Journal of Statistics Applications & Probability
An International Journal

http://dx.doi.org/10.18576/jsap/100206

Using Correlation Coefficient to Solve Outliers Problem in Regression Analysis, with Practical Application

Afrah Yahya AL-Rezami [1,2,*]

[1]Department of Mathematics, Al-Aflaj College of Science and Humanities Studies , Prince Sattam Bin Abdulaziz University, Al-Aflaj 710-11912, Saudi Arabia.
[2]Department of Statistics and Information, College of Commerce and Economics, Sana'a University, Yemen.

Received: 23 Oct. 2019, Revised: 12 Dec. 2019, Accepted: 23 Jan. 2020
Published online: 1 Jul. 2021

Abstract: A new algorithm is presented on the basis of the partial and multiple correlation coefficient 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. Regression analysis was applied to a phenomenon, whose results are known in advance (The relationship between Semester GPA and Cumulative GPA). The results were misleading . After checking Ordinary Least Squares (OLS)) assumptions, outliers were identified by scatter plot for the standardized predicted values against Standardized residual, Studentized deleted residual, Cook's D, and Hit Matrix. Influential cases were identified using box plot for overall influence measures (DFFITS, COVRATIO, and Cook's D.). Thereafter, outliers are estimated using the proposed algorithm, which is compared with OLS before discovery outliers, trimmed mean, and weighted least squares (WLS). These methods were compared based on [(P-Value for i), (Adjusted R^2), and assumptions of OLS]. The results proved that the proposed method is a robust solution for outliers estimation. Thus, it is recommended to use the proposed algorithm to estimate multiple outliers for any other similar phenomenon. (For example, the proposed method can be applied to a credit card transaction control system in a bank).

Keywords: Correlation coefficient, Influence measures, Real outliers, Regression analysis.

1 Introduction

The Ordinary Least Squares (OLS) method is the most common way to fit the regression model, but it cannot deal with data that contains of outliers. Therefore, one we cannot firmly stand on regression analysis results because OLS is said make no sense its assumptions. All major software packages (SAS, SPSS, R, MINITAB and STATA) provide both the model estimates and the diagnosis of the model fit. However, the wide popularity for the linear regression creates some problems. The problems of multiple linear regression models arise when there is an outlier in the data. Identifying and estimating outliers are an importantsteps in building the regression model. If outliers are identified and estimated, they will lead to a different model [1].

Sometimes, when natural phenomena are studied, the effect of one or more independent variables is insignificant. However, it is known that these variables only affect 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 is strong . 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, and produce misleading results. 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 [2].

In this regard, a new algorithm is presented based on the partial and multiple correlation coefficient 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 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 the research is to present a new idea for estimating outliers in independent

* Corresponding author e-mail: a.alrezamee@psau.edu.sa

variables and dependent variable using an easy algorithm to obtain the reliable model of prediction, when only these variables affect the dependent variable.

Multiple linear regression models

Multiple linear regression helps predict the values of a dependent variable,by identifying the values of independent variables with statistical significance. It can be expressed in the following form [3,4]:

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

Fit multiple linear regression model [5];

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

Where:
$\hat{y}$: Fitted response.
x_n: Independent variables.
n: Number of observations.
p: Number of model parameters.
β_n: Regression coefficient.
e_n : i^{th} residual
Estimation of Parameters with OLS models [6,7]:

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

The goodness-of-fit (OLS) regression [8]
$R-S_q \rightarrow R^2$ is known as the coefficient of determination. A commonly used measure of goodness of fit of a linear model can be measured as;

$$\text{Formula} \rightarrow R^2 = 1 - \frac{SSError}{SSTotal} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2}. \tag{4}$$

Where: $\bar{y}$: mean response.

$$\text{Formula} \rightarrow Adj.R^2 = 1 - \frac{MSError}{MSTotal} = 1 - (\frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2})(\frac{n-1}{n-p-1}). \tag{5}$$

$$\text{Formula} \rightarrow PredR^2 = 1 - \frac{PRESS}{SSTotal} = 1 - \left(\frac{\sum_1^n (\frac{e_i}{1-h_i})^2}{\sum_1^n (y_i - \bar{y}_i)^2}\right). \tag{6}$$

Where:
h_i: i^{th} diagonal element of $x(x`x)^{-1}x`$.

Unusual and Influential observations

Some definitions are presented to be reviewed.

Outliers

However, young and educated people tend to select their wives from different tribes, groups and clans. The groom or his family should present a dowry (bride-price), which usually consists of jewelry or any valuable thing, to the bride. This is an obligatory duty for the groom or his family towards the bride according to Islamic teachings. No marriage would be considered legal without this dowry, regardless of its value. The husband, according to Islamic teachings, is in charge of his family and must be the one who assumes the family's financial burdens even if his wife is employed or rich unless she makes concessions. For Muslims, Sharia remains largely un-codified, allowing for plural interpretations.

Extreme values are in the y-direction relative to the fitted regression line, or as an observation that has a large residual [9,10,11]. Rousseeew [12] explained how the single outlier changed from the direction of the lower squares. Huber [13] explained the effect of outliers on the OLS estimates by destroying the least squares.

Leverage

Extremes values are in the x- direction, or as an observation with high leverage. These values will pull the regression line towards it and can have a large effect on regression coefficients [14].

Influential observations
Influential observations can change the slope of the line , nd extensively affect 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 [15,16].

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 [17,18]. Formulas of diagnostic measures are , as follows:

Residuals are the distance between observed values and the predicted values [19,20].
The residual is defined as: $e_i = y_i - y_i$

Standardized residual (ZRESID) [21]

$$r_i = \frac{e_i}{\sqrt{S^2(1-h_i)}}. \tag{7}$$

Studentized Deleted Residuals [22]

$$t_i = e_i\left(\frac{n-p-1}{S(1-h_i)-e_i^2}\right)^{\frac{1}{2}}. \tag{8}$$

Leverages values (hi) of the ith observation as [23]:

$$h_i = \frac{1}{n} + \frac{(x_i-\bar{x})^2}{(n-1)S_x^2}. \tag{9}$$

Cook's distance
It combines information on the residual and leverage [24]. It identifies influential cases as it considers changes in all residuals when a case is omitted. It is calculated from the following relationship:

$$D_i = \frac{\sum_{j=1}^{n}(\hat{y}_j - \hat{y}_{j(i)})^2}{(k+1)S^2} = \frac{e_i^2}{pS^2}[\frac{h_i}{(1-h_i)^3}] = \frac{(b-b_{(i)})'x'x(b-b_{(i)})}{pS^2}. \tag{10}$$

Where
$b_{(i)}$: coefficient vector calculated after deleting the ith observation.

DFFITS [25];

$$DFFITS = \frac{\hat{y}_i - \hat{y}_{i(i)}}{\sqrt{S_{(i)}^2 h_{ii}}} = e_i(\frac{n-p-1}{S^2(1-h_i)-e_i^2})^{\frac{1}{2}}(\frac{h_i}{i-h_i})^{\frac{1}{2}} = t_i(\frac{h_i}{i-h_i})^{\frac{1}{2}} \tag{11}$$

Where $y_{i(i)}$: Fitted value calculated without the ith observation

COVRATIO [26]

$$COVRATIO_i = \frac{\det[(X_{(i)}^{'}X_{(i)})^{-1}S_{(i)}^2]}{\det[(X'X)^{-1}S^2]} = (\frac{1}{1-h_{ii}})(\frac{S_{(i)}^2}{S^2})^p \tag{12}$$

Where:
$\det[(X_{(i)}^{'}X_{(i)})^{-1}S_{(i)}^2]$: Determinant of the coefficient covariance matrix with observation i.
$\det[(X^{'}X)^{-1}S^2]$: Determinant of the covariance matrix for the full model.

DFBETAS

$$DFBETAS \rightarrow = \frac{\beta_k - \beta_{k(i)}}{\sqrt{S^2_{(i)} C_{kk}}} \tag{13}$$

Where:
C_{kk}: is the j^{th} diagonal element of $(x^{\backslash}x)^{-1}$.
$\beta_{k(i)}$: Regression coefficient computed without using j^{th} observation.

2 Proposed Work

In this section, the researcher has investigated the proposed method using a real data set. Influential observations should be examined carefully both in the dependent variable and independent variables. An algorithm was suggested to estimate the influential outliers in x_i, depending on the partial correlation coefficient between x_i and y, and the total mean for independent variables. Then, it can estimate the influential outliers in y, depending on the multiple correlation coefficient between x_i and y.

Estimating the outliers in the independent variables

If x_i is an independent variable; regression coefficient is insignificant and an independent variable contains one or multiple outliers, then the algorithm will be as follows: The partial correlation coefficient (R_{yx_i}) is calculated in a simple linear regression for the variable that contains an outlier observation. Calculating the sum of the averages of the independent variables for the same observation $\sum_{i=1}^{n} \bar{x}_{i_m}$, adopts the following formula:

$$x^*_{i_m} = \sum_{i=1}^{p} \bar{x}_{i_m}(R_{yx_m}) \tag{14}$$

Where:
$x^*_{i_m}$: Estimating outlier.
m: Outlier observed.
$\bar{x}_{i_m}$: Average independent variables for outlier (m).

Estimating the outliers in the dependent variable

If y_i is a dependent variable which contains one or multiple outliers, then the algorithm will be as follows:

$$y^*_j = \sum_{i=1}^{p} \bar{x}_{ij}(R_{yx_i}) \tag{15}$$

Where:
y^*_j: Outlier estimation.
R_{yx_i}: Multiple correlation coefficient.
$\bar{x}_{ij}$: Average independent variables for outlier (j).

3 Empirical Results

Data independent variables used in this study are represented as (x_i);
The semester average for level 1 (x_1)
The semester average for level 2 (x_2)
The semester average for level 3 (x_3)
The semester average for level 4 (x_4)
The semester average for level 5 (x_5)
The semester average for level 6 (x_6)
Dependent Variable(y): Cumulative Grade Point Average -GPA.
The Minitab program outputs are given below (Table 1).

Table 1: Descriptive Statistics for data

Var.	Mean	St. Dev	Var.	Min	Max	Trimmed Mean	M- Est
y	3.290	0.817	0.667	2.00	4.92	3.2702	3.2417
x_1	3.159	1.080	1.167	1.00	4.95	3.1767	3.2228
x_2	3.440	0.869	0.754	1.88	4.91	3.4437	3.4601
x_3	3.379	1.034	1.069	1.53	5.00	3.3914	3.4306
x_4	3.251	0.948	0.899	1.23	5.00	3.2534	3.2346
x_5	3.166	0.965	0.931	1.05	5.00	3.1621	3.1554
x_6	3.149	0.972	0.944	1.00	4.85	3.1710	3.1894

Through descriptive statistics, the absence of unusual values is observed. The smallest value was 1 and the highest value was 5, indicating that there were no errors in data collection or input. Trimmed mean shows smaller or larger means compared to the real mean. The difference difference between real mean and trimmed mean indicates the distortion in data due the presence of outliers. The M-estimator tells us about mean which is not affected by outliers.

Table 2: Fitting the regression model using (OLS) before regression diagnosis

Model Summary				
S	**R-sq**	**R-sq(adj**	**R-sq(pred)**	
0.422	76.08%	73.32%	68.80%	
Model Summary and Coefficients				
Term	**Coef (β_i)**	**SE Coef**	**T-Value**	**P-Value**
Constant	0.550	0.238	2.31	.025
x_1	0.089	0.071	1.25	.217
x_2	0.110	0.102	1.08	.286
x_3	0.315	0.095	3.32	.002
x_4	0.106	0.110	0.96	.341
x_5	0.152	0.097	1.56	.124
x_6	0.061	0.083	0.73	.468
Regression Equation:				
$\mathbf{y = 0.55 + 0.089 x_1 + 0.11 x_2 + 0.315 x_3 + 0.106 x_4 + 0.152 x_5 + 0.061 x_6}$				

Table (2) shows that the Semester average for the third level has a p-value less than the 0.05. This result indicates that this variable has a statistically significant effect on the cumulative average.However, the p-value for the other semesters averages indicates that there is not a statistically significant effect on the cumulative average. Although, the Cumulative average of the student is affected only by the Semesters averages, the value for proportion of total variation explained by regression ($Adj.R2 = 73.32\%$) was medium. Consequently, these results are misleading. This makes the study try to find a solution to this contradiction.

Assumptions of the OLS estimator

Many graphical methods and numerical tests have been developed over the years for regression diagnostics [27]. Statistical software facilitates accessibility to many of these methods . To fully check the assumptions of the regression, a normal P-P (probability plot), and a scatterplot. , consider the following assumptions:

Linearity

The relationships between the predictors and the response variable should be linear. Checking the linearity assumption is not so straightforward in the case of multiple regression. The most straightforward thing to do is to plot the dependent variable against each of the independent variables. Next, the study fitted the best fit line, the Loess curve to see if any nonlinear relationship could be detected. A scatterplot is a good means to identify how well a straight line fits the data (Figure 1).

Multicollinearity

Severe multicollinearity is problematic because it can increase the variance of the regression coefficients, and so make them unstable. To verify the absence of multiple linearity, Variance Inflation Factor (VIF) was used. The values of the inflation factor should be less than 10 (Figure 2).

Figure 2,shows that the Variance Inflation Factor (VIF) is less than 10. This is an evidence of the absence of multiple linear correlations between independent variables. This is also confirmed by the matrix plot.

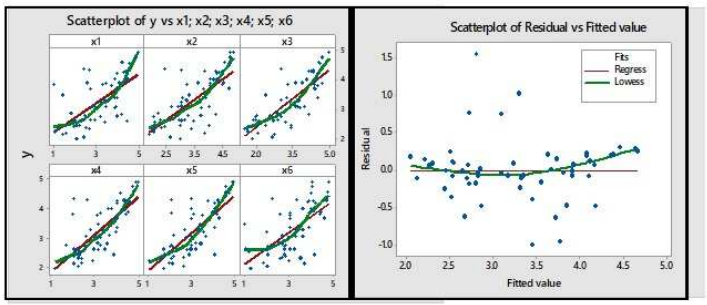


Figure (1)..

Fig. 1: Checking Linearity supposition.

From the loess curve, it appears that the relationship of fitted value against residuals is roughly linear around zero. It is estimated that the linearity assumption is satisfied.

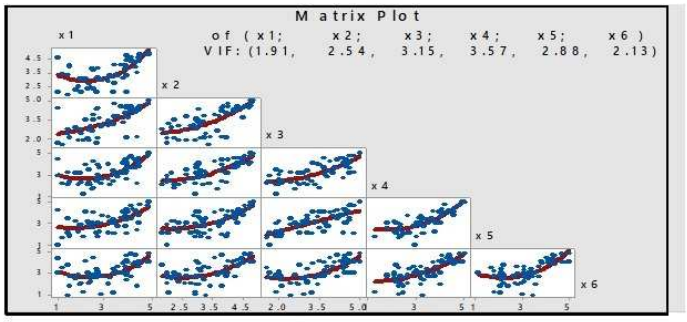


Figure (2)

Fig. 2: Checking multicollinearity suppositions

Homogeneity of variance

The residuals variance should be constant. The study has used the Levenes test and scatter plot for fitted values against residues [28,29] (Figure 3).

Figure 3 shows that there are points around zero, which is scattered uniformly. There is a clear indication that the residuals are homogeneous. This is confirmed by the Levenes test.

Normality

One of the common aspects to determine the normality of the data is the acceptance of the data through normal distribution. In this regard, a probability plot of residuals and the histogram will be used to identify the normality of the

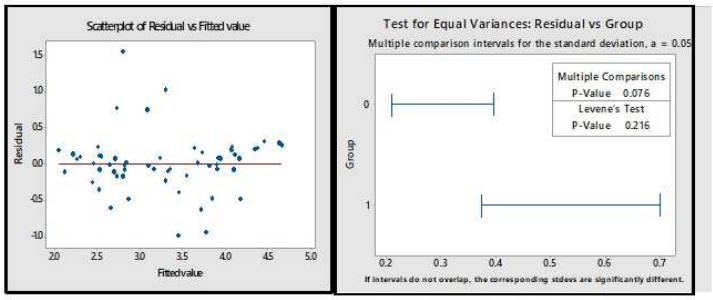


Figure (3)

Fig. 3: Checking Homogeneity of variance for residuals

given data. However, a statistical test is preferred, which is not entirely relied on graphs when testing the normality [30] (Figure 4).

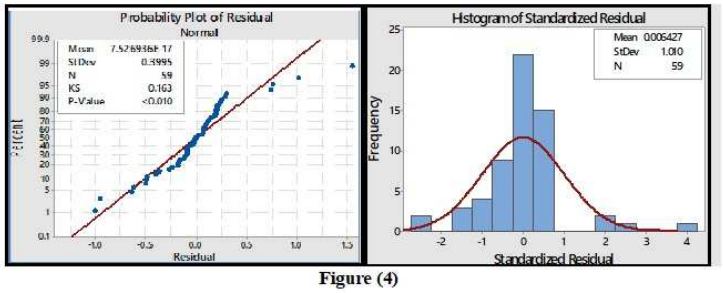


Figure (4)

Fig. 4: Checking Normality of The residuals .

The resulting probability plot shows that the points do not cluster around the line . This indicates that the residues are not distributed according to normal distribution. The histogram of the residuals indicates that some highly extreme residuals are worthy of investigation, where extreme values can be observed at the tail end of the distribution from right and left. This is confirmed by the Kolmogorov-Smirnov Test ($KS = 0.010 < 0.05$).

Independence

If the residuals are randomly distributed around zero, it means that there is no autocorrelation.Also Durbin-Watson statistic was used where a rule of values of$1.5 < D - W < 2.5$ indicates that there is no autocorrelation [31] (Figure 5).

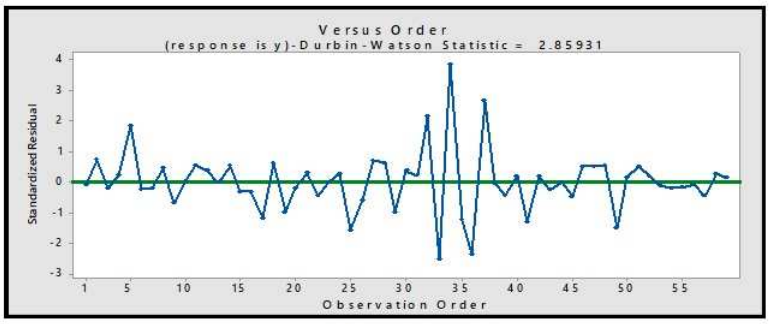


Figure (5)

Fig. 5: Checking Independence of residuals

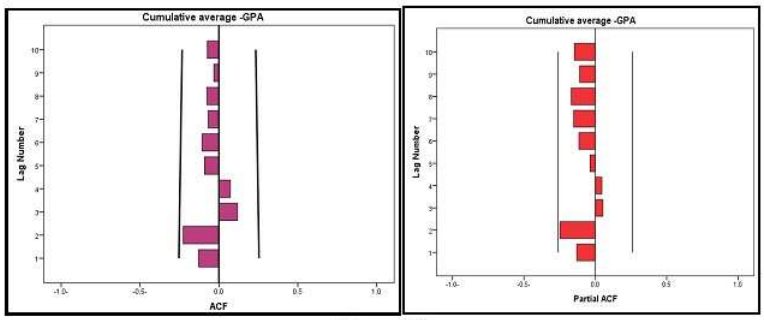


Figure (6)

Fig. 6: The auto and partial correlation function for stability model

The data are moving away from zero, especially in the middle, indicating that there is auto-correlation in the data. The value for D-W) shows that there is autocorrelation.

Checking the Stability of the Regression Model

The auto and partial correlation function were used for checking the stability of the regression model (Figure 6). The auto and partial correlation of the estimated model lies column within the area of confidence. This means that the model is stationary. Accordingly, the model has achieved all the assumptions of (OLS), except for two hypotheses (normality and independence of residuals). Thus, it should be ensured that the data must be entered correctly.

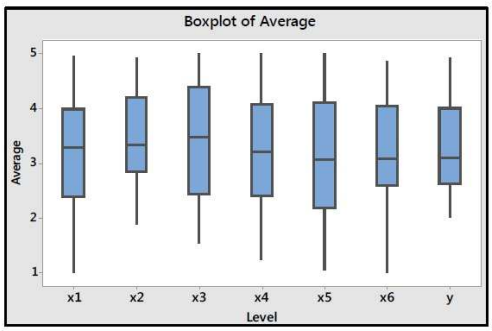
Diagnosis of Outliers

In the beginning, , one should get familiar with the data file and look for errors to collect and input data using box plot and Moses test [32] (Table 3). The box plot shows that there are no outliers in the data as confirmed by Moses' test. This indicates that there are no errors in data collection. Also, it shows that the median is not in the middle of the data for most variables. This indicates that there is a Skewness in the distribution of the variables.

Table 3: Identifying outliers using Box plot and 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



Identifying outliers using the residuals

The goal is to detect the cases which have large residuals (outliers), and the cases whose removal (influential cases),creates a different model. The distinction between these two kinds of cases is not always obvious. Both types of points are of great concern. There is a total of 59 residuals (Figure 7) [33].

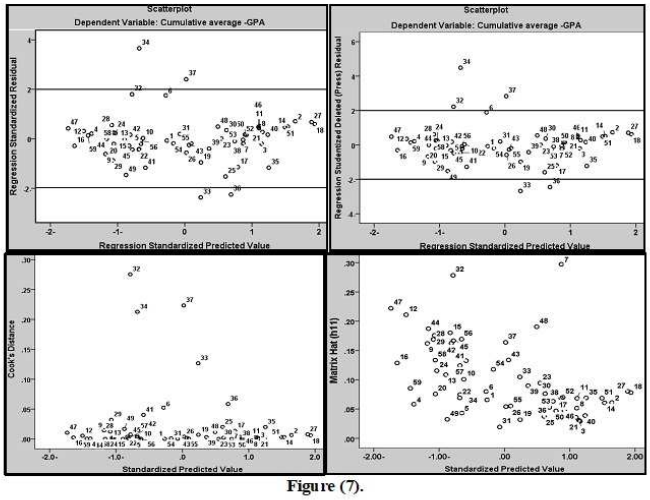


Figure (7).

Fig. 7: Identifying outliers using Several Measures

The scatter plot in Figure 7 shows standardized predicted values against standardized residua and Studentized deleted residual. The results indicate that some of extreme residuals are worthy of investigation, where the cases 32 ,33, 34, 36, and 37 are problematic. However, it is noted that the case no6 is suspicious and it is confirmed by the scatter plot for standardized predicted values against Cook's Distance and hat matrix.

Significance test of outliers

The study has used Grubbs' test and Dixon' test (Figure 8). The results showed that the cases diagnosed as outliers through the Grubbs' test had a significant effect on the regression coefficients. However, the cases that have been diagnosed as outliers through the Dixon's test did not have any effect.

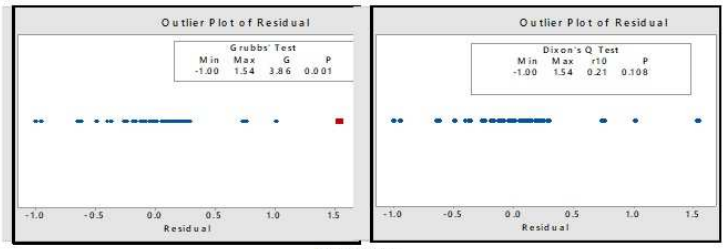


Figure (8) .

Fig. 8: Outliers Significance test .

Identifying Influential Observations in (y and x_i)

To identify whether outliers are influential or not, not necessarily that all outliers' observations are influential. In this regard, box plot will be used by overall measures of influence (DFFITS, COVRATIO, and Cook's D) to discover influential cases in y, and (DFBETA) to discover influential cases in xi [34]. The cases which form a star are influential, while the cases which form a circle are influentiall (Figure 9-13).

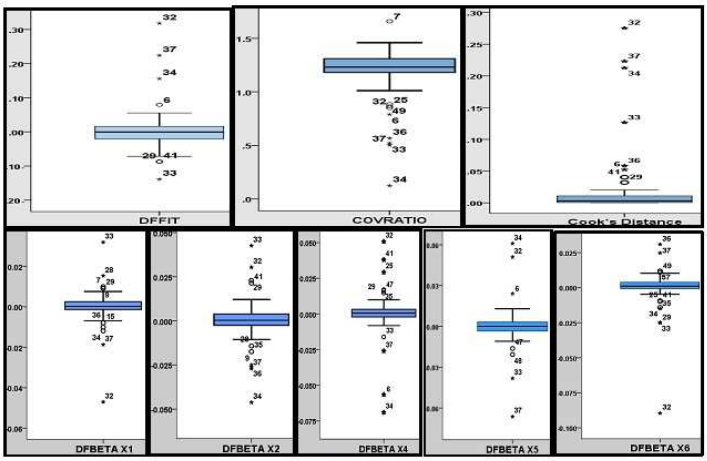


Figure (9).

Fig. 9: Box plot for overall measures to identify influential cases in y and xi

Application of the proposed algorithm

The multiple correlation between independent and dependent variables is used to find the partial correlation coefficient. The sum of the averages of the independent variables are presented below (Table 4-6).

Table 4: Partial and multiple correlation and Sum of the averages (x_i)

Var.	x_1	x_2	x_3	x_4	x_5	x_6	(x_i)
y	0.653	0.719	0.822	0.751	0.748	0.633	0.872
Sum of the averages (x_i)							
Cases	$\bar{x}_{i_6}$	$\bar{x}_{i_{25}}$	$\bar{x}_{i_{28}}$	$\bar{x}_{i_{29}}$	$\bar{x}_{i_{32}}$	$\bar{x}_{i_{33}}$	$\bar{x}_{i_{36}}$	$\bar{x}_{i_{37}}$	$\bar{x}_{i_{41}}$
$\sum_{i=1}^{p} \bar{x}_{im}$	3.0	3.8	2.6	2.34	2.47	3.48	3.74	3.04	2.57

Table 5: Fitting the regression model using the proposed algorithm

Model Summary				
S	**R-sq**	**R-sq (adj)**	**R-sq (pred)**	
0.198	94.64%	94.02%	93.31%	
Model Summary and Coefficients				
Term	**Coef (β_i)**	**SE Coef**	**T-Value**	**P-Value**
Constant	0.2490	0.1040	2.40	0.020
x_1	0.1030	0.0360	2.86	0.006
x_2	0.1231	0.0476	2.59	0.013
x_3	0.1938	0.0424	4.57	0.000
x_4	0.2526	0.0518	4.87	0.000
x_5	0.1043	0.0499	2.09	0.041
x_6	0.1405	0.0463	3.03	0.004

Regression Equation:

$$y = 0.249 + 0.103\,x_1 + 0.1231x_2 + 0.1938x_3 + 0.2526x_4 + 0.1043x_5 + 0.1405\,x_6$$

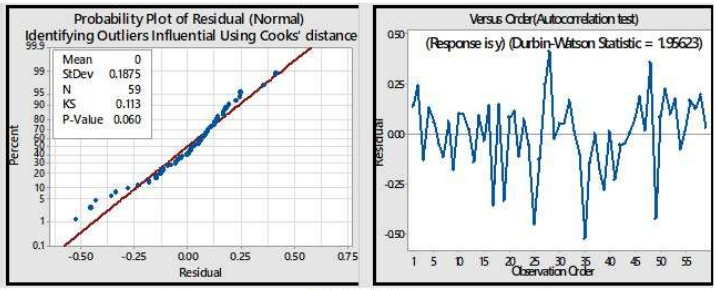


Figure (10)

Fig. 10: Checking hypothesis (OLS) after using the proposed algorithm

Table 6: Fitting the regression model after (OLS) after deleted outliers

Model Summary				
S	**R-sq**	**R-sq (adj)**	**R-sq (pred)**	
0.205	94.64%	93.79%	92.97%	
Model Summary and Coefficients				
Term	**Coef (β_i)**	**SE Coef**	**T-Value**	**P-Value**
Constant	0.1860	0.1210	1.54	0.129
x_1	0.1314	0.0369	3.56	0.001
x_2	0.1372	0.0516	2.66	0.011
x_3	0.1895	0.0504	3.76	0.000
x_4	0.2458	0.0582	4.23	0.000
x_5	0.1092	0.0522	2.09	0.042
x_6	0.1204	0.0467	2.58	0.013

Regression Equation:

$$y = 0.186 + 0.1314\,x_1 + 0.1372x_2 + 0.1895x_3 + 0.2458x_4 + 0.1092x_5 + 0.1204\,x_6$$

Table 7: Fitting the regression model using weighted least squares (WLS).

Model Summary				
S	**R-sq**	**R-sq (adj)**	**R-sq (pred)**	
1.31523	94.36%	93.71%	91.72%	
Model Summary and Coefficients				
Term	**Coef (β_i)**	**SE Coef**	**T-Value**	**P-Value**
Constant	0.5300	0.1470	3.61	0.001
x_1	0.1000	0.0278	3.60	0.001
x_2	0.0691	0.0446	1.55	0.127
x_3	0.3038	0.0525	5.79	0.000
x_4	0.1418	0.0569	2.49	0.016
x_5	0.0952	0.0435	2.19	0.033
x_6	0.1323	0.0451	2.93	0.005

Regression Equation:

$$y = 0.53 + 0.10\,x_1 + 0.0691x_2 + 0.3038x_3 + 0.1418x_4 + 0.0952x_5 + 0.1323\,x_6$$

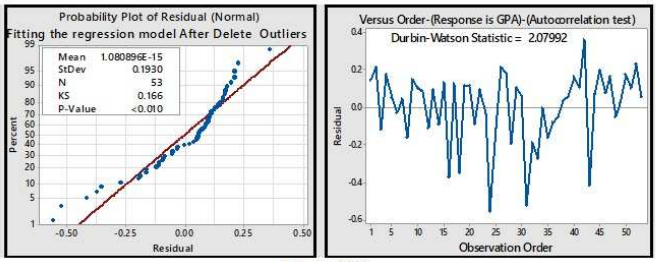


Figure (11)

Fig. 11: Checking the hypothesis of (OLS) after delete outliers

Fitting the regression model using weighted least squares (WLS) (Table 7) [35].

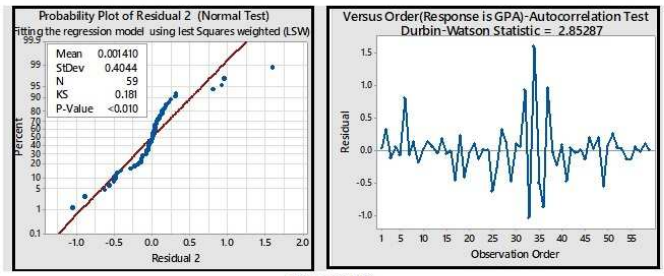


Figure (12).

Fig. 12: Checking the hypothesis of (OLS) after using (WLS)

Fitting the regression model using Trimmed Mean (Table 8) [36].

Table 8: Fitting the regression model using Trimmed Mean

Model Summary				
S	**R-sq**	**R-sq (adj0**	**R-sq (pred)**	
0.181	95.17%	94.62%	93.79%	
Model Summary and Coefficients				
Term	**Coef (β_i)**	**SE Coef**	**T-Value**	**P-Value**
Constant	0.1630	0.1060	1.54	0.130
x_1	0.1440	0.0324	4.44	0.000
x_2	0.1546	0.0452	3.42	0.001
x_3	0.1665	0.0408	4.08	0.000
x_4	0.2226	0.0519	4.29	0.000
x_5	0.1256	0.0458	2.74	0.008
x_6	0.1310	0.0412	3.18	0.002

Regression Equation:

$$y = 0.163 + 0.144\,x_1 + 0.1546x_2 + 0.1665x_3 + 0.2226x_4 + 0.1256x_5 + 0.131\,x_6$$

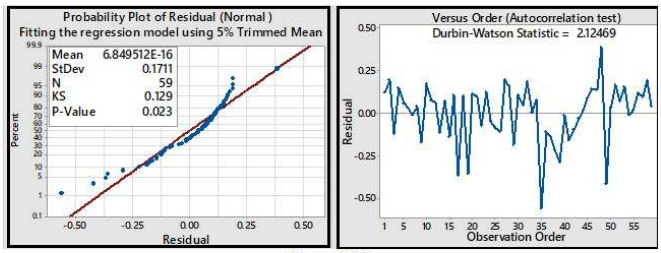


Figure (13)

Fig. 13: Checking the hypothesis of (OLS) after using Trimmed Mean

Table 9: Comparison of estimation methods

Method	Sig.	β_1	β_2	β_3	β_4	β_5	β_6	R^2 %	Norm.	D.W
OLS	P-Value	0.217	0.286	0.002	0.341	0.124	0.468	73.32	0.01	2.860
WLS	P-Value	0.001	0.127	0.000	0.016	0.033	0.005	93.71	0.01	2.853
Delete Outliers	P-Value	0.001	0.011	0.000	0.000	0.042	0.013	93.79	0.01	2.080
Trimmed Mean	P-Value	0.000	0.001	0.000	0.000	0.008	0.002	94.62	0.023	2.125
Proposed method	P-Value	0.006	0.013	0.000	0.000	0.041	0.004	94.02	0.060	1.956

Significance of regression coefficients was achieved using three methods (i.e. Proposed Method, Delete Outliers, Trimmed Mean). In addition, the highest coefficient of determination was (94.62%), using trimmed mean method. However the removal of outliers can be dangerous because it may end up destroying some of the most important information in the data. The hypothesis of the independence of residues was achieved in three methods (Proposed Method, Delete Outliers, Trimmed Mean). Only the normality hypothesis for residues was achieved using the proposed algorithm (Table 9).

4 Conclusion

The present paper adopted MATLAB, SPSS, and EVIEWS to perform the computations. All methods of estimation were compared using three standards (The significance of regression coefficients (P-Valueβ_i), adjusted determination coefficient (Adj.R^2), and achieving the assumptions of OLS]. They were applied to a real data. 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 significant multiple outliers in the data set. The study has found that the proposed algorithm can obtain highly efficient estimates of regression coefficients. Thus, we recommend diagnosing outliers before doing analysis and using the proposed algorithm to estimate multiple outliers in regression model .

Conflict of interest: There is no conflict of interest in this paper.

References

[1] V.Barnett and T. Lewis, Outliers in Statistical Data, John Wiley & Sons, Chichester, 49-95, (1994).

[2] F.H. M. Salleh, S. Zainudin and S.M. Arif, Multiple Linear Regression for Reconstruction of Gene Regulatory Networks in Solving Cascade Error Problems. Hindawi Limited , Advances in Bioinformatics, 2017,1-14, (2017) .
[3] Y. Zhou, Y. Cheng and T. Tong, A Least Squares Method for Variance Estimation in Heteroscedastic Nonparametric Regression. Journal of Applied Mathematics. 2014, 1-14, (2014).
[4] J. Neter, M. Kutner, C. Nachtsheim and W. Wasserman, Applied Linear Regression Models. Irwin, Chicago,66-95, (1997).
[5] [5]. R.Johnson, Estimating the Size of a Population, International Journal for Statistics and Data Science Teaching,16,50-52,(1994).
[6] R.J. Freund, W.J. Wilson, P. Sa, Regression Analysis, Academic Press, Elsevier. 119-131 ,(2006).
[7] J. Fan and L.S. Huang, Goodness-of-fit tests for parametric regression models, Journal of the American Statistical Association, 96 , 640-652, (2001).
[8] K.K. L. B Adikaram, M.A. Hussein, M. Effenberger and T. Becker. Outlier Detection Method in Linear Regression Based on Sum of Arithmetic Progression. The Scientific World Journal, 2014, 1-12, (2014).
[9] S. Weisberg, Applied Linear Regression , Wiley , Hoboken NJ, 200-255, (2013).
[10] A. Cerioli, M. Riani and Torti, F. Size and power of multivariate outlier detection rules, In Algorithms from and for Nature and Life, Berlin: Springer-Verlag, 3-17 , (2013)..
[11] P.J. Rousseeuw and A.M. Leroy, Robust Regression and Outlier Detection. John Wiley & Sons, Inc. New York, NY United States,250-262, (1987).
[12] P.J. Huber and E.M. Ronchetti, Robust Statistics, John Wiley & Sons, New York, 149-198, (1981).
[13] M. Alguraibawi, H. Midi and A. Imon, A new robust diagnostic plot for classifying good and bad high leverage points in a multiple linear regression model. Mathematical Problems in Engineering, 2015 , 1-12, (2015).
[14] F. Ghapani, A.R. Rasekh, M.R. Akhoond B and Babadi, Detection of Outliers and Influential Observations in Linear Ridge Measurement Error Models with Stochastic Linear Restrictions. Journal of Sciences, 26 , 355- 366 , (2015).
[15] R.D. Cook, Detection of influential observation in linear regression. Taylor & Francis, Ltd. ,Technometrics ,19,15-18,(1977).
[16] S. Turkan, M. Candan and O. Toktamıs, Outlier Detection by Regression Diagnostics in Large Data. Journal of Mathematics and Statistics, 41,147-155, (2012).
[17] A.H.M. Imon, A stepwise procedure for the identification of multiple outliers and high leverage points in linear regression, Pakistan Journal of Statistics, 21,(2005).
[18] F.G. Richard, Regression Analysis and its Application, A Data-Oriented Approach, Boca Raton, 42 -92, (2019).
[19] B.M. Greenwell, A.J. McCarthy, B.C. Boehmke and D. Liu, Residuals and Diagnostics for Binary and Ordinal Regression Models: An Introduction to the sure Package. The R Journal., 10, 381-394 (2018)
[20] R. Dennis Cook and Sanford Weisberg, Residuals and Influence in Regression. New York: Chapman and Hall, 10- 45, (1982).
[21] C.M. Judd, G.H. McClelland and C.S. Ryan, Data analysis: A model comparison approach to regression, ANOVA, and beyond, Routledg , New York, 102-132,(2017).
[22] D.A. Belsley, E. Kuh and R.E. Welsch. Regression Diagnostics, John Wiley & Sons. New York., 97-102, (1980).
[23] U. Balasooriya, Y.K. Tse and Y.S. Liew, An empirical comparison of some statistics for identifying outliers and influential observations in linear regression models. Journal of Applied Statistics. 14 , 177-184, (1987).
[24] R. Valliant, Regression Diagnostics in Survey Data. DC09 Stata Conference 15, Stata Users Group,1-12,(2009).
[25] A. Abuzaid, I. Mohamed, A.G. Hussin and A. Rambli, COVRATIO statistic for simple circular regression model, Chiang Mai J. Sci, 38, 321-330 , (2011).
[26] H.G. Müller, Goodness-of-fit diagnostics for regression models. Scandinavian Journal of Statistics , 19, 157-172 , (1992) .
[27] J.L. Gastwirth, Y.R. Gel and W. Miao, The Impact of Levenes Test of Equality of Variances on Statistical Theory and Practice, Institute of Mathematical Statistics, 24, 343-360, (2009).
[28] K.D.S. Young, Bayesian diagnostics for checking assumptions of normality, Journal of Statistical Computation and Simulation, 47,.167–180, (1993).
[29] S. Mohanasundaram, B. Narasimhan and G.S. Kumar, The Significance of Autocorrelation and Partial Autocorrelation on Univariate Groundwater Level Rise (Recharge) Time Series Modeling, JGWR, AGGS, India ,131-142,(2013).
[30] M. Nussbaum, Categorical and Nonparametric Data Analysis Choosing the Best Statistical Technique. Routledge, Taylor & Francis, 177-185,(2014).
[31] A. Cerioli, M. Riani and A.C. Atkinson, Controlling the size of multivariate outlier tests with the MCD estimator of scatter. Statistics and Computing Journal , 19,341, (2009).
[32] C. Chamnein and M.I. Don, Can the Box Plot be Improved, Songklanakarin Journal of Science and Technology , 27,649-657,(2005).
[33] X. Gao and Y. Fang, Penalized weighted least squares for outlier detection and robust regression, arXiv preprint arXiv:1603.07427, stat.ME,1, 1-27,(2016).
[34] P.J. De Jongh, T. De Wet and A.H. Welsh, Mallows-type bounded-influence-regression trimmed means, Journal of the American Statistical Association, 83 , 805-810, (1988).
[35] P. J. De Jongh , T. de Wet and A. H .Welsh , Mallows-Type Bounded-Influence-Regression Trimmed Means, Journal of the American Statistical Association , 83, 805-810 , (1988).
[36] Ali S. Hadi and Jeffrey S. Simonoff, Procedures for the Identification of Multiple Outliers in Linear Models, Journal of the American Statistical Association, 88, 1264-1272, (1993).1993.