IBBUL Lapai Journal of Applied and Natural Sciences

LAJANS Vol 7(1): 1-10

RESEARCH PAPER

The Performance of the Modified Detection Method in the Presence of High Leverage Collinearity Enhancing Observations

[1*]Abdulkarim, K., [2]Rasheed, B.A. and [1]Lasisi, K.E.

[1]Department of Mathematical Sciences, Abubakar Tafawa Balewa University Bauchi, Nigeria
[2]Department of Mathematics and Statistics, Federal Polytechnic, Bauchi, Nigeria

ABSTRACT

The presence of outliers and multicollinearity are inevitable in real data sets and they have an unduly effect on the parameter estimation of multiple linear regression models. It is now evident that outliers in the X-direction or high leverage points are another source of multicollinearity. These leverage points may enhance multicollinearity in a dataset. We call this high leverage collinearity enhancing observations. 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 short comings mentioned in the existing method by using DRGP(RFCH) instead of DRGP(MVE) used by LTSR-HLCIM. We study several criteria such as sample sizes, percentages and position of high leverage points which cause these leverages to change the multicollinearity pattern of collinear data sets. The results of simulation 0073tudy and real dataset indicates that the modified detection method out-performed the existing method in term of having the highest percentage of correct detection of good and bad leverage collinearity enhancing observation.

Keywords: Linear regression, Collinearity enhancing observation, multicollinearity.

INTRODUCTION

Multiple regression was first used by Sir Karl Pearson in 1908 (Pearson, 1908). In statistics, multiple regression analysis includes any techniques for modeling and analyzing the relationship between a response or dependent variable and more than one explanatory variables (also known as regressor, predictor and independent 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 (see Andersen, 2001; Montgomery *et al*., 2001; Gujarati, 2002; Kutner *et al*., 2005; 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.

The existence of a few anomalous points which often referred as outliers is inevitable in real data sets (Hampel et al., 1986). These atypical points may occur in the response variable (y-direction) or regressor variables (X-direction) in multiple regression model which are called vertical (regression) outliers or high leverage points, respectively. Regardless of their sources, outliers can render least squares estimations meaningless (Rousseeuw and Leroy, 1987).

More specifically, the presence of high leverage points, the points far from the rest of the data set in the X-direction, can mislead the conclusion of statistical analysis. Unfortunately, many statisticians are not aware that high leverage points may change the multicollinearity pattern of the data and will lead to misleading inferences.

However, (Bagheri and Habshah, 2015) developed four detection method for detecting High Leverage Collinearity enhancing observations. The LTSR-HLCIM successfully among the existing methods classified vertical

Received 12 August, 2022
Accepted 1 November, 2022
Address Correspondence to:
kamlakmashi@gmail.com

outliers, good leverage collinearity-enhancing observations, bad leverage collinearity-enhancing observations, vertical outliers and regular observations.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 an (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 Hawkins BraduKass Data (Hawkins et al., 1984) below in Table 1 consists of 75 observations and 3 independent variables. Hawkins et al., (1984) claimed that this data set had 14 HLPs. Bagheri et al. (2012) demonstrated that the first 14 observations in the original data as displayed in Table 1, caused multicollinearity evident by showing the value of Variance Inflation Factor (VIF=33.342 for X3) which is greater than 10. On the contrary, no multicollinearity was observed in their absence. However, this data is used to assess the proposed plot LTSR-HLCIM(DRGP(RFCH) over the existing method LTSR-HLCIM using the diagnostics plot..

Table 1: Multicollinearity Diagnostics for HawkinBraduKass Data

Diagnostic	Status	**1**	**2**	**3**
VIF > 5	Original data	13.432	23.853	33.432
	Without observations 1-14	1.012	1.017	1.027

Diagnostic Robust Generalized Potential based on (Minimum Volume Ellipsoid)

The commonly used diagnostic methods fail to identify high leverage points correctly when a group of high leverage points is present in a data set due to the masking and/or swamping effects (Peña and Yohai, 1995; Rousseeuw and Leroy, 1987). Rousseeuw (1985) introduced Robust Mahalanobis Distance (RMD), based on the Minimum Volume Ellipsoid (MVE). The DRGP algorithm comprises two steps where;

Step 1: Robust Mahalanobis Distance (RMD) based on MVE is used to detect the suspected HLPs (bad cases). (Mahalanobis, 1936) defined Mahalanobis Distance (MD) as a measure of deviation of a data point from its center. 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.

Subsequently, the (MD) for each observation is written as Equation 1:

$$MD_i = \sqrt{(t_i - T(X))' C(X)^{-1} (t_i - T(X))} i = 1,2, \dots, n \quad (1)$$

Where T(X) is the mean vector and C(X) is the variance covariance matrix (C). (Rousseeuw and Leroy, 1987) suggested using Robust Mahalanobis Distance (RMD) as a diagnostic tool for detection of HLPs by replacing the classical mean vector T(X), and classical covariance matrix, C(X) of *MDi* in Equation 1 by robust estimators such as Minimum Volume Ellipsoid (MVE) or Minimum Covariance Determinant (MCD) (Rousseeuw&Yohai, 1984), because the former estimators are not robust.

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

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 2: 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.

$$DRGP = (p_{ii}^{*}) = \begin{cases} h_{ii}^{(-D)} & for\ i \in D \\ \dfrac{h_{ii}^{(-D)}}{1 - h_{ii}^{(-D)}} & for\ i \in R \end{cases} \quad (3)$$

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 1: Identify the suspected HLPs by using *RMDi* for each i[th] observation based on RFCH (Equation 1):

RMDi=

$$\sqrt{(t_i - T_{RFCH}(X))' C_{RFCH}(X)^{-1}(t_i - T_{RFCH}(X))} i = 1,2,\dots,n \quad (4)$$

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 2: The DRGP $(p_{ii}^{*}) = \begin{cases} h_i^{(-D)} & for\ i \in D \\ \dfrac{h_i^{(-D)}}{1-h_i^{(-D)}} & for\ i \in R \end{cases} \quad (5)$

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

The main reason of developing a new measure of high leverage collinearity-influential measure is due to the fact that the commonly used measures failed to detect multiple high leverage collinearity influential observations.

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:

$$\delta_i^{(D)} = \begin{cases} log\dfrac{k_{(D)}}{k_{(D-i)}} & if\ i \in D, \quad num\{D\} \neq 1 \\ log\dfrac{k_{(i)}}{k} & if\ num\{D\} = 1,\ D = i,\ i = 1,2,\dots,n \\ log\dfrac{k_{(D+i)}}{k_{(D)}} & if\ i \in R, \end{cases} \quad (6)$$

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 1. Calculate DRGP(MVE) and DRGP(RFCH).

Step 2. 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 3. 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 Cut-off 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 Cutoff Point of Least Trimmed Square Residuals

The LTS is used for detection purposes, according to (Mohammed et al., 2015) the standardized least trimmed square residuals (LSTR) which is the horizontal cutoff point in both panels with 97.5% quantile of the chi-square distribution and 1 degree of freedom which is given as$(\pm\sqrt{\chi^2_{1,\ 0.975}})$. It can be further explain that when α=0.05, the significance cutoff of a chi-square distribution with 1 *df* is approximately 3.84, and that if we take the square root of 3.84, it's approximately 1.96, which is the 97.5 percentile.

The Existing 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 proposed (Improvised LTSR-HLCIM)

The 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 $\delta_i^{(D)}$.

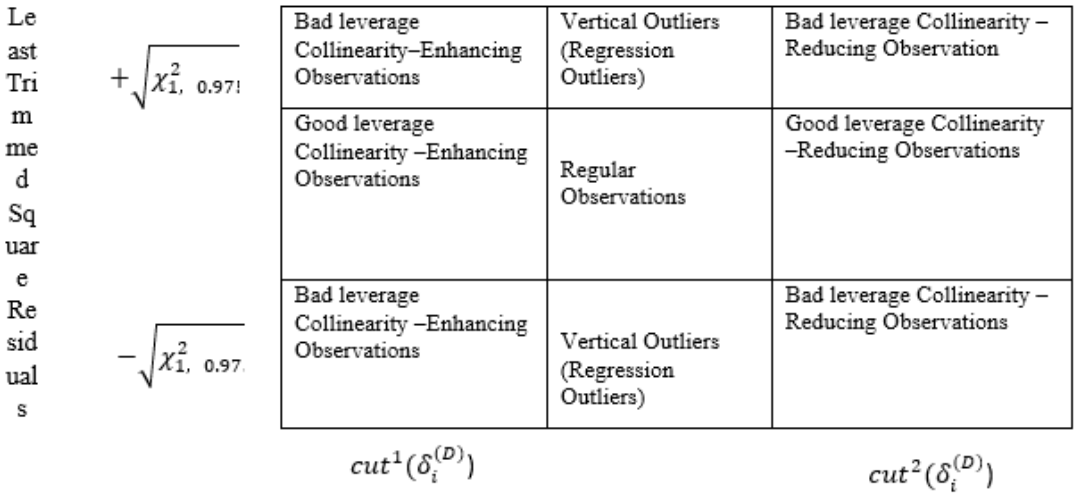

Figure 1: The Ballentine View of LTRS-HLCIM Plot.

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).

RESULTS AND DISCUSSION

Numerical study

In this section, the performance of the proposed and the existing methods, namely the (Improvised LTSR-HLCIM) and LSTR-HLCIM respectively are to be investigated through the usage of Hawkins-Bradu-Kass data set. The data set is taken from (Hawkins, Bradu, and Kass, 1984) and also available in robustbase package. The Hawkins-Bradu-Kass data set is constructed to have ten bad leverage points (cases 1−10) and four good leverage points (cases 11−14) (Habshah *et al.,* 2009). Hawkins-Bradu-Kass data is an artificial data set with ten bad leverage points (cases 1–10) and four good leverage points (cases 11–14) (Habshah *et al.*, 2009). Figure 1 presents LTSR-HLCIM diagnostic plot for Hawkins-Bradu-Kass data set. According to this figure part (a), cases 11-14 not only are good leverage points but also are good leverage collinearity-enhancing observations. Moreover, cases 1-10 are bad leverage points and bad leverage collinearity-enhancing observations. However, Figure 2 presents diagnostic plot for Hawkins-Bradu-Kass data set without the first 14 observations. It can be observed from Figure 1 part (b), that this data set does not have vertical outlier and high leverage collinearity enhancing observation. However (case 53) was masked in the presence of the 14 observations.

Kass data set. According to this figure part (c), cases 11-14 not only are good leverage points but also are good leverage collinearity-enhancing observations and cases 1-10 are bad leverage points and bad leverage collinearity-enhancing observations. Moreover, (case 53) was unmasked by the LTSR-HLCIM(DRGP(RFCH)). However, Figure 3 also presents diagnostic plot for Hawkins-Bradu-Kass data set without the first 14 observations. It can be observed from Figure 3 part (d), that this data set does not have vertical outlier or Collinearity enhancing observations.

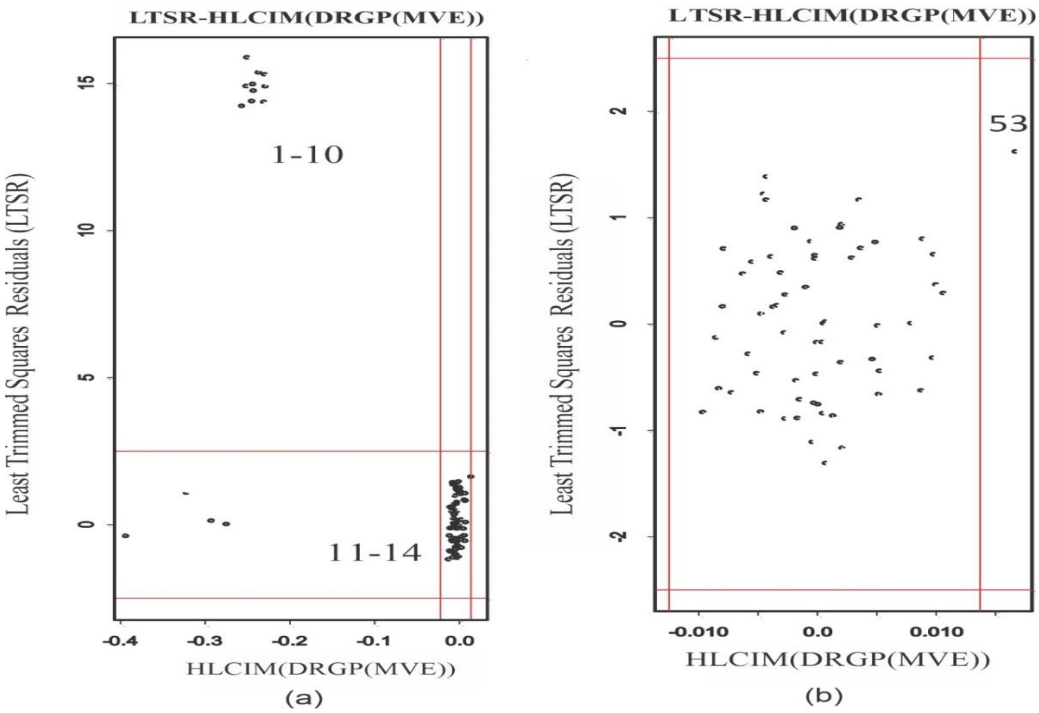

Figure 2: Diagnostic plot of (a) Hawkins-Bradu-Kass data set and (b) Hawkins-Bradu-Kass data set without the first 14 observations.

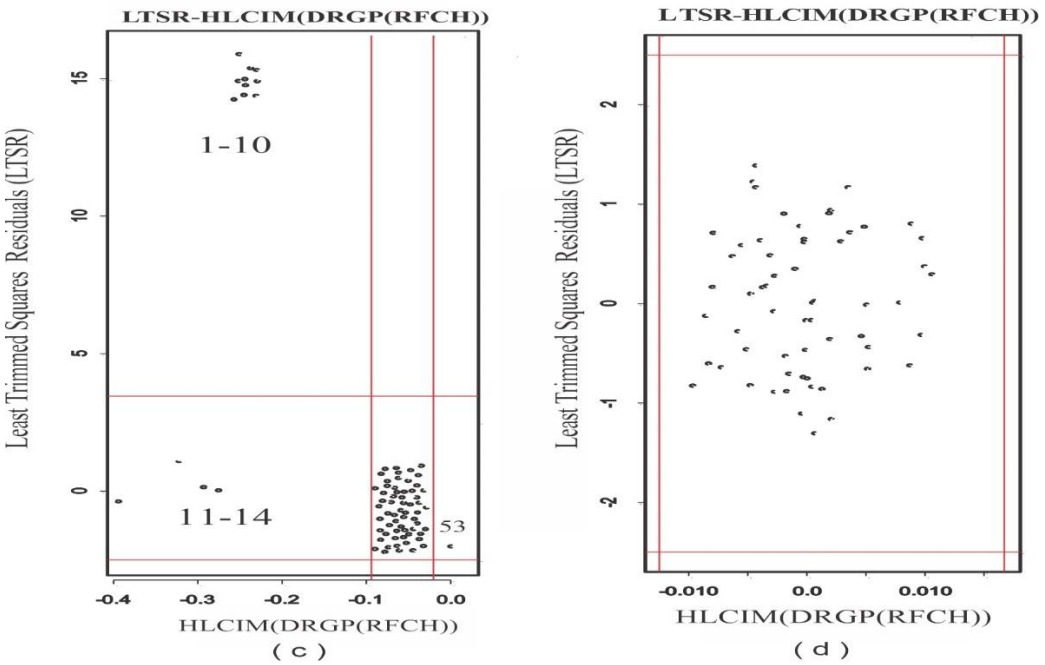

Figure 3: Diagnostic plot of (c) Hawkins-Bradu-Kass data set and (d) Hawkins-Bradu-Kass data set without the first 14 observations.

Monte Carlo simulation study

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 enhancing observations. To achieve this aim, 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-enhancing 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.

However according (Bagheri and Habshah, 2015), to generate high leverage collinearity-enhancing observations, each variable will firstly generated from Uniform $(0,1)$ to produce non-collinear data sets. This generated data is referred to as the regular observations. The last 100%α observations of the regular observations of each regressor will then replace with certain percentage of high leverage points to create high leverage collinearity-enhancing observations. To generate the high leverage points as collinearity-enhancing observations in non-collinear data sets, the values corresponding to the first high leverage point will be kept fixed at 10 and those of the successive values are created by multiplying the observations index, i, by 10. The Good leverage Collinearity-Enhancing Observation (GLCEO) was created in such a way that the High leverage Collinearity-Enhancing Observation (HLCEO) is generated without any vertical outlier. On the other hand, Bad leverage Collinearity Enhancing Observation (BLCEO) was created when both HLCEO and The simulated data without the BLCEO in Table 3a above showed non-existence of multicollinearity since all the VIF values are (<5) and as for the generated bad leverage collinearity-enhancing observations (BLCEO), the VIF* values (> 10) drastically increased in the presence of high leverage points. This indicates that high leverage points are the cause of multicollinearity. However, the existing plot LTSR-HLCIM which is based DRGP(MVE) detected the observations as (DBLCEO) on the average of 91.1 percent.

Moreover, the simulated data without the BLCEO in Table 3b also showed all the VIF values (<5) indicating none-presence of multicollinearity and the generated bad leverage collinearity-enhancing observations indicated all the VIF* values (> 10) drastically increased in the presence of high leverage points. This indicates that high leverage points are the cause of multicollinearity. It is also interesting to note that the proposed plot LTSR-HLCIM which is based DRGP(RFCH) detected almost all observations as (DBLCEO) on the average of 99.4 percent.

The simulated data without the GLCEO in Table 4a above showed all the VIF values (<5) indicating none-presence of multicollinearity and as for the generated good leverage collinearity-enhancing observations, the VIF* values (> 10) drastically increased in the presence of high leverage points. This indicates that high leverage points are the cause of multicollinearity. Furthermore, the existing plot LTSR-HLCIM which is based DRGP(MVE) detected the observations as (DGLCEO) on the average of 96.4 percent.

Moreover, the simulated data without the GLCEO in Table 4b above showed non-existence of multicollinearity since all the VIF values (<5) and as for the generated good leverage collinearity-enhancing observations of (0.05, 0.10 and 0.15), the VIF* values (> 10) drastically increased in the presence of high leverage points. This indicates that high leverage points are the cause of multicollinearity. It is also interesting to note that the existing plot LTSR-HLCIM which is based on DRGP(RFCH) detected almost all the as (DGLCEO) on the average of 99.6 percent.

Table 2: 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
BLCEO	bad leverage collinearity –enhancing observations
DBLCEO	The number of detected bad leverage collinearity –enhancing observations
GLCEO	Good leverage collinearity –enhancing observations
DGLCEO	The number of detected good leverage collinearity–enhancing observations

Table 3a: The number of abnormal observations in the simulated data sets with bad leverage collinearity enhancing observations using LTSR-HLCIM(DRGP(MVE)).

(n)		40			100			300	
∝	0.05	0.10	0.15	0.05	0.10	0.15	0.05	0.10	0.15
VIF	3.56	3.41	3.38	3.29	3.18	3.10	3.30	3.29	3.27
VIF*	23.68	60.09	107.56	166.13	131.70	375.29	704.68	1107.26	1210.35
RO	38	36	34	95	90	85	285	270	255
BLCEO	2	4	6	5	10	15	15	30	45
DBLCEO	1.74	3.91	5.95	4.10	9.04	14.5	12	28	42
%	*87%*	*97.8%*	*99.2%*	*82%*	*90.4%*	*96.7%*	*80%*	*93.3%*	*93.3%*

Table 3b: The number of abnormal observations in the simulated data sets with bad leverage collinearity enhancing 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	3.56	3.41	3.38	3.29	3.18	3.10	3.30	3.29	3.27
VIF*	23.70	61.00	108.56	167.14	133.67	380.37	710.69	1118.2	1256.5
RO	38	36	34	95	90	85	285	270	255
BLCEO	2	4	6	5	10	15	15	30	45
DBLCEO	1.99	3.99	5.97	4.98	9.89	14.59	15	30	45
%	*99.5%*	*99.8%*	*99.5%*	*99.6%*	*98.9%*	*97.3%*	*100%*	*100%*	*100%*

Table 4a: The number of abnormal observations in the simulated data sets with good leverage collinearity enhancing observations using LTSR-HLCIM(DRGP(MVE)).

(n)		40			100			300	
∝	0.05	0.10	0.15	0.05	0.10	0.15	0.05	0.10	0.15
VIF	3.46	3.50	3.13	3.64	3.29	3.10	3.30	3.29	3.27
VIF*	23.69	61.92	107.60	166.05	132.15	175.29	377.64	704.52	1107.30
RO	38	36	34	95	90	85	285	270	255
GLCEO	2	4	6	5	10	15	15	30	45
DGLCEO	1.74	4.00	5.95	4.90	9.89	14.5	14	29	44
%	*87%*	*100%*	*99.2%*	*98%*	*98.9%*	*96.7%*	*93.3%*	*96.7%*	*97.8%*

Table 4b: The number of abnormal observations in the simulated data sets with good leverage collinearity enhancing 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	3.46	3.50	3.13	3.64	3.29	3.10	3.30	3.29	3.27
VIF*	23.90	62.90	108.60	167.08	130.79	175.03	378.60	707.5	1107.87
RO	38	36	34	95	90	85	285	270	255
GLCEO	2	4	6	5	10	15	15	30	45
DGLCEO	2.00	4.00	5.99	4.98	10	14.99	15	30	45
%	*100%*	*100%*	*99.8%*	*99.6%*	*100%*	*97%*	*100%*	*100%*	*100%*

Conclusion

The main aim of this paper is to determine the high leverage collinearity enhancing observations and the rate masking and swamping effects using simulation and a real data set. The results indicated that the proposed measureDRGP(RFCH) can successfully detect almost the multiple high leveragecollinearity-enhancing observations in the data set and it also reduced the rate of Masking effect. Finally and importantly, two different kinds of regression diagnostic plots namely LTSR-HLCIM and Improvised LTSR-HLCIM were used. The merit of our proposed diagnostic plot is confirmed through applying them to different real data set and simulation study.

REFERENCES

Anderson, C. (2001). A Comparison of Five Robust Regression Methods with Ordinary Least Squares: Relative Efficiency, Bias, and Test of the Null Hypothesis, Unpublished Ph.D. thesis, University of North Texas, U.S.A.

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.

Gujarati, D.N. (2002). Basic Econometrics. 4th edition. New York: Macgraw-Hill.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). Robust Statistics. New York: Wiley.

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).*

Hawkins, D.M., Bradu, D. and Kass, G.V.(1984). Location of several outliers in multiple regression data using elemental sets. Technometrics. 26:197-208.

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.

Kutner, M.H., Nachtsheim, C.J., Neter, J. and Li, W. (2005). Applied Linear Regression Models. 5th edition. New York: MacGRAW-Hill.

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.

Peña, D. and Yohai, V.J. (1995). The detection of influential subsets in linear regression by using an influence matrix. Journal of Royal Statistical Society. B 57: 18–44.

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 Yohai, V. (1984). Robust regression by means of Sestimators, Robust and Nonlinear Time series Analysis. Lecture Notes in Statistics. 26: 256-272.

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.