SORT 39 (1) January-June 2015, 51-70

Diagnostic plot for the identification of high leverage collinearity-influential observations

Arezoo Bagheri[1] and Habshah Midi[2]

Abstract

High leverage collinearity influential observations are those high leverage points that change the multicollinearity pattern of a data. It is imperative to identify these points as they are responsible for misleading inferences on the fitting of a regression model. Moreover, identifying these observations may help statistics practitioners to solve the problem of multicollinearity, which is caused by high leverage points. A diagnostic plot is very useful for practitioners to quickly capture abnormalities in a data. In this paper, we propose new diagnostic plots to identify high leverage collinearity influential observations. The merit of our proposed diagnostic plots is confirmed by some well-known examples and Monte Carlo simulations.

MSC: 62-09, 62G35, 62J05, 62J20

Keywords: Collinearity influential observation, diagnostic robust generalized potential, high leverage points, multicollinearity.

1. Introduction

Multicollinearity is an exact or a near linear relationship among regressors in a multiple linear regression. According to Kamruzzaman and Imon (2002), high leverage points or observations that fall far from the majority of independent variables in a data set, are a prime source of multicollinearity. Hadi (1988) pointed out that this source of multicollinearity is a special case in collinearity-influential observations, which may change the multicollinearity pattern of data. They are referred to as high leverage collinearity-enhancing observations or high leverage collinearity-reducing observations (Habshah

Corresponding author: Habshah Midi

[1] National Population Studies & Comprehensive Management Institute, No. 3, 5th Street, Miramad Street, Motahari Street, Tehran, Iran. abagheri_000@yahoo.com

[2] Department of Mathematics, Faculty of Science/Institute for Mathematical Research, Universiti Putra Malaysia, 43400 Serdang , Selangor, Malaysia. habshah@upm.edu.my

Received: November 2013
Accepted: March 2015

et al., 2010; Habshah et al., 2011; Bagheri et al., 2012).With their presence, multiple linear regression models encounter serious problems (Habshah et al., 2009; Bagheri et al., 2009; Bagheri and Habshah, 2008). Hence it is very important to detect them so that appropriate steps can be taken to remedy such problems (Bagheri and Habshah, 2012-2011; Habshah et al., 2010).

Simple scatter plots are very useful in exploring the relationship between a response and a single explanatory variable as well as in detecting outliers. They are, however, ineffective in revealing the complex relationships or detecting the trend and data problems in multiple regression models. Partial plots, on the other hand, may be better substitutes for scatter plots in a multiple linear regression. This is because these plots illustrate the partial effects or the effects of a given predictor variable after adjusting for all the other predictor variables in a regression model.

There are two different kinds of partial plots, namely the partial residual and the partial regression or added variable plot (See partial plots in Myers, 1990 and also leverage plots in Sall, 1990; Leverage-Residual Plot of Gray, 1983) which are documented in the literature (Belsley et al., 1980; Cook and Weisberg, 1982). However, partial residual and partial regression plots are generally unable to detect multicollinearity. Overlaying both the partial residual and partial regression plots on the same plot, with the centered xi values on the x-axis, may in fact provide an alternative method to detect multicollinearity (Stine, 1995) by highlighting the amount of shrinkage in partial regression residuals. However, when high leverage points are the source of multicollinearity, these plots will be affected and as a result they will no longer be useful for diagnosing multicollinearity in a data set.

Unfortunately, to the best of our knowledge, we have not found any paper in the literature that establishes graphical methods for the identification of multicollinearity due to high leverage points. This gap in the literature has motivated us to propose appropriate plots that are able to classify observations according to regular observations, high leverage points, collinearity-influential observations and vertical outliers.

These plots will be examined in this paper which is organized into five sections. The next section, Section 2, reviews High Leverage Collinearity-Influential Measure (HLCIM) based on Diagnostic-Robust Generalized Potential (DRGP) which is referred to in this paper as HLCIM(DRGP). Section 3 introduces the newly proposed high leverage collinearity-influential observation regression diagnostic plots. Section 4 discusses both the performance of our proposed plots by using some real data sets and their merit according to Monte Carlo simulations. Finally, some concluding remarks are presented in Section 5.

2. Literature review

In the following section, high leverage collinearity-influential measure based on DRGP will be discussed. Firstly, the regression model can be defined as the following equation:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{1}$$

where $\boldsymbol{Y}$ is an $(n \times 1)$ vector of response or the dependent variable, $\boldsymbol{X}$ is an $(n \times p)$ matrix of predictors $(p \times 1)$, $\boldsymbol{\beta}$ is $(p \times 1)$ vector of unknown finite parameters to be estimated and ε is an $(n \times 1)$ vector of random errors. We allow $\boldsymbol{X}_j$ to denote the j^{th} column of the $\boldsymbol{X}$ matrix; therefore, $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_p]$. Additionally, we define multicollinearity in terms of the linear dependence of the columns of $\boldsymbol{X}$; thus, the vectors of $\boldsymbol{X}_1, \boldsymbol{X}_2, \dots, \boldsymbol{X}_p$ are linearly dependent if there is a set of constants $t_1, t_2, \dots, t_p$ that are not all zero, such as $\sum_{j=1}^{p} t_j \boldsymbol{X}_j = 0$. The problem of multicollinearity is said to exist when this equation holds approximately $\sum_{j=1}^{p} t_j \boldsymbol{X}_j \approx 0$.

Since multicollinearity is a problem that exists in a data set, there is no statistical test for its presence. Nonetheless, a statistical test can be substituted by a diagnostic method in order to indicate the existence and extent of multicollinearity in a data set. Belsley et al. (1980) proposed an approach for diagnosing multicollinearity based on a singular-value decomposition of a $(n \times p)$ $\boldsymbol{X}$ matrix as:

$$\boldsymbol{X} = \boldsymbol{UVD}' \tag{2}$$

where $\boldsymbol{U}$ is the $(n \times p)$ matrix in which the columns that are associated with the p non-zero eigenvalue of $(\boldsymbol{X}'\boldsymbol{X})$ is $(n \times p)$, $\boldsymbol{V}$ (the matrix of eigenvectors of $\boldsymbol{X}'\boldsymbol{X}$) is $(p \times p)$, $\boldsymbol{U}'\boldsymbol{U} = \boldsymbol{I}$, $\boldsymbol{V}'\boldsymbol{V} = \boldsymbol{I}$, and D is a $(p \times p)$ diagonal matrix with non-negative diagonal elements, k_j, $j = 1, 2, \dots, p$, which is called the singular-values of $\boldsymbol{X}$. The j^{th} Condition Index (CI) of the $\boldsymbol{X}$ matrix is defined as:

$$k_j = \frac{\lambda_{max}}{\lambda_i}, \quad j = 1, 2, \dots, p, \tag{3}$$

where $\lambda_1, \lambda_2, \dots, \lambda_p$ are the singular values of the $\boldsymbol{X}$ matrix. The largest value of k_j is defined as the Condition Number (CN) of the $\boldsymbol{X}$ matrix. Belsley (1991) stated that an $\boldsymbol{X}$ matrix between 10 and 30 indicates a moderate to strong multicollinearity, whereas a value of more than 30 reflects severe multicollinearity.

As previously mentioned, high leverage collinearity-influential observations are those observations that may disrupt the multicollinearity pattern of a data. Unfortunately, not many studies relevant to these issues are found in the literature. Hadi(1988) noted that not all high leverage points are collinearity-influential observations, but most collinearity-influential observations are points with high leverages. He proposed a measure for the identification of high leverage collinearity-influential observations based on the influence of the i^{th} row of $\boldsymbol{X}$ matrix on the condition index as:

$$\delta_i = \log \frac{k_{(i)} - k}{k}, \quad i = 1, 2, \ldots, n, \tag{4}$$

where $k_{(i)}$ is the eigenvalue of $\boldsymbol{X}_{(i)}$ when the i^{th} row of $\boldsymbol{X}$ matrix has been deleted. He pointed out that a large negative value of δ_i indicates that the i^{th} observation is a collinearity-enhancing observation, while a large positive δ_i value indicates a collinearity-reducing observation. Sengupta and Behimasankaram(1997) suggested a more preferable measure to Hadi's measure (Hadi ,1988) which is defined as follows:

$$l_i = \log \frac{k_{(i)}}{k}, \quad i = 1, 2, \ldots, n, \tag{5}$$

According to Bagheri et al. (2012), the performance of both δ_i and l_i is only good for the detection of a single high leverage collinearity influential observation. Moreover, there are some drawbacks in using δ_i or l_i because there are no given specific cutoff points to indicate which observations are collinearity-enhancing and which are collinearity-reducing. To rectify these problems, Bagheri et al. (2012) and Bagheri and Habshah (2012) proposed a high leverage collinearity-influential measure, namely HLCIM (DRGP), denoted as $\delta_i^{(D)}$ and which is defined as follows:

$$\delta_i^{(D)} = \begin{cases} \log \frac{k_{(D)}}{k_{(D-i)}} & \text{if} \quad i \in D \, and \, \neq \{D\} \neq 1 \\ \log \frac{k_{(i)}}{k} & \text{if} \quad \neq \{D\} \, and \, D = i, \; i = , 2, .., n \\ \log \frac{k_{(D+i)}}{k_{(D)}} & \text{if} \quad i \in R \end{cases} \tag{6}$$

where D is the suspected group of multiple high leverage points and R is the remaining good observations diagnosed by DRGP based on Minimum Volume Ellipsoid (MVE) (Habshah et al., 2009). The number of elements in the D group is denoted as $\neq \{D\}$. $k_{(i)}$ indicates the condition number of the $\boldsymbol{X}$ matrix without the i^{th} high leverage points. $k_{(D-i)}$ indicates the condition number of the $\boldsymbol{X}$ matrix without the entire D group minus the i_{th} high leverage points where i belongs to the suspected D group. $k_{(D+i)}$ refers to the condition number of the $\boldsymbol{X}$ matrix without the entire D group of high leverage points plus the i_{th} additional observation of the remaining group (For more information on high leverage diagnostic measures, please refer to Hadi, 1992 and Imon, 2002).

Bagheri et al. (2012) and Bagheri and Habshah(2012) proposed some cutoff points for $\theta_i, \quad i = 1, 2, \ldots, n$:

$$\text{cut}^1(\theta) = \text{Median}(\theta_i) - c\text{Mad}(\theta_i) \tag{7}$$

$$\mathrm{cut}^2(\theta) = \mathrm{Median}(\theta_i) + c\mathrm{Mad}(\theta_i) \quad (8)$$

where $\mathrm{cut}^1(\theta)$ is the cutoff point for collinearity-enhancing measure and $\mathrm{cut}^1(\theta)$ is the collinearity-reducing measure cutoff point. Median and Mean Absolute Deviation (MAD) stand for robust measures of central tendency and dispersion, respectively. θ_i can be δ_i , l_i , or $\delta_i^{(D)}$ and c is the chosen constant value of $3.|\theta_i| \geq |\mathrm{cut}^1(\theta)|$ for $\theta_i < 0$ and $\theta_i \geq \mathrm{cut}^2(\theta)$ for $\theta_i > 0$ is an indicator that the i_{th} observation is a high leverage collinearity-enhancing or -reducing observation, respectively.

Bagheri et al. (2012) pointed out that $\delta_i^{(D)}$ values which exceed the cutoff point and belong to the D groups are called high leverage collinearity-influential observations. On the other hand, those $\delta_i^{(D)}$ which exceed the cutoff point and belong to the R group are called collinearity-influential observations. Since the existence of these points have unduly effects on the parameter estimates, it is imperative to quickly identify them by using diagnostic plots. In this regard, new diagnostic plots to separate high leverage collinearity-influential observations from collinearity-influential observations are proposed.

3. Proposed diagnostic plots

Identifying outliers and high leverage points is a fundamental step in the least squares regression model building process. The usage of graphical tools is one of the easiest ways to quickly capture abnormal points in a data set. Rousseeuw and Van Zomeren (1990) proposed the usage of diagnostic plots and referred to them as an outlier map to classify observations into four types of data points, namely regular observations, good leverage points, vertical outliers and bad leverage points. The proposed outlier map plots the standardized residual ($\frac{r_i}{\hat{\sigma}_i}$,for $i = 1,2,\ldots,n$) versus Squared Robust Mahalanobis Distance based on (MVE)(RMD^2(MVE)) or Squared Robust Mahalanobis Distance based on Minimum Covariance Determinant (RMD^2(MCD)). The disadvantage of this plot is that it uses robust distance which has the tendency to declare more observations as high leverage points due to swamping effects (Habshah et al., 2009). Since robust distance fails to accurately identify high leverage points correctly while the DRGP is able to successfully identify their presence, in this paper we suggest the usage of DRGP in the construction of our proposed diagnostic plots.

The first proposed plot is similar to the outlier map of Rousseeuw and Van Zomeren (1990), except that the robust distance is substituted with the DRGP. As suggested by Rousseeuwand Van Zomeren (1990), the standardized Least Trimmed Squares Residuals (LTSR) residuals are plotted on the $\boldsymbol{Y}$-axis. We name the first proposed plot the LTSR-DRGP plot. First, each of the LTS residuals, r_i for $i = 1,2,..,n$, is standardized by $\hat{\sigma}$. The LTSR -DRGP plots the standardized LTS residuals against the DRGP. In the LTSR -DRGP plot, any observation which exceeds the $\boldsymbol{Y}$-axis boundaries

$(\pm\sqrt{X^2_{1,0.975}})$ is called a vertical outlier while any that exceeds the $\boldsymbol{X}$-axis boundaries $(\text{Median}(p^*_{ii}) + c\text{Mad}(p^*_{ii})$ where p^*_{ii} is the value of DRGP (Habshah et al., 2009) is called a good leverage point. When an observation exceeds both the y-axis and the x-axis boundaries, it is called a bad leverage point.

The second proposed plot is based on the newly developed diagnostic measure for the identification of multiple high leverage collinearity-influential observations, HLCIM(DRGP), denoted as $\delta_i^{(D)}$ as presented in Equation (6). We name this plot the DRGP-HLCIM plot. It plots the DRGP against the High Leverage Collinearity-influential Measure.

The third proposed plot is also based on HLCIM(DRGP). This plot is called the LTSR -HLCIM plot. In this plot, the Standardized LTS Residuals are plotted against the High Leverage Collinearity-influential Measure. Figures 1, 2 and 3 show the Venn diagram or Ballentine view of the LTSR-DRGP, the DRGP-HLCIM, and the LTSR-HLCIM plots, respectively. It is important to note that the proposed cutoff points are as follows:

$$\text{cut}^1(P^*_{ii}) = \text{Median}(P^*_{ii}) + c\text{Mad}(P^*_{ii}) \tag{9}$$

where P^*_{ii} is the DRGP. If the proposed $\delta_i^{(D)}$ in Equation 6 is employed, then $\text{cut}^1\left(\delta_i^{(D)}\right)$ and $\text{cut}^2\left(\delta_i^{(D)}\right)$ from Equations 7 and 8 are the cutoff points for detecting high leverage collinearity-enhancing and -reducing observations, respectively.

Figure 1 separates the data set into groups of regular observations, vertical (or regression) outliers, and good or bad leverage points. The figure groups the data set according to whether the observation is a high leverage point and/or a vertical outlier. Nevertheless, it does not take into consideration the multicollinearity pattern of a data set.

Figure 2 groups the data set according to whether the observation is a high leverage point or a collinearity-influential observation. Hence, it classifies the data set into groups

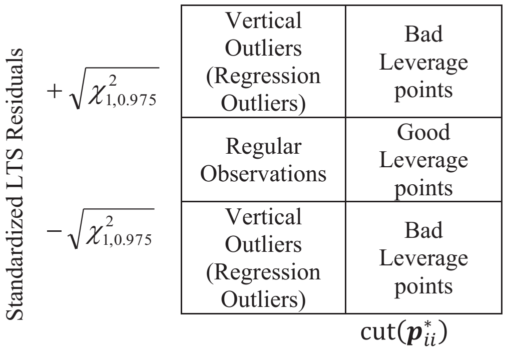
Standardized LTS Residuals
$+\sqrt{\chi^2_{1,0.975}}$
$-\sqrt{\chi^2_{1,0.975}}$
Vertical Outliers (Regression Outliers)
Bad Leverage points
Regular Observations
Good Leverage points
Vertical Outliers (Regression Outliers)
Bad Leverage points
$\text{cut}(\boldsymbol{p}^*_{ii})$

Figure 1: *The Venn Diagram or Ballentine View of LTSR-DRGP Plot.*

DRGP(MVE)

	High leverage Collinearity –Enhancing Observations	High leverage points	High leverage Collinearity –Reducing Observations
$\text{cut}(\boldsymbol{p}_{ii}^{*})$	Collinearity –Enhancing Observations	Regular Observations	Collinearity – Reducing Observations

$\text{cut}^1(\delta_i^{(D)})$ $\text{cut}^2(\delta_i^{(D)})$

HLCIM(DRGP)

Figure 2: *The Venn Diagram or Ballentine View of DRGP-HLCIM Plot.*

Standardized LTS Residuals

	Bad leverage Collinearity –Enhancing Observations or Collinearity –Enhancing Observations with large residual	Vertical Outliers (Regression Outliers)	Bad leverage Collinearity –Reducing Observations or Collinearity –Reducing Observations with large residual
$+\sqrt{\chi^2_{1,0.975}}$	Good leverage Collinearity –Enhancing Observations or Collinearity –Enhancing Observations	Regular Observations	Good leverage Collinearity –Reducing Observations or Collinearity –Reducing Observations
$-\sqrt{\chi^2_{1,0.975}}$	Bad leverage Collinearity –Enhancing Observations or Collinearity –Enhancing Observations with large residual	Vertical Outliers (Regression Outliers)	Bad leverage Collinearity –Reducing Observations or Collinearity –Reducing Observations with large residual

$\text{cut}^1(\delta_i^{(D)})$ $\text{cut}^2(\delta_i^{(D)})$

Figure 3: *The Venn Diagram or Ballentine View of LTRS-HLCIM Plot.*

of regular observations, high leverage points, high leverage collinearity-enhancing/reducing observations, and collinearity-enhancing/reducing observations.

This figure also does not take into consideration whether the observation is abnormal in the$\boldsymbol{Y}$-direction. Finally, Figure 3 classifies the data as regular observations, vertical outliers, good leverage collinearity-enhancing/reducing observations, collinearity-enhancing/reducing observations, bad leverage collinearity-enhancing/reducing observations as well as collinearity-enhancing/reducing observations with large residuals. One of the interesting features of this figure is that it takes into account the good leverage points which are also collinearity-influential observations. Most statisticians believe that good leverage points are not problematic since they are in the same fitted regression line as the other data set and they decrease the standard error of the parameter estimations because they increase the variability of $\boldsymbol{X}$ (see for instance Moller et al., 2005; Andersen, 2008). However, these points maybe collinearity-influential observations and like bad leverage points, they may be destructive to the regression analysis. A joint DRGP-HLCIM and LTSR-HLCIM plot can give a clearer view of the outlyingness of any points in the $\boldsymbol{X}$-direction or $\boldsymbol{Y}$-direction as well as the multicollinearity pattern of a data set. In the following section, the performance of our proposed diagnostic plots is measured by applying these plots to influential cases with authentic and well-known data sets.

4. Results and discussion

Numerical and Monte Carlo simulation results will be discussed in the following sub sections.

4.1. Numerical results

In this section, the performance of the proposed diagnostic plots, namely the LTSR-DRGP, the DRGP-HLCIM, and the LTSR-HLCIM are investigated through the usage of some commonly referred data sets such as the Hawkins-Bradu-Kass data, Commercial Properties data and Body Fat data sets. The first data set is taken from Hawkins, Bradu, and Kass(1984) while the second and third are taken from Kutner et al.(2005).

The Hawkins-Bradu-Kass data set is constructed to have ten bad leverage points $(cases\ 1-10)$ and four good leverage points $(cases\ 11-14)$ (Rousseeuw and Leroy, 1987; Habshah et al., 2009; Bagheri et al., 2012). Figure 4 presents the proposed diagnostic plots for the Hawkins-Bradu-Kass data set. According to parts (a) and (c) of this figure, cases $11-14$ are not only good leverage points but are also good leverage collinearity-enhancing observations. Moreover, cases $1-10$ are bad leverage points and bad leverage collinearity-enhancing observations. It is important to mention that cases 1-14 are all high leverage collinearity-enhancing observations (Figure 4, part (b)). Also, it is worth noting that even though cases $11-14$ are good leverage points, they are collinearity-enhancing observations. Hence, more attention is needed in the estimation of their parameters.

Figure 5 presents the diagnostic plots for the Hawkins-Bradu-Kass data set without the first 14 observations. It can be observed from parts (a) and (b) of Figure 5 that this data set does not have any vertical outliers nor any high leverage points. Nonetheless, it has one collinearity-reducing observation (case 53) which was masked in the presence of the first 14 observations.

Diagnostic plots for the original and modified Commercial Properties data set are presented in Figures 6 and 7, respectively. The original data set has 19 high leverage points (observations 1, 2, 3, 6, 7, 8, 17, 21, 26, 29, 37, 45, 53, 54, 58, 61, 62, 72 and 79) with only two (cases 6 and 62) bad leverage points (Figure 6 part (a)). Moreover, cases 9, 63, 64, 65, and 68 are vertical outliers. There are no high leverage collinearity-enhancing observations in this data set (Figure 6 part (b)). Parts(b) and (c) of Figure 6

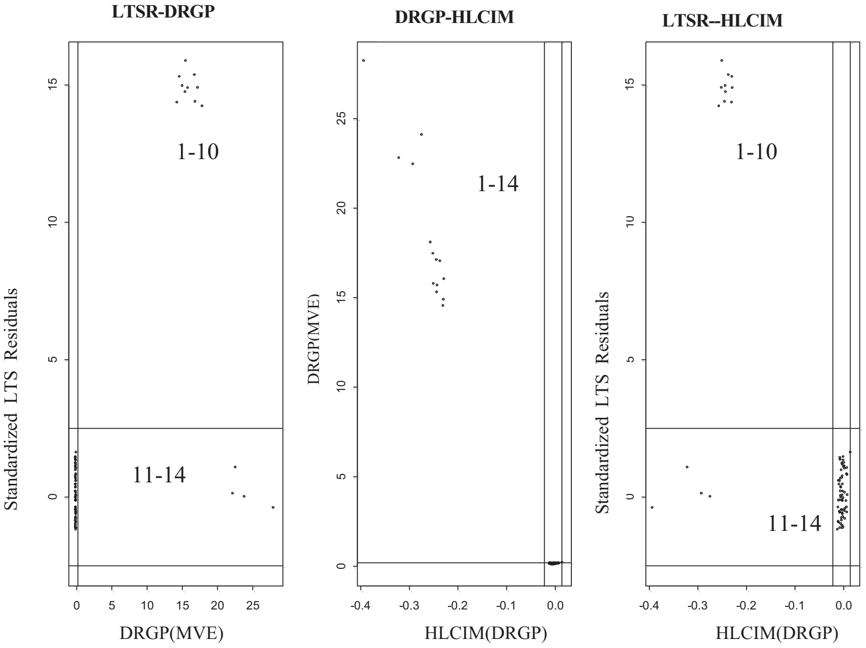

***Figure 4:** Diagnostic Plots of Hawkins-Bradu-Kass Data Set.*

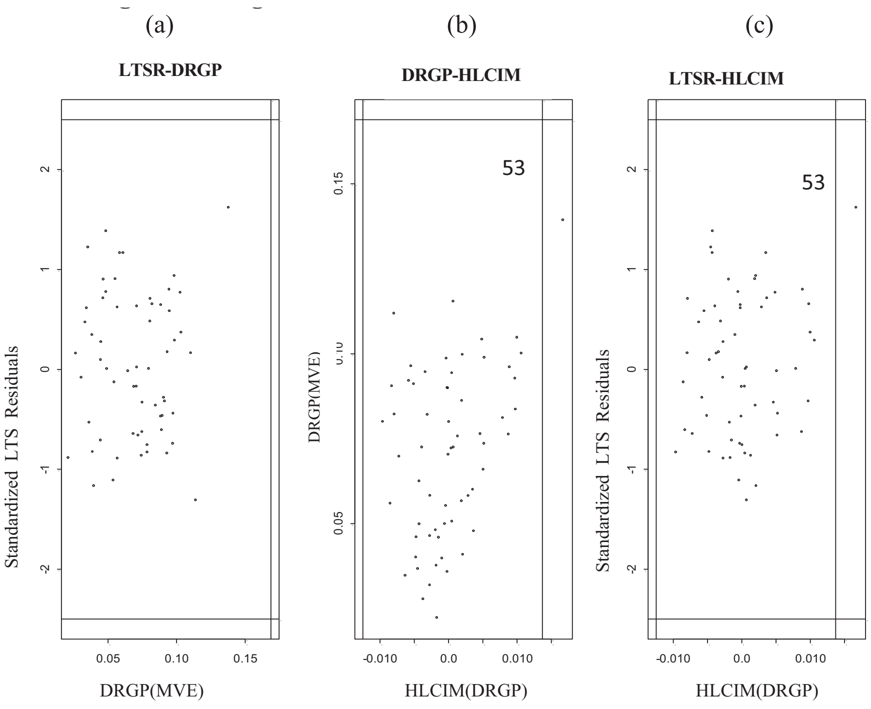

***Figure 5:** Diagnostic Plots of Hawkins-Bradu-Kass Data Set Without the First 14 Observations.*

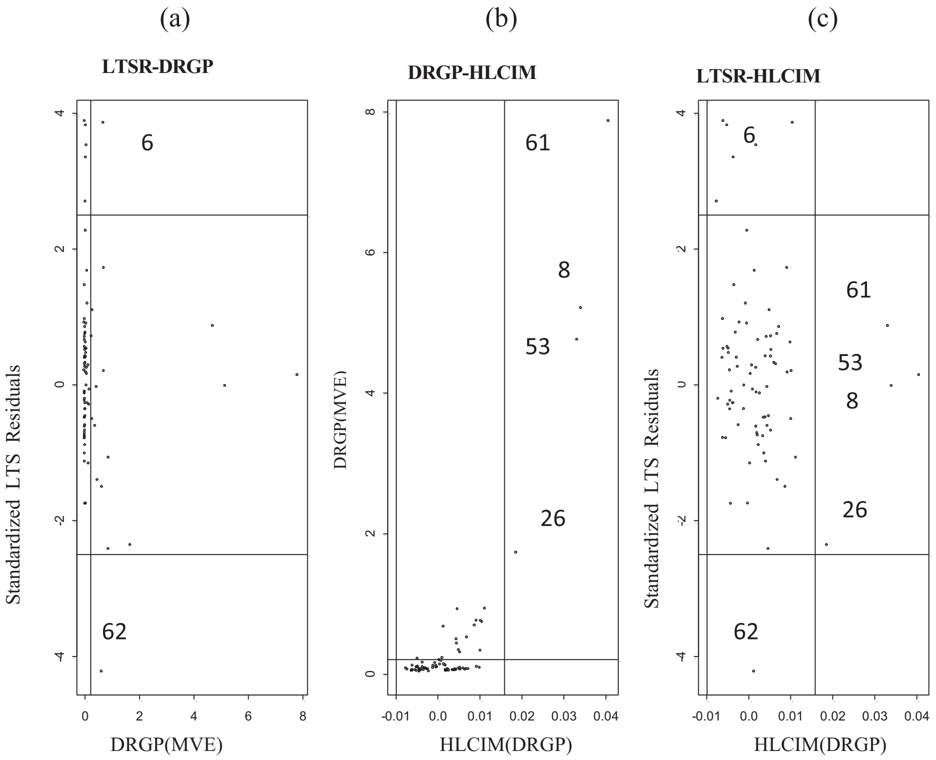

Figure 6: *Diagnostic Plots of Commercial Properties Data Set.*

reveal that cases 8, 26, 53, and 61 are high leverage collinearity-reducing observations and good leverage collinearity-reducing observations, respectively.

After modifying the Commercial Properties data set by replacing observations 1, 2, 3, 6, 7 and 8 in each of the explanatory variables by fixed values of 300, 200, 100, 300, 200, and 100, respectively, these observations became good leverage points (Figure 7 part (a)). Figure 7 part (a) also indicates that case 8 is a bad leverage point. All the modified cases of 1, 2, 3, 6, 7 and 8 are high leverage collinearity-enhancing observations (Figure 7, part (b)). According to Figure 7, part (c), case 8 is a bad leverage collinearity-enhancing observation while cases 1, 2 , 3, 6 and 7 are good leverage collinearity-enhancing observations. Hence, cases 1, 2, 3, 6, and 7 require more attention in order to prevent any misleading conclusions.

Figures 8 to 10 are diagnostic plots for the original and modified Body Fat data set. Part (a) of Figure 8 shows that the original Body Fat data set has four good leverage points (cases 5, 15, 1 and 3) and having zero vertical outliers. Only case 15 is a high leverage collinearity-reducing observation. It can be seen that case 13, a non high leverage, is also a collinearity-reducing observation (Figure 8 part (b)). Additionally, cases 15 and 13 are good leverage collinearity-reducing and collinearity-reducing observations, respectively (Figure 8 part (c)).

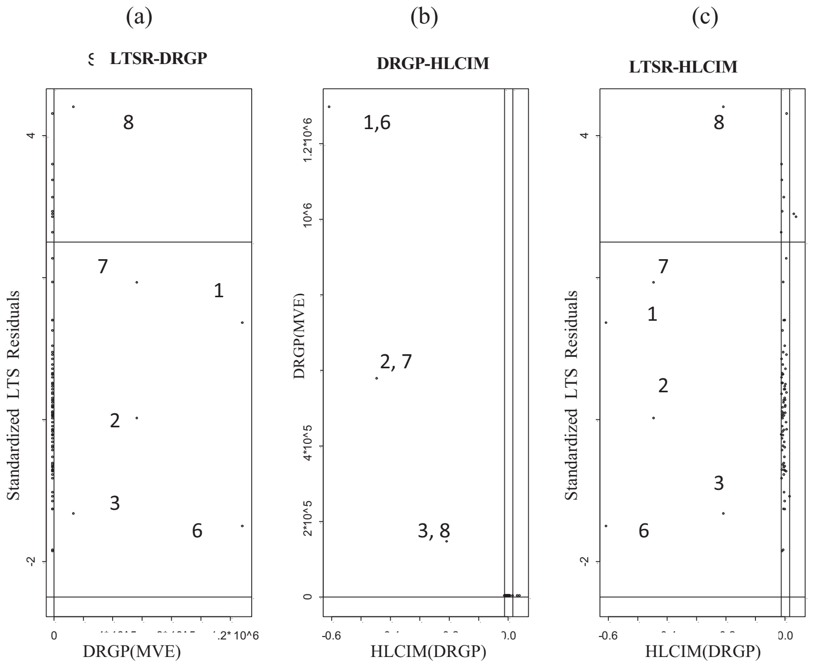

Figure 7: *Diagnostic Plots of Modified Commercial Properties Data Set.*

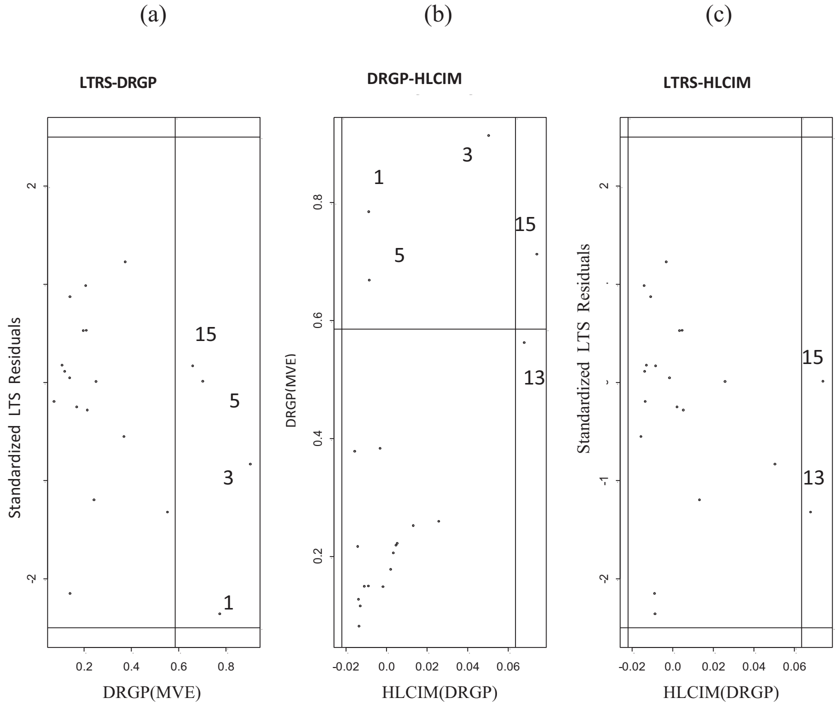

Figure 8: *Diagnostic Plots of Original Body Fat Data Set.*

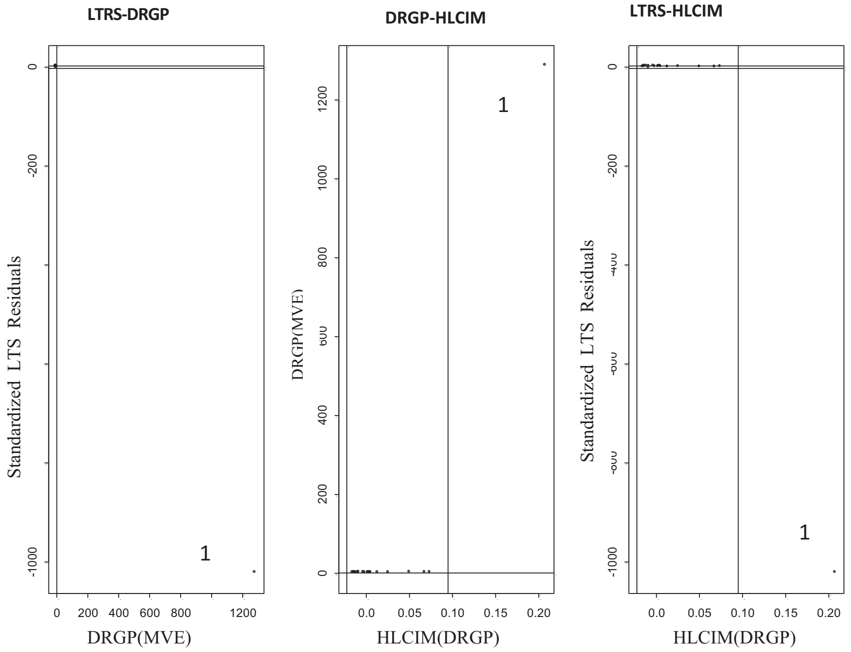

Figure 9: *Diagnostic Plots of Modified x_1 Body Fat Data Set.*

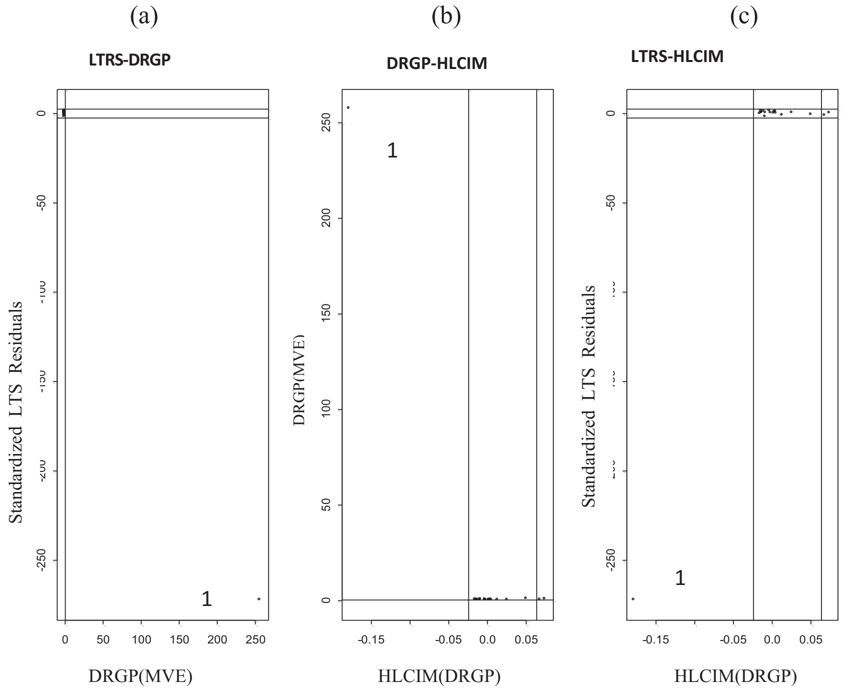

Figure 10: *Diagnostic Plots of Modified x_1 and x_2 in the Same Positions Body Fat Data Set.*

Figure 9 and 10 illustrates the modified Body Fat data set when the first observation of x_1 is fixed to 300 and when the first observation of x_1 and x_2 is fixed to 300, respectively. Figures 9 and 10, part (a), reveal that the added contaminated point is a bad leverage point. Moreover, according to Habshah et al. (2011) when the high leverage point only exists in x_1, case 1 becomes a high leverage collinearity-reducing observation (Figure 9 part (b)). Figure 10 part (b) however, shows case 1 as a high collinearity-enhancing observation when modification is for x_1 and x_2 in the same position. Furthermore, part (c) in Figure 9 shows that case 1 is a bad leverage collinearity-reducing observation while in part (c) of Figure 10 it is a bad leverage collinearity-enhancing observation.

4.2. Monte Carlo simulation study

In this section, a Monte Carlo simulation study was designed to assess the merit of our proposed diagnostic plots in terms of its ability to separate the data set according to regular observations, vertical outliers (regression outliers), collinearity-enhancing/reducing observations with large residuals, bad leverage collinearity-enhancing/reducing observations, good leverage collinearity-enhancing/reducing observations and collinearity-enhancing/reducing observations. To achieve this aim, non-collinear and collinear data sets with three regressors were generated in such a way that different scenarios were created, namely, high leverage collinearity-enhancing/reducing observations and vertical outliers. It is important to mention here that although the proposed diagnostics plots can detect collinearity-enhancing/reducing observations clearly, they were not explicitly generated. In each scenario, four samples of size 40, 60, 100, and 300 and different levels of high leverages of (the percentage of added contaminated cases) = 0.05,0.10, 0.15, 0.20 with unequal weights were considered.

In order to generate high leverage collinearity-enhancing observations, each variable was firstly generated from Uniform (0,1) to produce non-collinear data sets. This generated data is referred to as the regular observations. The last $100\%\alpha$ observations of the regular observations of each regressor were then replaced with certain percentage of high leverage points to create high leverage collinearity-enhancing observations. To generate the high leverage points as collinearity-enhancing with unequal weights in non-collinear data sets, the values corresponding to the first high leverage point were kept fixed at 10 and those of the successive values were created by multiplying the observations index, i, by 10.

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

$$x_{ij} = (1-\rho^2)z_{ij} + \rho z_{i(t1)} \tag{10}$$

where the z_{ij}, $i = 1,\ldots,n$; $j = 1,\ldots,t+1$; t=3, are independent standard normal random numbers. The value of ρ^2 or the correlation between the two explanatory variables, was

Table 1: *The abbreviations used in Tables 2-6.*

Abbreviations	Meaning
CN	the condition number of ***X*** matrix without high leverage points
CN*	the condition number of ***X*** matrix with high leverage points
RO	the number of simulated regular observations
VO	the number of simulated vertical outliers
DCEO	the number of detected collinearity-enhancing observations

set to be equal to 0.95 which causes 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.

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 vertical outliers were generated. Similarly, Good leverage Collinearity-Reducing Observation (GLCRO) was created only when High leverage Collinearity-Reducing Observation (HLCRO) was generated, while the Bad leverage Collinearity-Reducing Observation (BLCRO) was created when both HLCRO and vertical outliers were generated.

Table 1 shows the notations used in Tables 2-6 (D in the entire abbreviations indicates the number of detected observations by the proposed plots). We ran 10,000 simulations. The results based on their averages are presented in Tables 2 to 6. Due to space constraints, only the results for $n = 40$ and 300 are included. The conclusions of other results were consistent.

Let us first look at Table 2 when $\alpha = 0.00$. It can be seen that when there is no vertical outliers or high leverage points in the data, the value of CN=CN* and is less than 5.0, indicating that there is no multicollinearity problem. It is also interesting to note that our proposed plots can detect almost all observations as regular observations (on the average of 96 percent). The results in Table 2 also indicate that in the presence of vertical outliers and in the absence of high leverage points, the data sets do not have multicollinearity problems $(CN < 5.0)$. The results also suggest that the number of detected vertical outliers is reasonably close to the number of generated vertical outliers.

As for the generated bad/good leverage collinearity-enhancing observations data (see Tables 3-4), all the CN* values (> 30) drastically increased in the presence of high leverage points. This indicates that high leverage points are the cause of multicollinearity.

Table 2: *The number of detected abnormal observations in the simulated data sets with vertical outliers.*

n			40					300		
α	0.00	0.05	0.10	0.15	0.20	0.20	0.00	0.05	0.1	0.20
CN	3.54	3.54	3.39	3.39	3.39	3.29	3.29	3.29	3.29	3.29
CN*	3.54	3.54	3.39	3.39	3.39	3.29	3.29	3.29	3.29	3.29
RO	**40.00**	38.00	36.00	34.00	32.00	**300.00**	285.00	270.00	255.00	240.00
DRO	**38.42**	34.59	33.24	31.95	31.27	**298.75**	283.38	268.39	253.07	238.09
VO	0.24	**2.00**	**4.00**	**6.00**	**8.00**	0.00	**15.00**	**30.00**	**45.00**	**60.00**
DVO	0.00	**1.85**	**3.87**	**5.88**	**7.89**	0.00	**14.30**	**29.47**	**44.60**	**59.74**
DCEO-VO	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.07	0.13	0.17
DBLCEO	0.19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DGLCEO	0.05	0.19	0.25	0.25	0.25	1.06	0.00	0.00	0.00	0.00
DCEO	0.00	0.05	0.01	0.00	0.00	0.00	1.03	1.00	0.93	0.90
DCRO-VO	0.00	0.05	0.00	0.00	0.00	0.00	0.67	0.96	1.17	1.00
DBLCRO	0.76	0.05	0.12	0.13	0.00	0.10	0.00	00.00	0.00	0.00
DGLCRO	0.34	0.72	0.27	0.25	31.00	0.09	0.10	0.10	0.10	0.10
DCRO	0.00	2.5	2.24	1.54	0.28	0.00	0.49	0.01	0.00	0.00

Table 3: *The number of abnormal observations in the simulated data sets with bad leverage collinearity-enhancing observations.*

n			40				300	
α	0.05	0.10	0.15	0.20	0.05	0.10	0.15	0.20
CN	3.56	3.41	3.38	3.64	3.29	3.30	3.29	3.27
CN*	23.68	60.09	107.56	166.13	131.70	375.29	704.68	1107.26
RO	38.00	36.00	34.00	32.00	285.00	270.00	255.00	240.00
DRO	35.49	34.04	32.72	30.99	283.07	268.35	252.89	238.72
DVO	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DCEO-VO	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BLCEO	**2.00**	**4.00**	**6.00**	**8.00**	**15.00**	**30.00**	**45.00**	**60.00**
DBLCEO	**1.74**	**3.91**	**5.95**	**8.00**	**14.87**	**30.00**	**45.00**	**60.00**
DGLCEO	0.26	0.09	0.00	0.00	0.00	0.00	0.98	0.30
DCEO	0.44	0.18	0.00	0.00	0.00	0.00	0.00	0.00
DCRO-VO	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DBLCRO	0.00	0.00	0.00	0.00	0.00	0.00	00.00	0.00
DGLCRO	0.15	0.18	0.15	0.00	0.56	0.50	0.02	0.00
DCRO	1.62	1.60	1.18	1.01	1.50	1.15	1.11	0.98

On the other hand, all the CN* values (< 5.00) for the generated bad/good leverage-reducing observations (see Tables 5-6) dramatically reduced in the presence of high leverage collinearity-reducing observations, suggesting that high leverage points conceal the problem of multicollinearity. The large and small values of CN* confirm that the generated data are collinear and non-collinear data sets, respectively. It can be observed that the number of detected bad/good leverage collinearity-enhancing observations is fairly close to the simulated data. A similar conclusion can be made for the

Table 4: *The number of abnormal observations in the simulated data sets with good leverage collinearity-enhancing observations.*

n			40				300	
α	0.05	0.10	0.15	0.20	0.05	0.10	0.15	0.20
CN	3.46	3.50	3.13	3.64	3.29	3.30	3.29	3.27
CN*	23.69	61.92	107.60	166.05	132.15	377.64	704.52	1107.30
RO	38.00	36.00	34.00	32.00	285.00	270.00	255.00	240.00
DRO	35.62	34.36	32.65	31.00	283.16	268.53	253.91	239.01
DVO	0.15	0.00	0.08	0.00	0.00	0.00	0.00	0.00
DCEO-VO	0.22	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BLCEO	**2.00**	**4.00**	**6.00**	**8.00**	**15.00**	**30.00**	**45.00**	**60.00**
DBLCEO	**1.74**	**4.00**	**5.95**	**8.00**	**14.91**	**30.00**	**45.00**	**60.00**
DGLCEO	0.45	0.09	0.00	0.00	0.09	0.00	0.98	0.30
DCEO	0.44	0.18	0.00	0.00	0.00	0.00	0.00	0.00
DCRO-VO	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DBLCRO	0.00	0.00	0.00	0.00	0.00	0.00	00.00	0.00
DGLCRO	0.15	0.09	0.10	0.00	0.55	0.31	0.02	0.00
DCRO	1.63	1.46	1.17	1.00	1.20	1.16	1.07	0.99

Table 5: *The number of abnormal observations in the simulated data sets with bad leverage collinearity-reducing observations.*

n			40				300	
α	0.05	0.10	0.15	0.20	0.05	0.10	0.15	0.20
CN	39.40	38.91	36.33	41.17	36.80	37.13	38.34	38.97
CN*	13.12	4.46	1.06	1.13	1.02	1.01	1.01	1.00
RO	38.00	36.00	34.00	32.00	285.00	270.00	255.00	240.00
DRO	36.45	35.48	33.67	32.00	284.02	269.78	255.00	240.00
DVO	0.15	0.03	0.00	0.00	0.00	0.00	0.00	0.00
DCEO-VO	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DBLCEO	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DGLCEO	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DCEO	0.65	0.33	0.31	0.00	0.01	0.00	0.00	0.00
DCRO-VO	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BLCRO	**2.00**	**4.00**	**6.00**	**8.00**	**15.00**	**30.00**	**45.00**	**60.00**
DBLCRO	**1.85**	**3.96**	**5.98**	**8.00**	**15.00**	**30.00**	**45.00**	**60.00**
DGLCRO	0.10	0.11	0.04	0.00	0.98	0.22	0.00	0.00
DCRO	0.80	0.10	0.00	0.00	0.00	0.00	0.00	0.00

case of detecting bad/good leverage collinearity-reducing observations. It is very important to note that as the value of alpha increases, the degree of multicollinearity also increases/decreases.

Table 6: *The number of abnormal observations in the simulated data sets with bad leverage collinearity-reducing observation.*

n	40					300		
α	0.05	0.10	0.15	0.20	0.05	0.10	0.15	0.20
CN	38.39	40.70	36.33	40.13	36.76	37.16	38.34	37.28
CN*	1.84	2.49	1.06	1.04	1.02	1.01	1.01	1.00
RO	38.00	36.00	34.00	32.00	285.00	270.00	255.00	240.00
DRO	36.56	35.12	33.47	31.62	282.01	269.23	255.00	240.00
DVO	0.25	0.15	0.00	0.00	0.00	0.00	0.00	0.00
DCEO-VO	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DBLCEO	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DGLCEO	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
DCEO	0.31	0.39	0.33	0.26	0.00	0.00	0.00	0.00
DCRO-VO	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00
BLCRO	**2.00**	**4.00**	**6.00**	**8.00**	**15.00**	**30.00**	**45.00**	**60.00**
DBLCRO	**1.85**	**3.96**	**5.98**	**8.00**	**15.00**	**30.00**	**45.00**	**60.00**
DGLCRO	0.10	0.11	0.04	0.00	0.98	0.22	0.00	0.00
DCRO	0.66	0.34	0.20	0.12	2.99	0.77	0.00	0.00

5. Conclusions

Based on Rousseeuw and Van Zomeren (1990) and Rousseeuw and Van Driessen (1999) and their development of Residual-Distance and Distance to Distance plots, three new diagnostic plots are proposed; the LTSR-DRGP, DRGP-HLCIM, and LTSR-HLCIM. The LTSR-DRGP plot was able to identify regular observations, good or bad leverage points and vertical outliers. The DRGP-HLCIM plot was able to classify the observations as regular observations, high leverage points, high leverage collinearity-enhancing or collinearity- reducing observations and collinearity-enhancing or collinearity-reducing observations. Finally, the LTSR-HLCIM plot successfully distinguishes vertical outliers, good leverage collinearity-enhancing/reducing observations, collinearity-enhancing/reducing observations and bad leverage collinearity-enhancing/reducing observations and collinearity-enhancing/reducing observations with large residuals. Thus, the merits of our proposed diagnostic plots are confirmed, as reflected in their application to different authentic data sets and in the Monte Carlo simulation study.

References

Andersen, R. (2008). *Modern Methods for Robust Regression.* Sara Miller McCune: SAGE publications, USA.

Bagheri, A.,Habshah, M. and Imon, A. H. M. R. (2009). Two-step robust diagnostic method for identification of multiple high leverage points. *Journal of mathematics and Statistics*, 5, 97–106.

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

Bagheri, A., Habshah, M. (2009). Robust estimations as a remedy for multicollinearity caused by multiple high leverage points. *Journal of Mathematics and Statistics*, 5, 311–321.

Bagheri, A., Habshah, M. (2011). On the performance of robust variance inflation factors. *International Journal of Agricultural and Statistics Sciences*, 7, 31–45.

Bagheri, A., Habshah, M. (2012). On the performance of the measure for diagnosing multiple high leverage collinearity-reducing observations. *Mathematical Problems in Engineering*, Volume 2012, Article ID 531607, 16 pages.

Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). *Regression Diagnostics: Identifying : Influential Data and Sources of Collinearity.* Wiley, New York.

Belsley, D. A. (1991). *Conditioning Diagnostics: Collinearity and Weak Data in Regression.* Wiley, New York.

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression.* Chapman Hall, London.

Gray, J. B. (1983). The L-R plot: a graphical tool for assessing influence. *Proceedings of the statistical computing section. American statistical association*, 159–164.

Habshah, M. and Bagheri, A. (2013). Robust multicollinearity diagnostic measures based on minimum covariance determination approach. *Economics Computation and Economic Cybernetics Studies and Research*, 4.

Habshah, M., Norazan, M. R. and Imon, H. M. R. (2009). The performance of diagnostic-robust generalized potentials for the identification of multiple high leverage points in linear regression. *Journal of Applied Statistics*, 36, 507–520.

Habshah, M., Bagheri, A. and Imon, A. H. M. R. (2010). The application of robust multicollinearity diagnostic method based on robust coefficient determination to a non-collinear data. *Journal of Applied Sciences*, 10, 611–619.

Habshah, M., Bagheri, A. and Imon, A. H. M. R. (2011). High leverage collinearity-enhancing observation and its effect on multicollinearity pattern: Monte Carlo simulation study. *Sains Malaysiana*, 40, 1437–1447.

Hadi, A. S. (1988). Diagnosing collineariy-influential observations. *Computational Statistics and Data Analysis*, 7, 143–159.

Hadi, A. S. (1992). A new measure of overall potential influence in linear regression. *Computational Statistics and Data Analysis*, 14, 1–27.

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

Imon, A. H. M. R. (2002). Identifying multiple high leverage points in linear regression. *Journal of Statistical Studies. Special Volume in Honour of Professor Mir Masoom Ali.*, 3, 207–218.

Kamruzzaman, M. D. Imon, A. H. M. R. (2002). High leverage point: Another source of multicollinearity. *Pakistan Journal of Statistics,* 18, 435–448.

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

Lawrence, K. D. and Arthur, J. L. (1990). *Robust Regression: Analysis and Applications*. INC: Marcel Dekker.

Moller, S. F., Frese, J. V. and Bro, R. (2005). Robust methods for multivariate data analysis. *Journal of Chemometrics*, 19, 549–563.

Myers, R. H. (1990). *Classical and Modern Regression with Applications*. 2nd Edition. CA: Duxbury Press.

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

Rousseeuw, P. J. and Van Driessen, K.(1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41, 212–223.

Rousseeuw, P. and Van Zomeren, B. (1990). Unmasking multivariate outliers and leverage points. *Journal of American Statistical Association*, 85, 633–639.

Sall, J. (1990). Leverage plots for general linear hypothesis. *The American Statistician,* 44, 308–315.

Sengupta, D. and Bhimasankaram, P. (1997). On the roles of observations in collinearity in the linear model. *Journal of American Statistical Association,* 92, 1024–1032.

Stine, R. A. (1995). Graphical interpretation of variance inflation factors. *The American Statistician,* 49, 53–56.