IBBUL
Lapai Journal of Applied and Natural Sciences
LAJANS Vol 6(1): 161 - 169

RESEARCH PAPER

Application Of Nonlinear Regression Model to Coordinate Statistical Outliers and Leverage Points

[1*]Abdulkarim, K., [2]Rasheed, B.A., [1]Bawa, M.U., [1]Aliyu, I.D., [3]Abdullahi, S. and [4]Ibrahim, H.

[1]Department of Mathematical Sciences, Abubakar Tafawa Balewa University, Bauchi, Nigeria
[2]Department of Mathematics and Statistics, Federal Polytechnic, Bauchi, Nigeria
[3]Department of General Studies, Bauchi State College of Agriculture, Yelwa, Nigeria
[4]Department of Statistics, Jigawa State Polytechnic, Dutse, Nigeria

ABSTRACT

The presence of multicollinearity is unavoidable in data sets. It is obvious that high leverage points are another source of multicollinearity. These leverage points may reduce multicollinearity in a dataset. We call this high leverage collinearity reducing observations. Several criteria were studied to assess the merit of the proposed detection method over the existing method using different sample sizes, percentages and position of high leverage points which cause these leverages to change the multicollinearity pattern of non-collinear data set. However, the detection method LTSR-HLCIM which is based on DRGP(MVE) depends on minimum volume ellipsoid which has a long computing running time, computational complexity and the presence of swamping and masking effects. Another diagnostic measure was introduced to reduce the three shortcomings mentioned in the existing method by using DRGP(RFCH) in (Improvised LTSR-HLCIM) instead of DRGP(MVE) used by LTSR-HLCIM. The results of simulation study indicated that the modified detection method outperformed the existing method in term of having the highest percentage of correct detection of good and bad leverage collinearity enhancing observations of the small, medium and large sample sizes with three different levels of HLPs.

Keywords: Linear regression, Ordinary least squares, Collinearity reducing observation, High leverage points, Vertical outliers

INTRODUCTION

One of the predominant regression analysis techniques is the Ordinary Least Squares (OLS) method. Data analyzers prefer to apply OLS due to the universal acceptance, elegant statistical properties, and computational simplicity. This method minimizes the sum of squared of regression errors which are the estimate of distances between the responses predicted by the linear approximation and the observed responses in the data set. Unfortunately, the mathematical elegance that makes OLS so popular depends on a number of fairly restrictive and often unrealistic assumptions. Two of the assumptions that make OLS so attractive in terms of general model hypothesis and parameter significance testing, are normality of error distribution and independency of the explanatory variables. The normality assumption can be violated in the presence of one or more sufficiently outlying observations in the data set resulting in less reliable estimates of the model parameters (Andersen, 2008). The second condition that potentially impacts the reliability of the least squares estimation is multicollinearity, which is a perfect or near-linear dependency among the explanatory variables (X-direction) in multiple regression model. However, due to multicollinearity, it is difficult to get coefficient estimates with small standard error. Hence, the confidence intervals of the parameter estimates become wider and the t-values become insignificant which leads to the rejection of the existence of some of the important coefficients in the regression model.

The LTSR-HLCIM successfully among the existing methods classified vertical outliers,

Received 12 September, 2021
Accepted 03 December, 2021
Address Correspondence to:
kamlakmashi@gmail.com

good leverage collinearity-reducing observations, bad leverage collinearity-reducing observations, vertical outliers and regular observations using simulation data of small, medium and large sample sizes (Bagheri and Habshah, 2015). But the shortcoming of this method used Diagnostic Robust Generalized Potential (DRGP) which depends on Minimum Volume Ellipsoid (MVE) or Minimum Covariance Determinant (MCD) for obtaining the final estimator of location and scatter. This diagnostic measure has long computing running time, swamping and masking effects and computational complexity. Furthermore, (Habshah *et al.,* 2020) found out that the DRGP(ISE) is not very stable and its running time can be improved and also the rate of the small percentage of swamping and masking effect can also be reduced using another fast and robust diagnostic measure. Therefore, this has motivated us to propose (Improvised LSTR-HLCIM) which will reduce swamping and masking effects, computational complexity and has fast computing running time using Reweighted Fast Consistent and High Breakdown denoted as DRGP(RFCH) proposed by (Habshah *et al.,* 2020) instead of MVE used by LSTR-HLCIM.

METHODOLOGY

The non-collinear dataset used in this study created using simulation technique of small, medium and large sample sizes with three different levels of High Leverage Points (HLPs) to assess the merit of the modified method (Improvised LTSR-HLCIM) over the existing method (LTSR-HLCIM).

Diagnostic Robust Generalized Potential based on (Minimum Volume Ellipsoid)

Rousseeuw (1985) Introduced Robust Mahalanobis Distance (RMD), based on the Minimum Volume Ellipsoid (MVE). The DRGP algorithm comprises two steps where;

Step I: Robust Mahalanobis Distance (RMD) based on MVE is used to detect the suspected HLPs (bad cases). Let us write the ith vector of predictor variables as:

$$X_i' = (1, X_1, X_2, \dots, X_3) = (1, t_i),$$

Where t_i is a p-dimentional row vector. The mean vector and the variance covariance matrix are calculated as:

$\bar{t} = {}^1\!/_n \sum_{i=1}^n t_i$ and $c = \left(\frac{1}{n-1}\right)\sum_{i=1}^n (t_i - \bar{t})(t_i - \bar{t})'$, respectively.

However, (Rousseeuw and Leroy 1987) suggested using Robust Mahalanobis Distance (RMD) as a diagnostic tool for detection of HLPs.

$$\text{RMDi(MVE)} = \sqrt{(t_i - T_R(X))' C_R(X)^{-1} (t_i - T_R(X))} \quad i = 1,2,\dots,n \qquad (1)$$

Where $T_R(X)$, $C_R(X)$ are robust locations and shape estimate of the MVE and where RMDi(MVE) = Median (RMD_i(MVE)) + c MAD(RMD_i (MVE)) which is the non-parametric cutoff point to determine the suspect high leverage point.

Step II: Generalized potential will be employed to confirm whether or not the suspected HLPs are genuine HLPs. Even though the DRGP is very successful in identifying HLPs, its running time is very slow due to using MVE in the first step.

$$DRGP = (p_{ii}^*) = \begin{cases} h_{ii}^{(-D)} & for\ i \in D \\ \dfrac{h_{ii}^{(-D)}}{1 - h_{ii}^{(-D)}} & for\ i \in R \end{cases} \qquad (2)$$

Hat matrix plays an important role in determining directly outlying X observations. This matrix is traditionally used as a measure of leverage points in regression analysis which is defined as

$$H = X(X'X)^{-1}X',$$

The diagonal elements of the H matrix which are known as the leverage values are expressed by

Where, $h_{ii}^{(-D)} = x_i'(X'X)^{-1}x_i,$

If all members of the D set in (p_{ii}^*) > median (p_{ii}^*) + 3MAD(p_{ii}^*), we declared them as HLPs, otherwise, those observations are put back into the estimation subset R.

Diagnostic Robust Generalized Potential based on (Reweighted Fast Consistent and High Breakdown Estimator)

Olive and Hawkins (2010) developed Reweighted Fast Consistent and High breakdown (RFCH) estimators of location and scatter which was faster than the fast MCD developed by (Rousseeuw and Driessen, 1999).

The advantage of RFCH technique is that not only its computation is very fast which is even faster than Fast MCD (Zhang *et al.*, 2012), but it is $\sqrt{n}$ consistent estimators. The RFCH utilizes the consistent DGK (Devlin *et al.*, 1981) estimator and high breakdown Median Ball (MB) (Olive & Hawkins, 2008) estimators as attractors. The RFCH algorithms can be summarized as follows:

Step I: Identify the suspected HLPs by using *RMDi* for each i[th] observation based on RFCH:

$$RMDi = \sqrt{(t_i - T_{RFCH}(X))' C_{RFCH}(X)^{-1}(t_i - T_{RFCH}(X))} \quad i = 1,2,\dots,n \qquad (3)$$

As per Midi et al. (2009) the cut-off point is defined as follows;

Cut_off=median(*RMDi*)+3MAD($RMDi$),

Where,

MAD(RMDi)=medianabs(RMDi) – median(RMDi/0.6745)

We declare that any i^{th} case with Robust RMDi > cut_off point, is the suspected HLPs and include them in a deletion group, denoted as D Group, while the rest of the observations are kept in the R

group.

Step II: The DRGP $(p_{ii}^*) = \begin{cases} h_i^{(-D)} & for\ i \in D \\ \frac{h_i^{(-D)}}{1-h_i^{(-D)}} & for\ i \in R \end{cases}$ (4)

Compute the cut-off point *pii*. If all members of the D set in $(p_{ii}^*) >$ median(p_{ii}^*)+3$Q_n(p_{ii}^*)$, we declared them as HLPs, otherwise, those observations are put back into the estimation subset R.

Rousseeuw and Croux (1993) noted that to make Q_n a consistent estimator for Gaussian data the value of C should be chosen equals to 2.2219.

High Leverage Collinearity-Influential Observations Measure

High leverage collinearity influential observations are those high leverage points that change the multicollinearity pattern of a data. (Bagheri et al., 2012) proposed a high leverage collinearity-influential measure based on DRGP(MVE), namely HLCIM (DRGP), denoted as $\delta_i^{(D)}$ and which is defined as follows:

$$\delta_i^{(D)} = \begin{cases} log\frac{k_{(D)}}{k_{(D-i)}} & if\ i \in D, \quad num\{D\} \neq 1 \\ log\frac{k_{(i)}}{k} & if\ num\{D\} = 1,\ D = i,\ i = 1,2,\dots,n, \\ log\frac{k_{(D+i)}}{k_{(D)}} & if\ i \in R, \end{cases} \qquad (5)$$

Where *D* is the suspected group of multiple high leverage points and *R* is the remaining good observations to be diagnosed by DRGP(MVE) (Habshah *et al.*, 2009) and DRGP(RFCH), p_{ii}^*. However, the MVE will be substituted with the ISE for the modified method i.e. HLCIM(DRGP(RFCH)). (Bagheri *et al.*, 2012) summarized the algorithm of HLCIM(DRGP) in three steps as follows.

Step I: Calculate DRGP(MVE) and DRGP(RFCH).

Step II: Compute high leverage collinearity-influential values $\delta_i^{(D)}$ as follows.

(i) If only a single member in the D group, the size of R is (n−1), and D=i, calculate $log\frac{k_{(i)}}{k}$ where (k_i) indicates the condition number of the X matrix without the *ith* high leverage points.

(ii) If more than one member in the D group, calculate $log\frac{k_{(D)}}{k_{(D-i)}}$ where (k_{D-1}) indicates the condition number of the X matrix without the entire D group minus the *ith* high leverage points, where *i* belongs to the suspected D group.

(iii) For any observation in the R group, compute $log\frac{k_{(D+i)}}{k_{(D)}}$ where (k_{D+1}) refers to the condition number of the X matrix without the entire group of D high leverage points plus the *ith* additional observation of the remaining group.

Step III. If any $\delta_i^{(D)}$values for *i=1,2,…n* does not exceed the cutoff points, put back the *ith* observation to the R group. Otherwise, D group is the high leverage collinearity enhancing observations.

The Cutoff Point of High Leverage Collinearity Influential Measure

The cutoff points to be used for the HLCIM for the identification of high leverage collinearity-enhancing observations were proposed by (Bagheri et al., 2012) and (Bagheri and Habshah, 2012) for $\theta_i, i = 1,2, \dots, n$:

$cut^1(\theta) = Median(\theta_i) - 3Mad(\theta_i)$;
$cut^2(\theta) = Median(\theta_i) + 3Mad(\theta_i)$

Where $cut^1(\theta)$ is the cutoff point for collinearity-enhancing measure and $cut^2(\theta)$ is the collinearity-reducing measure cutoff point. Median and Mean Absolute Deviation (MAD) stand for robust measures of central tendency and dispersion, respectively. More so, θ_i can be $\delta_i^{(D)}$ and c is the chosen constant values of 3.

The LTSR-HLCIM

However, the existing method to be used is based on the diagnostic measure for the identification of multiple high leverage collinearity-influential observations, each of the LTS residuals, r_i for $i = 1,2,3 \dots n,$ is standardized by $\hat{\sigma}$. The LTSR– HLCIM method plots the standardized least trimmed square residuals (LSTR) which is the horizontal cutoff point as suggested by (Rousseeuw and Van Zomeren, 1990) with 97.5% quantile of the chi-square distribution with also 1 degrees of freedom $(\pm\sqrt{\chi^2_{1,\ 0.975}})$ against the HLCIM(DRGP(MVE)) on the X-axis denoted as $\delta_i^{(D)}$.

The Improvised LTSR-HLCIM

The proposed (Improvised LTSR-HLCIM) method is based on the diagnostic measure for the identification of multiple high leverage collinearity-influential observations, each of the LTS residuals, r_i for $i = 1,2,3..n,$ is standardized by $\hat{\sigma}$. Therefore, in the proposed method, each of the LTS residuals, r_i for $i = 1,2,3..n$ is standardized by $\hat{\sigma}$. The LTSR–HLCIM(DRGP(RFCH)) plots the standardized LTS residuals (LTSR) which is the horizontal cutoff point in both panels $(\pm\sqrt{\chi^2_{1,\ 0.975}})$ on the Y-axis against the HLCIM(DRGP(RFCH)) on the X-axis denoted as $\delta_i^{(D)}$.

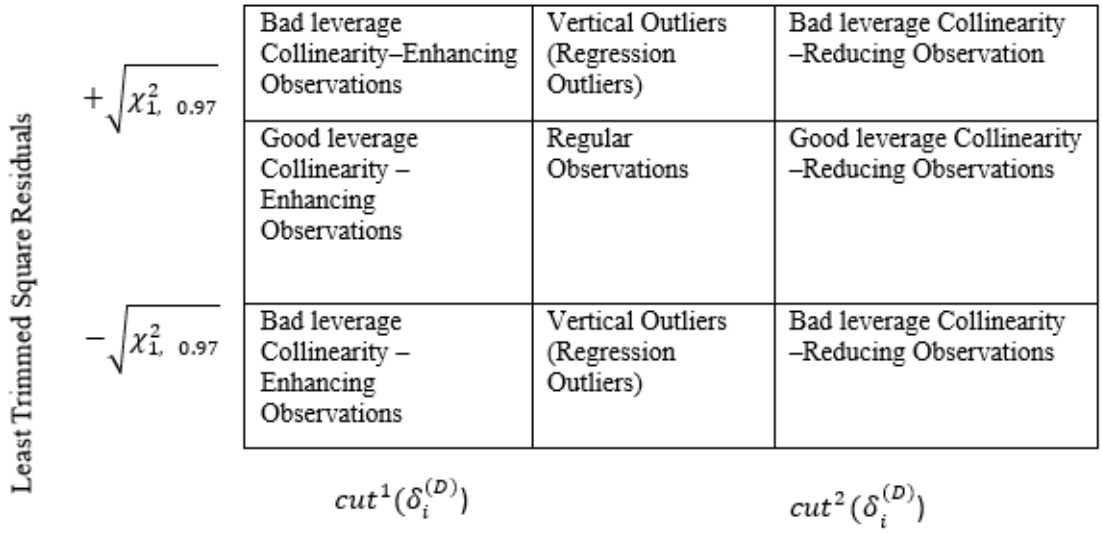

Figure 1: The Venn diagram of LTRS-HLCIM Plot.

Diagnostic Method for Multicollinearity

The most common used measure of multicollinearity is the Classical Variance Inflation Factor (VIF) (Wonsuk et al., 2014). The VIF can be calculated using the equation.

$VIF = \frac{1}{1-R_J^2}$, where $J = 1,2, \dots, p-1$

Where R_J^2 denote the coefficient of determination when xj is regressed on all other

predictor variables in the model. VIF provides an index that measures how much the variance of an estimated regression coefficient is increased because of the multicollinearity. As per practical experience, if any of the VIF >5, it is an indication that the associated regression coefficients are poorly estimates because of multicollinearity (Montgomery *et al*., 2001).

Monte Carlo simulation Technique

A Monte Carlo simulation was used to assess the merit of our proposed method over the existing method in terms of its ability to excellently detect collinearity reducing observations. To achieve the merit of the (Improvised LTSR-HLICM), we follow the simulation procedure used by (Bagheri and Habshah, 2015). Considering a linear relation $y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e_i$ where $i = 1,2,\dots,n$. Three explanatory variables $(p = 3)$ were generated from Uniform $(0,1)$ to produce non-collinear data sets in which the true parameters were set at $\beta_0 = \beta_1 = \beta_2 = \beta_3 = 1$ and $\varepsilon_i \sim N(0, \delta_I^2)$ in such a way that different scenarios were created, namely, high leverage collinearity-reducing observations and vertical outliers. In each scenario, small, medium and large samples of size 40, 100 and 300 with three different levels of high leverages points (HLP) of 5%, 10% and 15% for all the sample sizes considered at 1000 replications.

As per (Lawrence and Arthur, 1990), high leverage collinearity-reducing observations are created by generating three collinear regressors on the outset:

$$X_{ij} = (1-\rho^2)Z_{ij} + \rho Z_{i4} \qquad i = 1,\dots,n; \quad j = 1,\dots,3.$$

Where, the X and Z are independent standard normal random numbers. The value of ρ^2 represents the correlation between the two explanatory variables, is set to 0.95 which cause high collinearity between regressors. High leverage collinearity-reducing observations in collinear data sets were then created by replacing the first $100(\frac{\alpha}{2})$ percent observations of X_1 and the last $100(\frac{\alpha}{2})$ percent observations of X_2 with high leverage points.

To create vertical outliers, a dependent variable from a Uniform (0, 1) was firstly generated. For each sample size, a certain percentage of outliers was generated by randomly deleting a certain percentage of 'good' observations and replacing them with 'bad' data points. The first outlier is kept fixed at 100 (10^2) and the successive values are created by multiplying the observations index,i by 10

RESULTS AND DISCUSSION

Table 1: The Abbreviations used in the simulation result for the evaluation of the two methods under study.

Abbreviations	Meaning
VIF	The Variance Inflation Factor without high leverage points
VIF**	The Variance Inflation Factor with high leverage points
RO	regular observations
DRO	The number of detected regular observations
VO	vertical outliers
DVO	The number of detected vertical outliers
BLCRO	bad leverage collinearity –reducing observations
DBLCRO	The number of detected bad leverage collinearity –reducing observations
GLCRO	Good leverage collinearity –reducing observations
DGLCRO	The number of detected good leverage collinearity–reducing observations

Table 2: The number of detected abnormal observations in the simulated data sets with vertical outlier using LTSR-HLCIM(DRGP(MVE)).

(n)			40			100			300	
∝	0.00	0.05	0.10	0.15	0.05	0.10	0.15	0.05	0.10	0.15
VIF	3.44	3.44	3.29	3.29	3.29	3.29	3.29	3.29	3.29	3.29
VIF*	3.44	3.44	3.29	3.29	3.29	3.29	3.29	3.29	3.29	3.29
RO	40	38	36	34	95	90	85	285	270	255
DRO	38.42	34.59	33.24	31.95	93.50	88.75	83.24	283.38	268.39	253.07
%	*96.1%*	*91.1%*	*92.3%*	*93.9%*	*98.4%*	98.6%	*97.9%*	*99.4%*	*99.4%*	*99.2%*
VO	0	2	4	6	5	10	15	15	30	45
DVO	0	1.85	3.87	5.88	4.18	8.91	14.30	14.30	29.47	44.60
%	100%	*92.5%*	*96.8%*	*98.0%*	*83.6%*	*89.1%*	*95.3%*	*95.3%*	*98.2%*	*99.1%*

Table 3: The number of detected abnormal observations in the simulated data sets with vertical outlier using LTSR-HLCIM(DRGP(RFCH)).

(n)			40			100			300	
∝	0.00	0.05	0.10	0.15	0.05	0.10	0.15	0.05	0.10	0.15
VIF	3.17	3.17	3.02	3.02	3.02	3.02	3.02	3.02	3.02	3.02
VIF**	3.17	3.17	3.02	3.02	3.02	3.02	3.02	3.02	3.02	3.02
RO	40	38	36	34	95	90	85	285	270	255
DRO	40	38	35.99	33.99	94.56	89.81	84.30	285	270	255
%	*100%*	*100%*	*99.9%*	*99.9%*	*99.5%*	99.8%	*99.2%*	*100%*	*100%*	*100%*
VO	0	2	4	6	5	10	15	15	30	45
DVO	0	1.95	3.99	5.98	4.98	9.99	15	15	30	45
%	100%	*97.5%*	*98.8%*	*99.7%*	*99.6%*	*99.9%*	*100%*	*100%*	*100%*	*100%*

Table 4: The number of abnormal observations in the simulated data sets with bad leverage collinearity reducing observations using LTSR-HLCIM.

(n)		40			100			300	
∝	0.05	0.10	0.15	0.05	0.10	0.15	0.05	0.10	0.15
VIF	39.40	38.91	36.33	41.17	36.80	39.10	37.13	38.34	38.97
VIF**	13.12	4.46	1.06	1.13	1.03	1.02	1.01	1.01	1.00
RO	38	36	34	95	90	85	285	270	255
BLCRO	2	4	6	5	10	15	15	30	45
DBLCRO	1.85	3.96	5.98	4.91	9.96	14.55	15	30	45
%	*92.5%*	*99%*	*99.7%*	*98.2%*	*99.6%*	*97%*	*100%*	*100%*	*100%*

Table 5: The number of abnormal observations in the simulated data sets with bad leverage collinearity reducing observations using (Improvised LTSR-HLCIM).

(n)		40			100			300	
∝	0.05	0.10	0.15	0.05	0.10	0.15	0.05	0.10	0.15
VIF	39.48	38.81	36.83	41.43	36.12	39.32	37.76	38.21	38.96
VIF**	15.12	6.46	3.06	3.13	3.03	3.02	3.01	3.01	3.00
RO	38	36	34	95	90	85	285	270	255
BLCRO	2	4	6	5	10	15	15	30	45
DBLCRO	1.99	4	5.99	5	9.99	15	15	30	45
%	99.5%	100%	99.8%	100%	99.9%	100%	100%	100%	100%

Table 6: The number of abnormal observations in the simulated data sets with good leverage collinearity reducing observations using LTSR-HLCIM.

(n)		40			100			300	
∝	0.05	0.10	0.15	0.05	0.10	0.15	0.05	0.10	0.15
VIF	38.39	40.70	36.33	40.13	36.76	37.16	38.34	37.20	37.28
VIF**	1.84	2.49	1.06	1.04	1.02	1.01	1.01	1.01	1.00
RO	38	36	34	95	90	85	285	270	255
GLCRO	2	4	6	5	10	15	15	30	45
DGLCRO	1.85	3.93	5.9	4.30	8	13	13	29	44
%	92.5%	98.3%	98.3%	86%	80%	86.7%	86.7%	96.7%	97.8%

Table 7: The number of abnormal observations in the simulated data sets with good leverage collinearity reducing observations using LTSR-HLCIM(DRGP(RFCH)).

(n)		40			100			300	
∝	0.05	0.10	0.15	0.05	0.10	0.15	0.05	0.10	0.15
VIF	39.30	41.71	36.33	40.13	37.79	38.16	39.32	38.25	38.28
VIF**	1.85	1.80	1.08	1.06	1.04	1.03	1.03	1.03	1.01
RO	38	36	34	95	90	85	285	270	255
GLCRO	2	4	6	5	10	15	15	30	45
DGLCRO	1.90	3.99	5.99	5	10	14	15	30	45
%	95%	99.8%	99.8%	100%	100%	93.3%	100%	100%	100%

When α = 0.00 in Table 2 above, It can be seen that there is no vertical outliers or high leverage points in the data, the value of VIF=VIF** which are less than 5, indicating that there is no multicollinearity problem. It is also interesting to note that our proposed plot (LTSR-HLCIM(DRGP(MVE)) can detect almost all observations as regular observations (DRO) (on the average of 96.6 percent). Furthermore, the results in Table 3a also indicated that in the presence of vertical outliers and in the absence of high leverage points, the data sets do not have multicollinearity problems (VIF**< 5). The results also suggest that the average percentages of detected vertical outliers (DVO) is reasonably close to the number of generated vertical outliers (VO) on the average of 94.8 percent.

Furthermore, when α = 0.00 in Table 3 above, It can be seen that there is no vertical outliers or high leverage points in the data, the value of VIF=VIF** which are less than 5, indicating that there is no multicollinearity problem. It is also interesting to note that our proposed plot (LTSR-HLCIM(DRGP(RFCH)) can detect almost all observations as regular observations (DRO) (on the average of 99.8 percent). Furthermore, the results in Table 3a also indicated that in the presence of vertical outliers and in the absence of high leverage points, the data sets do not have multicollinearity problems (VIF**< 5). The results also suggest that the average percentages of detected vertical outliers (DVO) is reasonably close to the number of generated vertical outliers (VO) on the average of 99.6 percent.

The simulated data without the BLCRO in Table 4 above showed all the VIF values are (>10) indicating presence of multicollinearity. However, in regards to the generated bad leverage collinearity-reducing observations, the VIF** values drastically reduced in the presence of high leverage points. This indicates that high leverage points conceal the problem of multicollinearity. Furthermore, the existing plot LTSR-HLCIM which is based DRGP(MVE) detected the observations as (DBLCRO) on the average of 98.4 percent.

The simulated data without the BLCRO in Table 5 above showed all the VIF values are (>10) indicating presence of multicollinearity. However, in regards to the generated bad leverage collinearity-reducing observations, the VIF** values drastically reduced in the presence of high leverage points. This indicates that high leverage points conceal the problem of multicollinearity. Furthermore, the existing plot LTSR-HLCIM which is based DRGP(RFCH) detected almost all the observations as (DBLCRO) on the average of 99.9 percent.

The simulated data without the GLCRO in Table 6 above showed all the VIF values are (>10) indicating presence of multicollinearity. However, in regards to the generated good leverage collinearity-reducing observations, the VIF** values drastically reduced in the presence of high leverage points. This indicates that high leverage points conceal the problem of multicollinearity. Furthermore, the existing plot LTSR-HLCIM which is based DRGP(MVE) detected the observations as (DGLCRO) on the average of 91.4 percent.

The simulated data without the GLCRO in Table 7 above showed all the VIF values are (>10) indicating presence of multicollinearity. However, in regards to the generated bad leverage collinearity-reducing observations, the VIF** values drastically reduced in the presence of high leverage points. This indicates that high leverage points conceal the problem of multicollinearity. Furthermore, the existing plot LTSR-HLCIM which is based DRGP(RFCH) detected almost all the observations as (DGLCRO) on the average of 98.7 percent.

Conclusion

The purpose of this paper is to assess the merit of the proposed detection method (Improvised LTSR-HLCIM) for the identification of the high leverage collinearity reducing observations and how to reduce the rate masking and swamping effects using fast and robust diagnostic method. The results indicated that the proposed diagnostic measure DRGP(RFCH) can successfully detect almost the multiple high leverage collinearity-reducing observations in the simulation data set. However, two different kinds of regression diagnostic plots namely LTSR-HLCIM and (Improvised LTSR-HLCIM) were used. The merit of our proposed diagnostic plot (Improvised LTSR-HLCIM) is confirmed through applying them with simulation data.

REFERENCES

Andersen, R. (2008). Modern methods for robust regression. The United States of America: Sara Miller McCune. SAGE publications.

Bagheri, A., Habshah, M. and Imon, A. H. M.R. (2012). A novel collinearity - influential observation diagnostic measure based on a group deletion approach. Communications in Statistics-Simulation and Computation, 41, 1379–1396.

Bhageri A., & Habshah, M. (2015). Diagnostic plot for the identification of high leverage collinearity-influential observations. SORT 39 (1) January-June 2015, 51-70.

Devlin, S. J., Gnanadesikan, R., & Kettenring, J. R. (1981). Robust estimation of dispersion matrices and principal components. Journal of the American Statistical Association, 76(374), 354-362.

Habshah M., Hasan T., Jayanthi A., and Hassan U (2020). Fast and Robust Diagnostic Technique for the Detection of High Leverage Points. *Pertanika J. Sci. & Technol. 28 (4): 1203 - 1220 (2020).*

Habshah, M., Norazan, M.R., and Imo, A.H.M.R (2009): The performance of Diagnostic - Robust Generalized Potentials for Identification of Multiple high leverage points in linear regression. *Journal of applied Statistics*. 36(5):507-520.

Montgomery, D. C., Peck, E. A. and Viving, G.G. (2001). Introduction to linear regression Analysis. 3rd edition. New York: John Wiley and sons.

Muhammed, A, Habshah, M and Imon, A.H.M.R (2015). A New Robust Diagnostic Plot for Classifying Good and Bad High Leverage Points in a Multiple Linear Regression Model. *Mathematical Problems in Engineering.* Volume 2015, Article ID 279472, 12 pages.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. National Institute of Science of India, 2(1), 49-55.

Olive, D. J., & Hawkins, D. M. (2008). High breakdown multivariate estimators. Retrieved September 5, 2019, from https://www.researchgate.net/profile/David_Olive2/publication/240737720_High_Breakdown_Multivariate_Estimators/links/0a85e53234b7db7f90000000.pdf.

Olive, D. J., & Hawkins, D. M. (2010). Robust multivariate location and dispersion. Retrieved September 5, 2019, from https://www.researchgate.net/profile/David_Olive2/publication/228434748_Robust_multivariate_location_and_dispersion/links/02bfe51015be5c88ca000000.pdf.

Pearson, K. (1908). On the generalized probable error in multiple normal correlation. Biometrika. 6: 59-68.

Rousseeuw, P.J. and Leroy, A.M. (1987). Robust Regression and Outlier Detection. New York: Wiley.

Rousseeuw, P.J. (1985). Multivariate estimation with high breakdown point. Mathematical and Statistical Applications. B: 283-297.

Rousseeuw, P. J. and Croux, C. (1993). Alternatives to the median absolute values. Journal of the American Statistical Association. 88: 1273-83.

Rousseeuw, P and Van Zomeren, B. (1990). Unmaking multivariate outliers and leverage points. Journal of America Statistical Association, 85, 633-639.

Wonsuk, Y., Robert, M. and James, W. (2014). A study effects of Multicollinearity in the Multivariate Analysis. International Journal of Applied Science and Technology, Vol. 4, No.5; October 2014 pp. 9-19.

Zhang, J., Olive, D., & Ye, P. (2012). Robust covariance matrix estimation with canonical correlation analysis. International Journal of Statistics and Probability, 1(2), 119-136.