm.	R^2	D-W
.264	. .036	.931	2.09	.198	0.062	0.94	1.96	.264	0.00	0.894	1.82
Heteroskedasticity Test ARCH			= 0.79	Heteroskedasticity Test ARCH			= 0.835	Heteroskedasticity Test ARCH			= 0.808
Breusch-Godfrey Serial Correlation			= 0.42	Breusch-Godfrey Serial Correlation			= 0.884	Breusch-Godfrey Serial Correlation			= 0.636

4.2. Fitting the regression model using the proposed algorithm

The resulting probability plot shows in Fig. 6 that the points cluster around the line; this indicates that the residues are distributed according to normal distribution.

This is confirmed Kolmogorov-Smirnov and Jarque-Bera test. And we observe there is no autocorrelation between the residues, and also regression model has stability. Formula (10 and 11) was used to estimate regression parameters. The result indicates in Table 6 that all the variables have a statistically significant effect on the cumulative average (This no contradicts reality).





Sample: 1 59
Included observations: 59

Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.027	-0.027	0.0458	0.831
		2	-0.051	-0.052	0.2095	0.901
		3	0.079	0.077	0.6120	0.894
		4	-0.105	-0.104	1.3329	0.856
		5	-0.235	-0.237	5.0245	0.413
		6	-0.081	-0.120	5.4677	0.485
		7	0.105	0.097	6.2349	0.513
		8	0.060	0.095	6.4883	0.593
		9	-0.049	-0.077	6.6583	0.673
		10	0.224	0.141	10.342	0.411
		11	0.012	-0.007	10.353	0.499

Fig. 6: Checking assumptions (OLS) after using the proposed algorithm

Table 6: Fitting the regression model using the proposed algorithm

Model Summary and Coefficients				
S.E. of regression	Adjusted R^2	F-statistic	Prob. (F-statistic)	Durbin-Watson
0.198	0.9402	153.0416	0.000	1.956232
$\boldsymbol{\beta_i}$	Coefficient	Std. Error	t-Statistic	Prob.
$\boldsymbol{\beta_0}$	0.248595	0.103724	2.396707	0.0202
$\boldsymbol{\beta_1}$	0.103035	0.035985	2.863261	0.0060
$\boldsymbol{\beta_2}$	0.123082	0.047601	2.585721	0.0126
$\boldsymbol{\beta_3}$	0.193795	0.042387	4.572048	0.0000
$\boldsymbol{\beta_4}$	0.252619	0.051842	4.872848	0.0000
$\boldsymbol{\beta_5}$	0.104303	0.049882	2.090995	0.0414
$\boldsymbol{\beta_6}$	0.140472	0.046339	3.031399	0.0038

4.3. Comparison of the proposed algorithm with some Robust Regression methods

The proposed algorithm will be compared with some robust regression methods, [Weighted Least Squares (WLS), Fully Modified Least Squares (FMOLS), and Least Median of Squares (LMS)]. From Table 7.

The statistically significant was achieved for all regression parameters, and the normality hypothesis for residues was achieved by using the proposed method only. In addition, the proposed method has the highest coefficient of determination was (Adj R^2=0.9402) and the lest standard error (S.E.= 0.198).

Table 7: Comparison of the proposed algorithm with some Robust Regression methods

	Proposed Method		WLS		FMOLS		LMS	
β_i	Prob.	Parameters Accuracy	Prob.	Parameters Accuracy	Prob.	Parameters Accuracy	Prob.	Parameters Accuracy
β_1	0.0202	Adj. R^2= 0.9402	0.001	Adj. R^2= 0.937	0.1332	Adj. R^2= 0.730791	0.0695	Adj. R^2= 0.638659
β_2	0.0060	S.E. =0.198	0.127	S.E.=1.3152	0.1192	S.E.=0.4273	0.0008	S.E.= 0.4376
β_3	0.0126		0.000		0.0000		0.1257	
β_4	0.0000	Norm=0.06	0.016	Norm=0.010	0.1691	Norm= 0.00	0.0084	Norm=0.000
β_5	0.0000	Stability	0.033	No Stability	0.0077	No Stability	0.1213	No Stability
β_6	0.0414		0.005		0.0769		0.0388	

5. Conclusion

The study has used MINITAP, SPSS, and EVIEWS to perform the computations. All methods of estimation were compared using three standards [The significant of regression parameters, adjusted determination coefficient (Adj.R^2), Standard Error (S.E.) of regression and achieve the assumptions of OLS]. Regression analysis was applied to a phenomenon, whose results are known in advance (The relationship between Semester average and Cumulative average). Since there is no method could correctly treat outliers 100%.

The results of this study proved that the proposed method is a robust solution for outliers estimation. Most importantly, the method is a solution for estimating multiple significant outliers in the data set. The study has found that the proposed algorithm performs well to obtain highly efficient estimates of regression coefficients. The proposed method can be applied on others similar phenomena (e.g., The proposed method can be applied to a credit card transaction control system in a bank, which aims to detect fraud, to detect unusual purchases, as an outlier, compared to the normal behavior of the customer of the cardholder. Another example delays in the delivery of orders to homes, such as when there is a delay in the delivery of 20 orders in one day, and therefore restaurant management can use the algorithm to solve the problem). Also, the novelty of this study can be observed by investigating testing the significance of outliers as most of the previous researchers were interested in detecting and addressing the outliers, without checking its significance. The research also found that the cause of the existence of outliers in the data was errors in the university website, and also the data has a high torsion to the right. Therefore, it is recommended to using the proposed algorithm to estimate multiple outliers on any phenomenon, whose results are known in advance, and also doing designing software Packages statistical for the proposed algorithm.

Acknowledgment

This project was supported by the Deanship of Scientific Research at Prince Sattam Bin Abdulaziz University under the research project 2017/01/8307.

Compliance with ethical standards

Conflict of interest

The authors declare that they have no conflict of interest.

References

Abuzaid A, Hussin AG, and Rambli A (2011). COVRATIO statistic for simple circular regression model. Chiang Mai Journal of Science, 38(3): 321-330.

Adikaram KKLB, Hussein MA, Effenberger M, and Becker T (2014). Outlier detection method in linear regression based on sum of arithmetic progression. The Scientific World Journal, 2014: 821623.
https://doi.org/10.1155/2014/821623
PMid:25121139 PMCid:PMC4121229

Alguraibawi M, Midi H, and Imon AHM (2015). A new robust diagnostic plot for classifying good and bad high leverage points in a multiple linear regression model. Mathematical Problems in Engineering, 2015: 279472. https://doi.org/10.1155/2015/279472

Altland HW (1999). Regression analysis: Statistical modeling of a response variable. Technometrics, 41(4): 367-368. https://doi.org/10.1080/00401706.1999.10485936

Belsley DA, Kuh E, and Welsch RE (1980). Regression diagnostics: Identifying influential data and sources of collinearity. John Wiley and Sons, New York, USA. https://doi.org/10.1002/0471725153

Cerioli A, Riani M, and Torti F (2013). Size and power of multivariate outlier detection rules. In: Lausen B, Van den Poel D, and Ultsch A (Eds.), Algorithms from and for nature and life: 3-17. Springer, Cham, Switzerland. https://doi.org/10.1007/978-3-319-00035-0_1

Choonpradub C and McNeil D (2005). Can the box plot be improved? Songklanakarin Journal of Science and Technology, 27(3): 649-657.

Cook RD (1977). Detection of influential observation in linear regression. Technometrics, 19(1): 15-18. https://doi.org/10.1080/00401706.1977.10489493

Cook RD and Weisberg S (1982). Residuals and influence in regression. Chapman and Hall, New York, USA.

Freund RJ, Wilson WJ, and Sa P (2006). Regression analysis statistical modeling of a response. Elsevier, Edinburgh, London, UK.

Gad AM and Qura ME (2016). Regression estimation in presence of outliers: A comparative study. International Journal of Probability and Statistics, 5(3): 65-72.

Greenwell BM, McCarthy AJ, Boehmke BC, and Liu D (2018). Residuals and diagnostics for binary and ordinal regression models: An introduction to the sure package. The R Journal, 10(1): 381-394. https://doi.org/10.32614/RJ-2018-004

Huber PJ and Ronchetti EM (1981). Robust statistics. John Wiley and Sons, New York, USA. https://doi.org/10.1002/0471725250

Ibrahim SA and Yahya WB (2017). Effects of outliers and multicollinearity on some estimators of linear regression model. Nigeria Statistical Society, 1: 204-209.

Judd CM, McClelland GH, and Ryan CS (2017). Data analysis: A model comparison approach to regression, ANOVA, and beyond. Routledge, Abingdon, UK. https://doi.org/10.4324/9781315744131

Kim JT (2000). An order selection criterion for testing goodness of fit. Journal of the American Statistical Association, 95(451): 829-835. https://doi.org/10.1080/01621459.2000.10474274

Müller HG (1992). Goodness-of-fit diagnostics for regression models. Scandinavian Journal of Statistics, 19: 157-172.

Neter J, Kutner M, Nachtsheim C, and Wasserman W (1996). Applied linear regression models. 3rd Edition, Irwin, Chicago, USA.

Nussbaum EM (2014). Categorical and nonparametric data analysis: Choosing the best statistical technique. Routledge, Abingdon, UK. https://doi.org/10.4324/9780203122860

Park CG, Kim I, and Lee YS (2012). Error variance estimation via least squares for small sample nonparametric regression. Journal of Statistical Planning and Inference, 142(8): 2369-2385. https://doi.org/10.1016/j.jspi.2012.02.050

Rahman SK, Sathik MM, and Kannan KS (2012). Multiple linear regression models in outlier detection. International Journal of Research in Computer Science, 2(2): 23-28. https://doi.org/10.7815/ijorcs.22.2012.018

Rahmatullah Imon AHM and Ali MM (2005). Simultaneous identification of multiple outliers and high leverage points in linear regression. Journal of the Korean Data and Information Science Society, 16(2): 429-444.

Richard F, Gunst M, and Robert L (2019). Regression analysis and its application: A data-oriented Approach. CRC Press, Boca Raton, USA. https://doi.org/10.1201/9780203741054

Rousseeuw PJ and Leroy AM (1987). Robust regression and outlier detection. John Wiley and Sons, New York, USA. https://doi.org/10.1002/0471725382

Salleh FHM, Arif SM, Zainudin S, and Firdaus-Raih M (2015). Reconstructing gene regulatory networks from knock-out data using Gaussian noise model and Pearson correlation coefficient. Computational Biology and Chemistry, 59: 3-14. https://doi.org/10.1016/j.compbiolchem.2015.04.012 **PMid:26278974**

Srivastava MS and Lee GC (1984). On the distribution of the correlation coefficient when sampling from a mixture of two bivariate normal densities: Robustness and the effect of outliers. Canadian Journal of Statistics, 12(2): 119-133. https://doi.org/10.2307/3315176

Turkan S, Meral CC, and Oniz T (2012). Outlier detection by regression diagnostics based on robust parameter estimates. Hacettepe Journal of Mathematics and Statistics, 41(1): 147-155.

Valliant R (2012). Regression diagnostics in survey data. Joint Program in Survey Methodology, University of Maryland and University of Michigan, USA.

Weisberg S (2013). Applied linear regression. 4th Edition, Wiley, Hoboken, USA.