Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 279472, 12 pages
http://dx.doi.org/10.1155/2015/279472

Research Article

A New Robust Diagnostic Plot for Classifying Good and Bad High Leverage Points in a Multiple Linear Regression Model

Mohammed Alguraibawi,[1,2] Habshah Midi,[1] and A. H. M. Rahmatullah Imon[3]

[1] *Faculty of Science and Institute for Mathematical Research, UPM, 43400 Serdang, Malaysia*
[2] *Al-Dewanyia Technical Institute, ATU, Dewanyia, Iraq*
[3] *Department of Mathematical Sciences, Ball State University, Muncie, IN 47306, USA*

Correspondence should be addressed to Habshah Midi; habshahmidi@gmail.com

Received 24 September 2015; Accepted 19 October 2015

Academic Editor: Eric Florentin

Identification of high leverage point is crucial because it is responsible for inaccurate prediction and invalid inferential statement as it has a larger impact on the computed values of various estimates. It is essential to classify the high leverage points into good and bad leverage points because only the bad leverage points have an undue effect on the parameter estimates. It is now evident that when a group of high leverage points is present in a data set, the existing robust diagnostic plot fails to classify them correctly. This problem is due to the masking and swamping effects. In this paper, we propose a new robust diagnostic plot to correctly classify the good and bad leverage points by reducing both masking and swamping effects. The formulation of the proposed plot is based on the Modified Generalized Studentized Residuals. We investigate the performance of our proposed method by employing a Monte Carlo simulation study and some well-known data sets. The results indicate that the proposed method is able to improve the rate of detection of bad leverage points and also to reduce swamping and masking effects.

1. Introduction

In regression problems, several versions of outliers exist such as residual outliers, vertical outliers, and high leverage points. Any observation that has large residual is referred to as a residual outlier. Ve rtical outliers (VO) or y-outliers are those observations which are extreme or outlying in y-coordinate. On the other hand, high leverage points (HLPs) are those observations which are extreme or outlying in X-coordinate. HLPs can be classified into good leverage points (GLPs) and bad leverage points (BLPs). GLPs are those outlying observations in the explanatory variables that follow the pattern of the majority of the data, while BLPs are the opposite. BLPs have a larger impact on the computed values of various estimates. On the other hand, GLPs contribute to the efficiency of an estimate (see [1–4]). As such, only BLPs should be weighted down while GLPs should not be given low weights in the computation of weighting function in any robust method. Nonetheless, it is now evident that most robust methods attempt to reduce the effect of outliers by weighting the outliers down, irrespective of whether they are good or bad leverage points.

There are a number of good papers in the literature on the detection of HLPs (see [5–8]). However, those detection methods are mainly focused only on the identification of HLPs without taking into consideration their classification into good and bad. It is very important to make the classification, as only the BLPs are responsible for the misleading conclusion about the fitting of the regression model.

It is not easy to capture the existence of several versions of outliers in multiple regression analysis by using a graphical method ([6]). If only one independent variable is being considered, the four types of outliers can easily be observed from a scatter plot of y against the x variables. However, for more than one predictor variable, it is difficult to detect these outliers from a scatter plot. Not much work has been focused on classifying HLP's into good and bad leverage points. Rousseeuw and Van Zomeren [2] have proposed a robust diagnostic plot or outlier map which is more efficient than the nonrobust plot for classifying observations into four types of data points, namely, regular or good observations, vertical outliers, GLPs, and BLPs. We suspect that this plot fails to detect multiple outliers and high leverage points. Pison and Van Aelst [9], suggested a new plot using robust distance obtained from robust location and scale estimators to identify outliers and empirical influences and to find the influence of observations. Similar to Rousseeuw and Van Zomeren plot, they construct a graphical tool for multivariate models. Although Pison and Van Aelst plot can be very useful in model building stage to evaluate the quality of a fit based on the number of detected outliers, it does not focus on classifying unusual data into VO, GLP, and BLP (see [9]). Hubert et al. 2005 introduced a new diagnostic plot based on robust principle component analysis (PCA) denoted by ROBPCA (see [10]). The ROBPCA can distinguish between 4 types of observations for high dimensional data. The regular observations are close to the PCA subspace. GLP lies close to the PCA space but far from the regular observations. We can also have VO, which have a large orthogonal distance to the PCA space. The BLPs have a large orthogonal distance whose projection on the PCA subspace is remote from the typical projections. To draw the diagnostic plot or outliers map, on the horizontal axis, we plot the robust score distance of each observation and on the vertical axis we draw the orthogonal distance of each observation. They suggested that the cut-off point for horizontal axis is $\sqrt{\chi^2_{p,0.975}}$ and approximately cut-off value for vertical axis equals $(\hat{\mu} + \hat{\sigma} z_{0.975})$, where $\hat{\mu}$ and $\hat{\sigma}$ are the estimated mean and standard deviation, respectively. The ROBPCA method is efficient for high dimensional data but not for low dimensional data.

As such, we propose a suitable plot, in this regard. The Modified Generalized Studentized Residuals for the identification of multiple vertical outliers and multiple high leverage points are discussed in Section 2. In Section 3, we propose a new robust diagnostic plot for classifying observations into the four categories. The proposed plot is based on the Modified Generalized Studentized Residuals (MGt_i). A simulation study and three numerical examples are presented in Sections 4 and 5, respectively. Finally, some concluding remarks are given in Section 6.

2. The Modified Generalized Studentized Residuals for the Identification of Multiple Outliers and Multiple High Leverage Points

Consider a multiple linear regression model as follows:

$$y = X\beta + \varepsilon, \tag{1}$$

where y is an $(n \times 1)$ vector of observation of dependent variables, X is an $(n\times p)$ matrix of independent variables, β is a $(p\times 1)$ vector of unknown regression parameters, ε is an $(n\times 1)$ vector of random errors with identical normal distribution of $\varepsilon \sim \mathrm{NID}(0, \sigma^2)$, and p is the number of independent variables. The linear regression model in (1) can be rewritten as follows:

$$y_i = x_{ij}^T \beta_j + \varepsilon_i, \quad i = 1, 2, \ldots, n; \ j = 1, 2, \ldots, p. \tag{2}$$

The ordinary least squares (OLS) estimates for linear regression in (1) are given by

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T y \tag{3}$$

and the ith residuals can be expressed in terms of the true disturbance as follows:

$$\hat{\varepsilon}_i = y - \hat{y} = (1 - H)\,\varepsilon, \tag{4}$$

where $H = X(X^T X)^{-1} X^T$ is a projection matrix or hat matrix denoted by "H." The diagonal elements of "H" matrix are called the hat values denoted by h_{ii}, given by

$$h_{ii} = x_i^T \left(X^T X\right)^{-1} x_i, \quad i = 1, 2, \ldots, n. \tag{5}$$

The h_{ii} values are often used as a classical diagnostic method to identify the high leverage points. However, the h_{ii} mostly fails to detect HLPs due to the fact that it suffers from the masking effect ([6, 7, 12, 13]). However, the h_{ii} mostly fails to detect HLPs due to the fact that it suffers from the masking and swamping effects. The masking problem happens when we fail to detect some outliers that are hidden by other outliers (outlier is identified wrongly as an inlier), whereas the swamping happens when we detect some inliers as outliers ([14]). According to Barnett and Lewis (1994), the masking is "the tendency for the presence of extreme observations not declared as outliers to mask the discordancy of more extreme observations under investigation as outliers" ([15]). On the other hand, the errors corresponding to the outliers may be very big, and this may lead to swamping problem. It means that the clean observations are declared as outliers by mistake. Hadi [13] suggested a single case deleted measure called Potentials matrix. The diagonal elements of Potential matrix denoted by "p_{ii}" are given by (see [1, 13, 16])

$$p_{ii} = x_i^T \left(X_{(i)}^T X_{(i)}\right)^{-1} x_i, \quad i = 1, 2, \ldots, n, \tag{6}$$

where $X_{(i)}$ is the matrix X excluding the ith row. We can rewrite p_{ii} as a function of h_{ii} as

$$p_{ii} = \frac{h_{ii}}{1 - h_{ii}}. \tag{7}$$

Although the potential matrix is efficient to identify a single outlier, it is not successful to identify multiple HLPs (see [3, 17]). To remedy this problem, Imon [17] proposed the Generalized Potentials denoted by "GP" as a diagnostic method for multiple HLPs. The GP diagnostic method is able to detect multiple HLPs, but it is not adequately effective in identifying the exact number of HLPs. This is due to the choice of the initial basic subset, which is prone to masking effects. Habshah et al. [3] developed the Diagnostic Robust Generalized Potential (DRGP) to improve the rate of detection of HLPs. They divided data into two sets, remaining data which contains clean data and deletion set which contains all suspect data. Let us denote a set of good cases (remaining in the analysis) by "R" and a set of bad cases (deleted from the analysis) denoted by "D." The DRGP consists of two steps, whereby in the first step the robust method is used to identify the suspected HLPs. In the second step, the generalized potential diagnostic approach is used to confirm our suspicion. Habshah et al. [3] pointed out that the low leverage points (if any) are put back into the estimation of the remaining subset, sequentially (the observation with the smallest p_{ii} will be substituted first), and then recompute the p_{ii} values. This process is continued until all members of the deletion set have been checked to determine whether or not they can be declared as HLPs.

The suspected HLPs are determined by the robust Mahalanobis distance (RMD), based on the minimum volume ellipsoid (MVE) developed by Rousseeuw and Leroy [6] as

$$\mathrm{RMD}_i = \sqrt{[X - T(X)]^T [C(X)]^{-1} [X - T(X)]}, \quad i = 1, 2, \ldots, n, \tag{8}$$

where $T(X)$ and $C(X)$ are robust locations and shape estimates of the MVE, respectively. Habshah et al. [3] suggested using the following cut-off value for the robust Mahalanobis distance:

$$\mathrm{Median}\left(\mathrm{RMD}_i\right) + 3\mathrm{MAD}\left(\mathrm{RMD}_i\right), \tag{9}$$

where MAD is the median absolute deviation.

Suppose that "R" contains $(n - d)$ cases after $d < (n - p)$ cases where "D" have been deleted. Once the remaining set has been determined, the second steps of DRGP are carried out to confirm the suspected HLPs by using the GP, denoted by p_{ii}^* and defined as

$$p_{ii}^* = \begin{cases} h_{ii}^{(-D)}, & \text{for } i \in D, \\ \dfrac{h_{ii}^{(-D)}}{1 - h_{ii}^{(-D)}}, & \text{for } i \in R, \end{cases} \tag{10}$$

where

$$h_{ii}^{(-D)} = x_i^T \left(X_{(-D)}^T X_{(-D)}\right)^{-1} x_i, \quad i = 1, 2, \ldots, n. \tag{11}$$

Observations, in which the p_{ii}^* values are larger than the following threshold,

$$p_{ii}^* > \mathrm{Median}\left(p_{ii}^*\right) + c\mathrm{MAD}\left(p_{ii}^*\right), \tag{12}$$

where c can be taken as a constant value of 2 or 3, are declared as HLPs. The studentized residuals (internally studentized residuals) and R-studentized residual (externally studentized residuals) are widely used measures for the identification of outliers (see Cook and Weisberg [18]). The studentized residuals are defined as follows:

$$r_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}, \quad i = 1, 2, \ldots, n, \tag{13}$$

where $\hat{\sigma} = [\hat{\varepsilon}^T \hat{\varepsilon}/(n - p - 1)]$ is the standard deviation estimator of the residuals. The special case of (13) is called the R-studentized (denoted by t_i), given by (Chatterjee and Hadi [12])

$$t_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1 - h_{ii}}}, \quad i = 1, 2, \ldots, n, \tag{14}$$

where $\hat{\sigma}_{(i)}$ is the standard deviation estimator of the residuals excluding the ith case. These two measures (r_i and t_i) are also not very successful in detecting multiple outliers due to masking and swamping effects ([3, 17]). Imon [17] suggested a Generalized Studentized Residual (denoted by Gt_i) based on a group of deletions to identify multiple outliers. The generalized version of regression diagnostics first requires the selection of deletion group "D" that contains all suspected influential cases and the remaining cases "R." The suspected influential cases consider outliers and HLPs separately, whereby outliers and HLPs are identified using the robust Reweighted Least Squares (RLS) residuals (Rousseeuw and Leroy [6]) and GP (Imon [8]), respectively. The union of the set of suspected outliers and the set of suspected HLPs points forms the members of the deletion set, which has d observations. However, the initial basic subset of Gt_i is not very stable since it is based on the GP which suffers from masking and swamping effects. In this regard, the DRGP is employed to remedy this problem. The Modified Generalized Studentized Residuals (MGt_i) are formulated based on the RLS and the DRGP as initial estimators. Once the "R" set is identified, the vector of the estimated parameters in the remaining groups, denoted by $\hat{\beta}_{(R)}$, is defined as

$$\hat{\beta}_{(R)} = \left(X_R^T X_R\right)^{-1} X_R^T y_R. \tag{15}$$

The residual for ith deletion observation is given by

$$\hat{\varepsilon}_{i(R)} = y_i - x_i^T \hat{\beta}_{(R)}, \quad i = 1, 2, \ldots, n. \tag{16}$$

The ith externally studentized residual t_i^* for the remaining groups "R" is given by

$$t_i^* = \frac{y_i - x_i^T \hat{\beta}_{(R)}}{\hat{\sigma}_{R-i}\sqrt{1 - h_{ii(R)}}} = \frac{\hat{\varepsilon}_{i(R)}}{\hat{\sigma}_{R-i}\sqrt{1 - h_{ii(R)}}}. \tag{17}$$

Thus, the diagonal elements of the hat matrix are given by

$$h_{ii(R)} = x_i^T \left(X_R^T X_R\right)^{-1} x_i, \quad i = 1, 2, \ldots, n. \tag{18}$$

By utilizing the results of Rao [19], an additional point i in the "R" set is defined as

$$h_{ii(R+i)} = x_i^T \left(X_R^T X_R + x_i x_i^T\right)^{-1} x_i = \frac{h_{ii(R)}}{1 + h_{ii(R)}} \tag{19}$$

and the corresponding estimate is given by

$$\begin{aligned} \hat{\beta}_{(R+1)} &= \left(X_R^T X_R + x_i x_i^T\right)^{-1} \left(X_R^T y_R + x_i y_i\right) \\ &= \hat{\beta}_{(R)} + \frac{\left(X_R^T X_R\right)^{-1} x_i}{1 + h_{ii(R)}} \hat{\varepsilon}_{i(R)}. \end{aligned} \tag{20}$$

Hence, the formulation of the externally studentized residual for $i \notin R$ is defined as

$$t_i^* = \frac{y_i - x_i^T \hat{\beta}_{(R)}}{\hat{\sigma}_R\sqrt{1 - h_{ii(R+1)}}} = \frac{\hat{\varepsilon}_{i(R)}}{\hat{\sigma}_R\sqrt{1 + h_{ii(R)}}}. \tag{21}$$

Subsequently, the Modified Generalized Studentized Residuals (MGt_i) for the whole data set are formulated by combining (17) and (21) as follows:

$$\mathrm{MGt}_i = \begin{cases} \dfrac{\widehat{\varepsilon}_{i(R)}}{\widehat{\sigma}_{R-i}\sqrt{1 - h_{ii(R)}}}, & \text{for } i \in R, \\ \dfrac{\widehat{\varepsilon}_{i(R)}}{\widehat{\sigma}_{R}\sqrt{1 + h_{ii(R)}}}, & \text{for } i \notin R, \end{cases} \quad (22)$$

where $\widehat{\sigma}_R$ is the standard deviation of R group and $\widehat{\sigma}_{R-i}$ is the standard deviation of R group excluding the ith case.

3. New Diagnostic Plots for Classifying Observations into Four Categories

Rousseeuw and Van Zomeren [2] proposed a robust diagnostic plot which is more effective than the nonrobust plot for classifying observations into regular observations, vertical outliers, GLPs, and BLPs. Rousseeuw and Van Zomeren plot draws the standardized least median of square residual (LMS) against the robust Mahalanobis distance (RMD) based on the minimum volume ellipsoid (MVE); this plot is denoted by (LMS-RMD). The nonrobust plot draws the Studentized OLS residuals (t_i) against the Mahalanobis distance (MD), and we called this plot as (OLS-MD) plot. We suspect that the robust LMS-RMD diagnostic plot is not very effective in classifying the observations into respective categories since it is based on the robust Mahalanobis distance, which suffers from swamping effects. Moreover, this plot uses studentized residual which is not very successful in identifying multiple outliers. Habshah et al. [3] showed that the DRGP was very successful in detecting multiple HLPs. In addition, we anticipate that the newly proposed MGt_i is able to detect multiple outliers. As such, we proposed improving the classification method of Rousseeuw and Van Zomeren [2] by plotting MGt_i versus DRGP as shown in Table 1. Our proposed diagnostic plot is called (MGt-DRGP) plot. The basic rules for classification observation by using the new proposed method are as follows (see Habshah and Mohammed [20]):

(i) regular observation (RO): an observation is declared as a "RO" if $|\mathrm{MGt}_i| \le 2.5$ and $p^*_{ii} \le \mathrm{Median}(p^*_{ii}) + c\mathrm{MAD}(p^*_{ii})$;

(ii) vertical outlier (VO): an observation is declared as a "VO" if $|\mathrm{MGt}_i| > 2.5$ and $p^*_{ii} \le \mathrm{Median}(p^*_{ii}) + c\mathrm{MAD}(p^*_{ii})$;

(iii) GLPs: an observation is declared as a GLP if $|\mathrm{MGt}_i| \le 2.5$ and $p^*_{ii} > \mathrm{Median}(p^*_{ii}) + c\mathrm{MAD}(p^*_{ii})$;

(iv) BLPs: an observation is declared as a BLP if $|\mathrm{MGt}_i| > 2.5$ and $p^*_{ii} > \mathrm{Median}(p^*_{ii}) + c\mathrm{MAD}(p^*_{ii})$.

4. Monte Carlo Simulation Study

In this section, a Monte Carlo simulation study is designed to evaluate the performance of our new proposed method, MGt-DRGP plot in classifying observation into regular observations, vertical outliers, and good and bad HLPs. Here, the MGt-DRGP plot is compared with some existing plots, namely, the OLS-MD and LMS-RMD plots. The performances of these plots are evaluated based on the rate of correct detection of BLPs and the rate of masking and swamping effects. A good plot is the one that has a higher percentage of correct detection of BLPs and smaller rate of masking and swamping effects. The experiments consider two explanatory variables, $p = 2$ and 3. In each experiment, four samples of size $n = 20, 40, 100$, and 200 and different percentages of BLPs ($\alpha = 0.05, 0.10, 0.15$, and 0.20) are considered. The regular observations are generated according to a normal distribution with a mean equaling 0 and a variance equaling 1. In order to generate HLPs in a data set, the first "100 α%" observations of the regular data in X and y variables are replaced with a certain percentage of BLPs. To generate BLPs, the first value of a BLP is kept fixed at 10 and sequential values are created by multiplying the values index, j, by 2. Each experimental run involves 5000 replications. The results of the Monte Carlo simulation study are summarized in Tables 2 and 3. They present the percentage of correct detection of BLPs and the masking and swamping rates at different levels of contamination and different sample sizes.

Table 1: Scatter plot of DRGP against Modified Generalized Studentized Residuals.

	DRGP	
Modified Generalized Standard Residual	Vertical outliers	Bad leverage points
	Regular observations	Good leverage points
	Vertical outliers	Bad leverage points

The results clearly indicate that the MGt-DRGP plot has a superior ability to identify the correct number of BLPs compared to OLS-MD and LMS-RMD plots regardless of the number of regressor variables, contamination rate, and size of samples. At a low level of contamination ($\alpha = 0.05$), the MGt-DRGP plot has perfectly identified the BLPs without any masking or swamping effects. The LMS-RMD plot has a high percentage of detection of BLPs. However, it also has a high percentage of swamping due to the weakness of the RMD method, which tends to swamp some low leverage points. Although the OLS-MD has a small low rate of swamping, its performance is very poor for detection of BLPs due to the masking effects.

At moderate and high levels of contamination "$\alpha = 0.10, 0.5$, and 0.20," the MGt-DRGP plot still outperforms other plots. On the other hand, the performance of LMS-RMD plots decreases in terms of having a smaller percentage of correct detection and increasing rate of masking and swamping effects. It can be seen that the performance of the OLS-MD plot is very bad as the level of contamination increases. The MGt-DRGP plot consistently has a higher percentage of correct detection and the lowest rate of swamping and masking, irrespective of the number of independent variables p, level of contamination, and sample size.

Table 2: Percentage of correctly identified BLPs, masking, and swamping for simulation data ($p = 2$).

Cont. level	n	% Correct detection			% Masking			% Swamping		
		OLS-MD	LMS-RMD	MGt-DRGP	OLS-MD	LMS-RMD	MGt-DRGP	OLS-MD	LMS-RMD	MGt-DRGP
5%	20	75	100	100	25	0.0	0.0	5.0	15.5	0.0
	40	50	100	100	60	0.0	0.0	2.2	14.8	0.0
	100	56	100	100	44	0.0	0.0	0.0	6.40	0.0
	200	40	100	100	60	0.0	0.0	0.9	4.60	0.0
10%	20	50	100	100	50	0.0	0.0	5.0	10.0	0.0
	40	50	90	100	60	10	0.0	0.0	17.5	0.0
	100	30	87.3	100	70	12.7	0.0	0.0	5.0	1.2
	200	30	92.3	100	70	7.7	0.0	1.0	4.0	1.0
15%	20	0.0	100	100	100	0.0	0.0	5.0	10.0	0.0
	40	16.7	66.7	97.0	83.3	33.3	3.0	0.0	15.0	0.0
	100	13.3	70	96.2	86.7	30.0	3.8	0.0	7.30	1.1
	200	23.3	85	99	76.7	15.0	1.0	0.0	2.0	1.5
20%	20	0.0	50	100	80	50	0.0	5.0	23.0	0.0
	40	12.5	62.5	95.7	81.5	38.5	4.3	0.0	15.0	2.1
	100	15	67	97.1	85	33.0	2.9	0.0	8.20	0.8
	200	20	82	98.1	80	18	1.9	1.0	3.70	0.5

Table 3: Percentage of correctly identified BLPs, masking, and swamping for simulation data ($p = 3$).

Cont. level	n	% Correct detection			% Masking			% Swamping		
		OLS-MD	LMS-RMD	MGt-DRGP	OLS-MD	LMS-RMD	MGt-DRGP	OLS-MD	LMS-RMD	MGt-DRGP
5%	20	50	100	100	50	0.0	0.0	0.3	32.3	0.0
	40	45	95.6	100	55	4.4	0.0	0.0	19.2	0.0
	100	20	95.2	100	80	4.4	0.0	1.0	18.7	0.0
	200	20	96	100	80	4.0	0.0	2.4	13.6	0.0
10%	20	0.0	90	100	100	10	0.0	0.0	26.7	1.5
	40	25	82	99	75	18	1.0	0.0	17.5	1.0
	100	30	84.4	97.4	70	13.6	2.6	1.0	17.0	1.0
	200	30	88	98	70	12	2.0	2.0	9.4	1.5
15%	20	0.0	77.2	100	100	22.8	0.0	0.0	26.7	0.0
	40	14	84.2	94	86	15.8	6.0	0.0	18.6	2.5
	100	14	79	94.2	86	21	5.8	0.0	15.0	2.0
	200	7	82.1	98	93	17.9	2.0	1.5	6.5	2.0
20%	20	0.0	75.3	99	100	24.7	1.0	0.0	16.7	0.0
	40	12	67.9	93.3	88	32.1	6.7	0.0	11	2.5
	100	10	63	93.1	90	37	6.9	0.0	10.0	2.0
	200	8	70	97.1	92	30	2.9	1.0	8.0	1.5

5. Example and Discussion

In this section, three examples are considered to assess the performance of our proposed diagnostic method.

5.1. Artificial Data Set. Our first example is an artificial data set given by Kamruzzaman and Imon [11]. This data set has three variables and contains 20 observations that are generated independently as "uniform (0, 1)." We suppose that the third variable is the response variable and the rest are the predicted variables. The boxplot in Figure 1 shows that this artificial data set does not have any outlier or high leverage point. We would like to evaluate the performance of our proposed diagnostic method in two situations. Firstly, we modified the data to see the effect of one BLP on the plots. To create BLP, the ith case of the response and predictor variables (x_1, x_2, and y) are replaced by outliers, where clean observations are replaced by arbitrary large values displayed in the brackets in Table 4. In this regard, the original data is modified by substituting one good observation with one BLP (case 20). In the second situation, we would like to observe the combined effect of VO, GLP, and BLP. In this situation, the original data is modified by replacing 2 good observations with 2 VO (cases 1 and 2), 2 GLPs (cases 3 and 4), and 1 BLP (case 20). To create VO, only the ith observation of response variable (y) is replaced by arbitrary large value and to create

Table 4: Original and modified (in the brackets) artificial data set by Kamruzzaman and Imon [11].

Ind.	x_1	x_2	Y
1	0.9917	0.7067	0.9820 (5)
2	0.7006	0.8301	0.6167 (5)
3	0.3949 (5)	0.7586 (5)	0.8862
4	0.9618 (7)	0.5460 (7)	0.6400
5	0.0042	0.0504	0.7311
6	0.3044	0.3952	0.5180
7	0.3521	0.3140	0.8409
8	0.0993	0.9953	0.8676
9	0.4072	0.4604	0.5702
10	0.1105	0.8435	0.8864
11	0.8282	0.9434	0.2216
12	0.0859	0.3367	0.3921
13	0.9283	0.6090	0.4289
14	0.4926	0.8648	0.6607
15	0.4069	0.2590	0.2028
16	0.0227	0.1896	0.6677
17	0.4129	0.6420	0.3723
18	0.8919	0.8459	0.0634
19	0.1397	0.5453	0.3354
20	0.6295 (5)	0.5354 (5)	0.5630 (5)

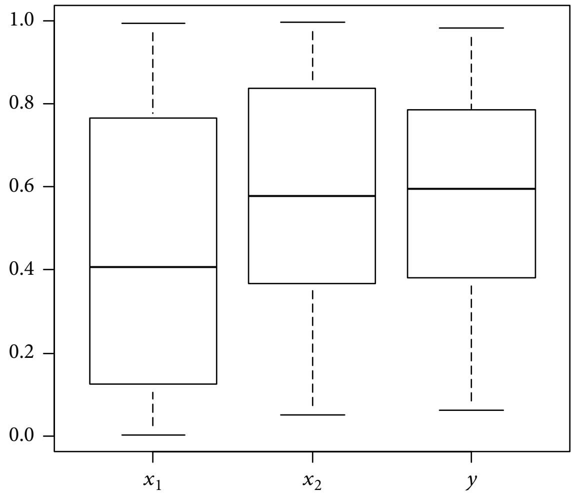


Figure 1: Boxplot for artificial data set.

GLP; the ith case of predicted variables (x_1 and x_2) is replaced by arbitrary large values. Table 4 exhibited the original and the modified artificial data

The three plots (OLS-MD, LMS-RMD, and MGt-DRGP) were then applied to both modified artificial data sets. The values of the diagnostic methods and preceding plots are displayed in Table 5 and Figure 2, respectively. Since only a single BLP is present in the first modified data, we can clearly see that all the classical and robust diagnostic plots are able to correctly detect the BLP (case 20) as shown in Figures 2(a), 2(b), and 2(c). However, for second modified data, it is interesting to observe that only the MGti-DRGP plot is able to detect and classify the 5 outlying observations (cases 1, 2, 3, 4, and 20) into VO, GLPs, and BLPs correctly (see Figure 2(f)). Although the LMS-RMS plot is able to detect and classify those 5 outlying observations correctly, its swamped 4 observations (cases 8, 10, 15, and 18) are as shown in Figure 2(e). The classical OLS-MD can only detect 1 outlying observation (case 2) and mask 4 observations (cases 1, 3, 4, and 20) as shown in Figure 2(d).

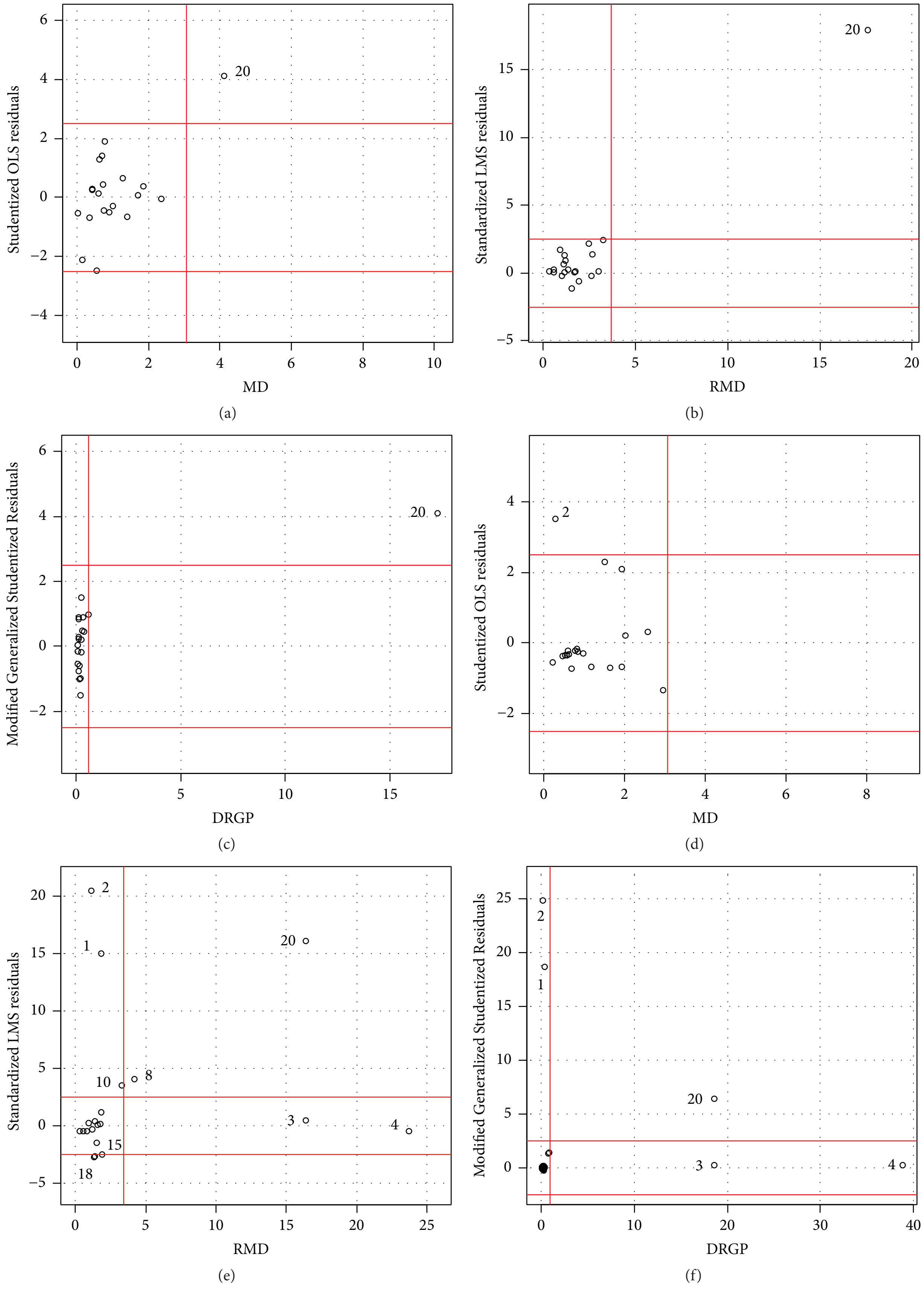
Next, we would like to justify which plot has identified or has classified the suspect observations correctly. Since removing suspect observations leads to a drastic change in the coefficient estimates, a good classification plot is the one which corresponds to the highest percentage changes for various estimates. The percentage of change in estimate (PCE) is computed as

$$\text{PCE} = \left| \frac{\widehat{\theta}_{\text{Proposed}} - \widehat{\theta}_{\text{Original}}}{\widehat{\theta}_{\text{Original}}} \right| \times 100\%, \qquad (23)$$

where $\widehat{\theta}_{\text{Original}}$ is the OLS parameter estimates of the original data and $\widehat{\theta}_{\text{Proposed}}$ is the OLS estimate for data set excluding suspected cases (VO and BLPs) and the "$|\cdot|$" is the absolute value. Another criterion of a good plot is that, after deleting the suspect bad influential observations, there is a significant reduction in the standard error of the estimates. Table 6 presents the regression coefficients, standard error for coefficients, coefficient of determination, F-test, and the PCE values for the second modified data in different situations, when we remove a single outlier (case 2) that is detected by OLS-MD plot and when we remove the outlying observations (cases 1, 2, 8, 10, 15, 18, and 20 and cases 1, 2, and 20) that are detected by LMS-RMD and MGt-DRGP plots, respectively. The results of Table 6 clearly show that the suspected influential observations that are detected by MGt-DRGP plot are removed, resulting in the smallest standard error for coefficients and having the largest PCE values.

5.2. Phosphorus Data Set. The second example is the phosphorus data set, which is taken from Snedecor and Cochran [21]. The concentrations of phosphorus in parts per million in each of 18 soils were measured to investigate the source from which the corn plants obtain their phosphorus. This data set has two predictor variables; x_1 is the concentrations of inorganic phosphorus in the soil and x_2 is the concentrations of organic phosphorus in the soil. The response variable y is the phosphorus content of corn grown in the soil at 20°C. Chatterjee and Hadi [1] showed by using the Studentized OLS residuals that this data set has a single outlier (case 17). Table 7 shows the absolute values for MD, RMD, DRGP, t_i, Studentized LMS residuals, and MGt$_i$. The MD did not detect any HLP, whereas the RMD detected 2 HLPs (cases 1 and 6) and the DRGP detect 3 HLPs (cases 1, 6, and 10). The t_i detected only one outlier (case 17), whereas the Studentized LMS residuals and MGt$_i$ detected 2 outliers (cases 10 and 17).

Next, we would like to see the classification made by OLS-MD, LMS-RMD, and MGt-DRGP plots. From Figures 3, 4, and 5, we can see clearly that the OLS-MD plot identified one vertical outlier (case 17). However, the LMS-RMD plot



(a)

(b)

(c)

(d)

(e)

(f)

Figure 2: The OLS-MD, LMS-RMD, and MGt-DRGP plots for the first situation (plots a, b, and c) and the second situation (plots d, e, and f) of modified artificial data set.

Table 5: Values of MD, RMD, DRGP, t_i, MLS residuals, and MGt$_i$ and their cut-off points (in the brackets) for the second modified Artificial data set.

Case number	$\lvert MD_i \rvert$ (3.058)	$\lvert RMD_i \rvert$ (3.397)	$\lvert DRGP_i \rvert$ (0.792)	$\lvert t_i \rvert$ (2.5)	$\lvert$Stand. MLS res.$_i \rvert$ (2.5)	$\lvert MGt_i \rvert$ (2.5)
1	1.50	1.83	0.37	2.30	15.04	18.69
2	0.27	1.11	0.15	3.52	20.51	24.81
3	1.92	16.38	18.60	0.69	0.48	0.29
4	2.95	23.70	38.86	1.34	0.48	0.26
5	0.77	1.76	0.40	0.23	0.15	0.05
6	0.54	0.52	0.09	0.35	0.48	0.01
7	0.84	0.92	0.14	0.24	0.23	0.19
8	2.57	4.16	0.72	0.30	4.08	1.33
9	0.57	0.27	0.08	0.35	0.44	0.06
10	2.02	3.27	0.83	0.20	3.50	1.47
11	0.22	1.50	0.25	0.55	1.52	0.03
12	0.62	1.16	0.16	0.33	0.32	0.11
13	1.63	1.87	0.36	0.71	2.48	0.09
14	0.82	1.82	0.23	0.16	1.19	0.20
15	1.17	1.38	0.21	0.68	2.64	0.20
16	0.60	1.37	0.24	0.21	0.40	0.03
17	0.46	0.84	0.09	0.38	0.48	0.03
18	0.69	1.30	0.23	0.74	2.77	0.14
19	0.96	1.59	0.18	0.29	0.09	0.12
20	1.92	16.38	18.60	2.09	16.13	6.42

Table 6: The regression estimates, standard errors (in the brackets), and PCE for the second modified Artificial data set.

Variables	Full data	Regular data					
		Remove only the suspect BLP detected by OLS-MD plot [cases 20]		Remove only the suspect BLP detected by LMS-RMD plot [cases 1, 2, 8, 10, 15, 18, and 20]		Remove only the suspect BLP detected by MGt-DRGP plot [cases 1, 2, and 20]	
		Estimate	PCE	Estimate	PCE	Estimate	PCE
$X1$	1.026	0.889	13.35	0.217	78.85	−0.377	136.74
	(1.412)	(1.292)	8.50	(0.319)	77.41	(0.196)	86.12
$X2$	−0.865	−0.934	7.98	−0.190	78.03	0.421	148.67
	(1.457)	(1.332)	8.58	(0.326)	77.63	(0.203)	86.07
Constant	1.185*	1.243*	4.89	0.545**	54.01	0.431**	63.63
	(0.549)	(0.502)	8.56	(0.080)	85.43	(0.078)	85.79
Observation	20	19		13		17	
R^2	0.067	0.030	55.22	0.149	122.39	0.264	294.03
Residual SE	1.817 (df = 17)	1.661 (df = 16)	8.59	0.204 (df = 10)	88.77	0.236 (df = 14)	87.01
F-statistics	0.614	0.246	59.93	0.874	42.35	2.513	309.28

Note: $^*p < 0.05$; $^{**}p < 0.01$.

identified 2 vertical outliers (cases 10 and 17) with 2 GLPs (cases 1 and 6), whereas the MGt-DRGP plot can classify the observations into their respective categories, where cases 1 and 6 are classified as GLPs with one vertical outlier (case 17) and one BLP (case 10). Similar to the artificial data, we would like to assess the performance of our proposed plot. The results of Table 8 demonstrate the standard deviation for regression coefficients in different situations, when we remove a single outlier (case 17) that is detected by OLS-MD plot and when we remove influential observations (cases 10 and 17) which are detected by LMS-RMD and MGt-DRGP plots. Also, we discussed the effect of removing BLPs (cases 10 and 17) and removing GLPs (cases 1 and 6) that are identified by both LMS-RMD and MGt-DRGP plots. The results of the analysis indicate that omitting influential observations (cases 10 and 17) has the greatest effect on the standard error of the parameter estimates. It is interesting to observe that keeping the GLPs (cases 1 and 6) has reduced the standard error of estimates compared to removing these two observations from the data set.

Table 7: Values of MD, RMD, DRGP, t_i, MLS residuals, and MGt$_i$ and their cut-off points (in the brackets) for Phosphorus data set.

Case number	$\lvert \text{MD}_i \rvert$ (3.058)	$\lvert \text{RMD}_i \rvert$ (4.551)	$\lvert \text{DRGP}_i \rvert$ (0.519)	$\lvert t_i \rvert$ (2.5)	$\lvert \text{Stand. MLS res.}_i \rvert$ (2.5)	$\lvert \text{MGt}_i \rvert$ (2.5)
1	1.88	6.02	0.89	0.13	0.71	0.78
2	1.51	1.55	0.24	0.05	0.10	0.21
3	1.70	1.61	0.34	0.40	0.90	0.05
4	1.12	2.73	0.28	0.04	0.25	0.07
5	1.33	1.17	0.20	0.66	0.74	0.92
6	2.62	7.86	1.58	0.78	0.25	1.00
7	0.38	2.23	0.16	0.20	0.24	0.55
8	0.84	0.66	0.12	0.80	1.61	0.82
9	1.07	1.16	0.18	0.68	1.56	0.67
10	1.28	4.02	0.77	1.86	2.88	2.81
11	0.38	0.69	0.07	0.14	0.01	0.02
12	1.13	0.98	0.17	0.28	0.25	0.12
13	1.10	0.83	0.15	1.32	1.38	0.86
14	1.01	0.88	0.15	0.29	0.21	0.05
15	1.24	1.52	0.20	0.38	0.21	0.30
16	0.99	2.73	0.26	0.44	0.97	0.42
17	1.56	1.44	0.26	5.36	5.12	6.22
18	1.78	1.90	0.39	0.83	0.25	0.27

Table 8: The regression estimates, standard errors (in the brackets), and PCE for Phosphorus data set.

Variables	Original data		Regular data					
			Remove suspect observations detected by OLS-MD plot [case 17]		Remove suspect GLP and BLP detected by both MGt-DRGP and LMS-RMD plots [cases 1, 6, 10, and 17]		Remove only the suspect BLP detected by MGt-DRGP and LMS-RMD plots [cases 10, 17]	
	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
Inorg.	1.790	0.557	−1.290	0.343	1.720	0.505	1.211	0.285
Organic	0.087	0.415	−0.111	0.249	−0.394	0.440	0.088	0.218
Constant	56.251	16.311	66.465	9.850	72.230	12.117	60.910	8.398

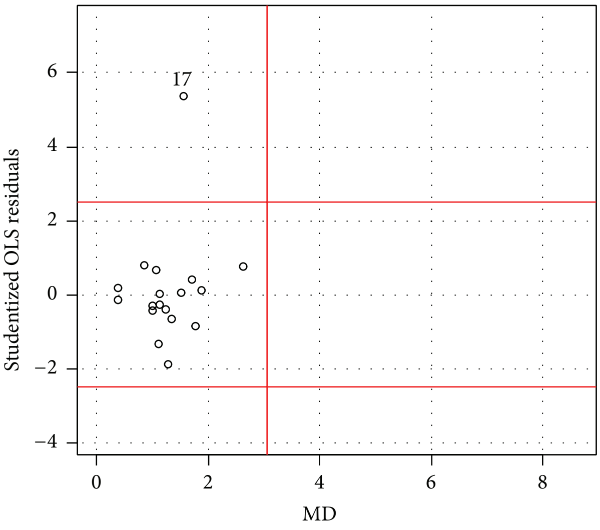


Figure 3: The studentized OLS residual against MD for the Phosphor Data Set.

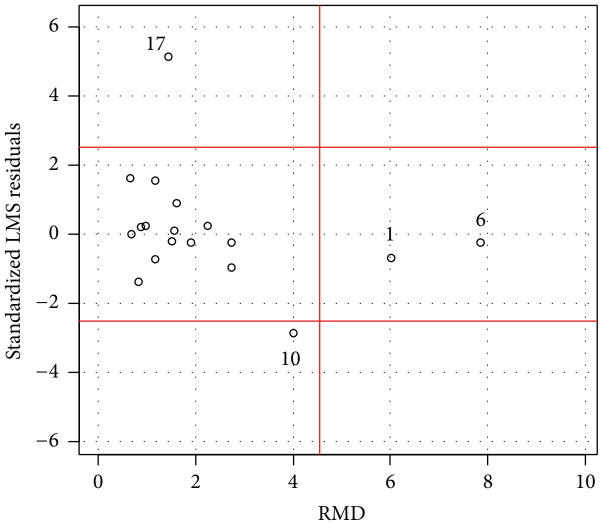


Figure 4: The standardized LMS residual against RMD for the Phosphor Data Set.

Table 9: RMD, DRGP, t_i, MLS residuals, and MGt$_i$ and their cut-off points (in the brackets) for Aircraft data set.

Case number	$\lvert\text{MD}_i\rvert$ (3.582)	$\lvert\text{RMD}_i\rvert$ (4.640)	$\lvert\text{DRGP}_i\rvert$ (0.365)	$\lvert t_i\rvert$ (2.5)	$\lvert$Stand. MLS res.$_i\rvert$ (2.5)	$\lvert\text{MGt}_i\rvert$ (2.5)
1	1.76	1.70	0.05	0.89	0.06	0.37
2	1.53	2.52	0.05	1.26	0.14	0.79
3	1.55	1.59	0.05	1.32	0.66	1.46
4	1.57	1.52	0.05	0.72	0.08	1.71
5	1.11	1.43	0.11	0.17	0.08	0.69
6	2.17	2.54	0.24	0.94	0.18	0.24
7	1.42	2.22	0.09	0.56	0.19	0.83
8	1.91	2.43	0.07	0.94	0.08	0.35
9	2.09	1.93	0.04	0.05	0.79	0.16
10	1.96	2.53	0.29	0.96	1.61	0.87
11	1.64	1.68	0.08	0.29	1.62	1.04
12	0.64	1.22	0.13	1.97	0.14	1.62
13	0.88	1.24	0.11	0.15	0.42	0.01
14	4.29	26.51	0.12	0.03	1.96	0.17
15	0.78	4.16	0.04	0.06	0.19	0.01
16	1.66	3.64	0.24	0.09	3.56	4.74
17	2.09	2.41	0.17	1.97	0.08	0.19
18	1.19	1.53	0.11	0.24	1.64	1.77
19	2.29	2.01	0.41	0.65	1.66	4.35
20	1.55	7.96	0.01	1.09	0.08	0.37
21	2.42	2.15	0.45	0.48	0.84	2.12
22	3.42	7.47	0.50	4.78	9.61	11.73
23	1.11	1.34	0.08	0.28	0.16	0.68

Table 10: Standard errors of regression coefficients for Aircraft data set.

	Original data		Regular data					
Variables			Remove only the suspect BLP detected by OLS-MD plot [case 22]		Remove only the suspect BLP detected by LMS-RMD plot [cases 16, 22]		Remove only the suspect BLP detected by MGt-DRGP plot [cases 16, 19, and 22]	
	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
Aspect. ratio	−3.853	1.763	−3.272	1.191	−3.049	0.919	−3.287	0.850
Life to drag ratio	2.488	1.187	1.873	0.808	1.210	0.649	1.404	0.602
Weight of the plane	0.003	0.0005	0.002	0.0004	0.001	0.0004	0.001	0.0004
Max. thrust	−0.002	0.0005	−0.001	0.0004	−0.001	0.0003	−0.001	0.0003
Constant	−3.791	10.116	4.621	7.023	9.501	5.578	10.804	5.148

5.3. Aircraft Data Set. The last example is the Aircraft data set, which is taken from Gray [22]. This data set contains 23 cases with four predictor variables (aspect ratio, life to drag ratio, weight of the plane, and maximal thrust); the response variable is the Cost. From the results of Table 9, we can see that the t_i identified one outlier (case 22), whereas the Standardized LMS residuals and MGt$_i$ identified cases 16, 22 and cases 16, 19, and 22 as outliers, respectively. Moreover, the MD identified one HLP (case 14) while RMD and DRGP detected cases 14, 20, and 22 and cases 19, 22 as HLPs, respectively.

The classification of data into regular data, vertical outliers, and good and bad leverage points is shown in Figures 6, 7, and 8. It can be observed from Figure 6 that the nonrobust plot (OLS-MD) identified one vertical outlier (case 22) and one GLP (case 14). The LMS-RMD plot in Figure 7 detected one vertical outlier (case 16), BLP (case 22), and 2 GLP (cases 14, 20), while the MGt-DRGP plot in Figure 8 identified one vertical outlier (case 16), two BLPs (cases 19 and 22), and one GLP (case 21).

It can be seen from Table 10 that the standard errors of the estimates when removing observations 16, 22, and 19 (identified by MGt-DRGP) are smaller than those when removing case 22 and cases 16 and 22, which are identified by OLS-MD and LMS-RMD, respectively.

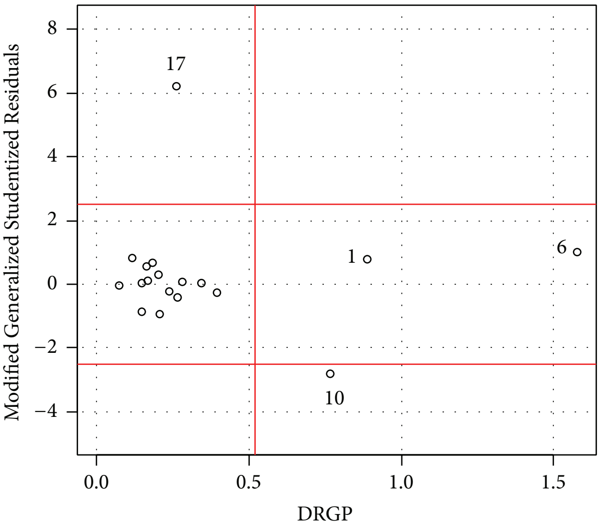


Figure 5: The Modified Generalized Standard Residual against DRGP for Aircraft data.

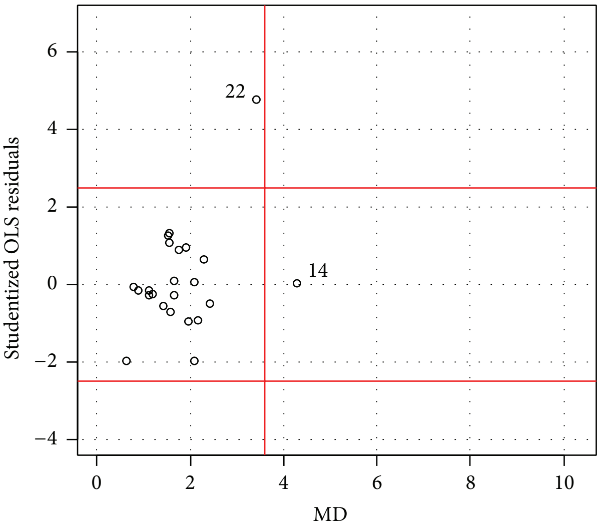


Figure 6: The Studentized OLS residual against MD for the Aircraft data set.

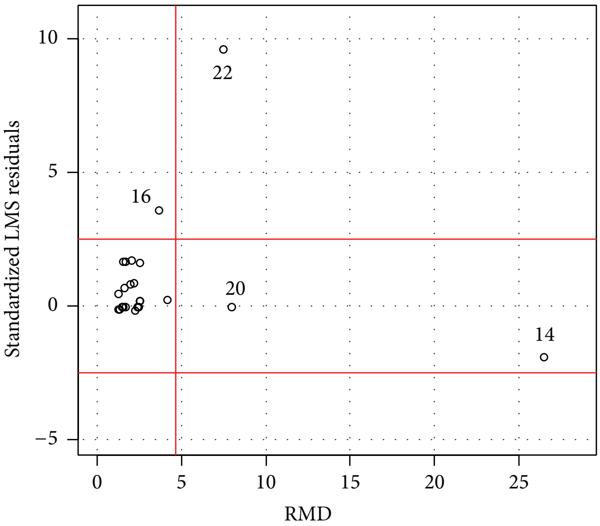


Figure 7: The Standardized LMS residual against RMD for the Aircraft data.

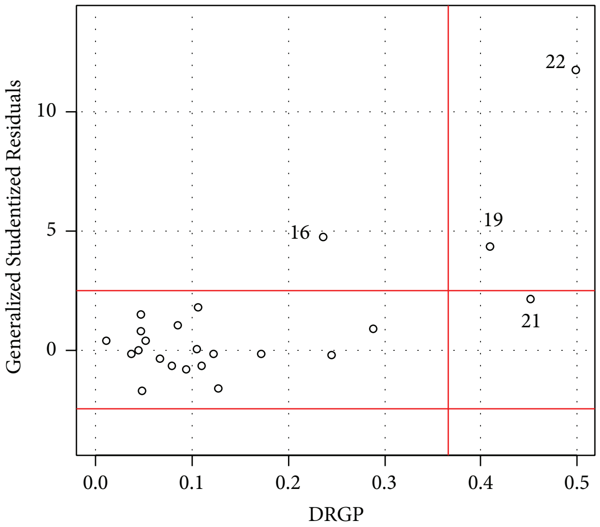


Figure 8: The Modified Generalized Studentized Residual against DRGP for Aircraft data.

6. Conclusion

In this paper, we proposed a new method for the identification of BLPs by means of a diagnostic plot. Most of the time, the classical OLS-MD plot fails to correctly identify the BLPs. The robust LMS-RMD plot is also not very successful in classifying observations into four categories. In this regard, we propose a new MGt-DRGP plot which is very successful in classifying observations into regular observations, vertical outliers, and good and bad leverage points. The Monte Carlo simulation study clearly shows that the MGt-DRGP plot can detect BLPs correctly with very low rates of masking and swamping. It is interesting to observe that the OLS-MD suffers from the masking problem and LMS-RMD suffers from the swamping problem.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] S. Chatterjee and A. S. Hadi, *Sensitivity Analysis in Linear Regression*, John Wiley & Sons, Ltd, New York, NY, USA, 1988.

[2] P. J. Rousseeuw and B. C. Van Zomeren, "Unmasking multivariate outliers and leverage points," *Journal of the American Statistical Association*, vol. 85, no. 411, pp. 633–639, 1990.

[3] M. Habshah, M. R. Norazan, and A. H. M. R. Imon, "The performance of diagnostic-robust generalized potentials for the identification of multiple high leverage points in linear

regression," *Journal of Applied Statistics*, vol. 36, no. 5, pp. 507–520, 2009.

[4] A. Bagheri and H. Midi, "Diagnostic plot for the identification of high leverage collinearity-influential observations," *Sort: Statistics and Operations Research Transactions*, vol. 39, no. 1, pp. 51–70, 2015.

[5] D. A. Belsley, E. Kuh, and R. E. Welsch, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, John Wiley & Sons, New York, NY, USA, 1980.

[6] P. J. Rousseeuw and A. M. Leroy, *Robust Regression and Outlier Detection*, Wiley Series in Probability and Mathematical Statistics, Wiley, New York, NY, USA, 1987.

[7] A. C. Atkinson, "Fast very robust methods for the detection of multiple outliers," *Journal of the American Statistical Association*, vol. 89, no. 428, pp. 1329–1339, 1994.

[8] A. H. M. R. Imon, "Identifying multiple high leverage points in linear regression," *Journal of Statistical Studies*, vol. 3, pp. 207–218, 2002.

[9] G. Pison and S. Van Aelst, "Diagnostic plots for robust multivariate methods," *Journal of Computational and Graphical Statistics*, vol. 13, no. 2, pp. 310–329, 2004.

[10] M. Hubert, P. J. Rousseeuw, and K. V. Branden, "ROBPCA: a new approach to robust principal component analysis," *Technometrics*, vol. 47, no. 1, pp. 64–79, 2005.

[11] M. Kamruzzaman and A. H. M. R. Imon, "High leverage point: another source of multicollinearity," *Pakistan Journal of Statistics*, vol. 18, no. 3, pp. 435–448, 2002.

[12] S. Chatterjee and A. S. Hadi, "Influential observations, high leverage points, and outliers in linear regression," *Statistical Science*, vol. 1, no. 3, pp. 379–393, 1986.

[13] A. S. Hadi, "A new measure of overall potential influence in linear regression," *Computational Statistics & Data Analysis*, vol. 14, no. 1, pp. 1–27, 1992.

[14] D. Peña and V. J. Yohai, "The detection of influential subsets in linear regression by using an influence matrix," *Journal of the Royal Statistical Society—Series B: Methodological*, vol. 57, no. 1, pp. 145–156, 1995.

[15] V. Barnett and T. Lewis, *Outliers in Statistical Data*, Wiley, New York, NY, USA, 3rd edition, 1994.

[16] A. Bagheri and H. Midi, "On the performance of the measure for diagnosing multiple high leverage collinearity-reducing observations," *Mathematical Problems in Engineering*, vol. 2012, Article ID 531607, 16 pages, 2012.

[17] A. H. M. R. Imon, "Identifying multiple influential observations in linear regression," *Journal of Applied Statistics*, vol. 32, no. 9, pp. 929–946, 2005.

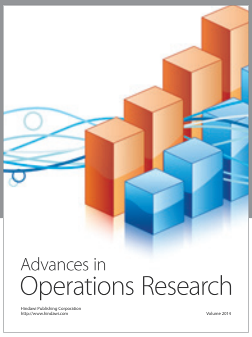
[18] R. D. Cook and S. Weisberg, *Residuals and Influence in Regression*, Chapman & Hall, New York, NY, USA, 1982.

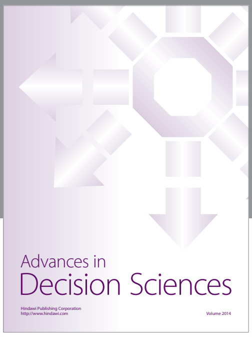
[19] C. R. Rao, *Linear Statistical Inference and Its Applications*, John Wiley & Sons, New York, NY, USA, 1965.

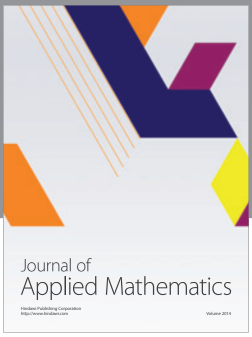
[20] M. Habshah and A. M. Mohammed, "Identification of good and bad high leverage points in multiple linear regression model," in *Proceedings of the 17th International Conference on Mathematical Methods, Computational Techniques and Intelligent Systems (Mamectis '15)*, pp. 147–153, Tenerife, Spain, January 2015.

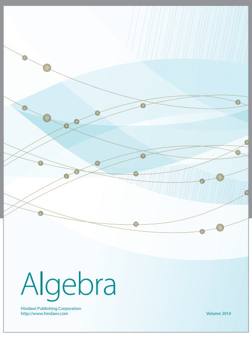
[21] G. W. Snedecor and W. G. Cochran, *Statistical Methods*, The Iowa State University Press, Ames, Iowa, USA, 6th edition, 1967.

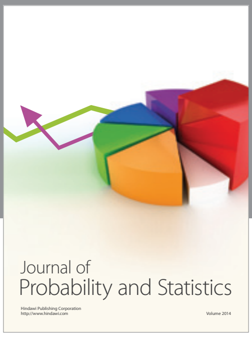
[22] J. B. Gray, "Graphics for regression diagnostics," in *Proceedings of the Statistical Computing Section*, pp. 102–107, American Statistical Association (ASA), Washington, DC, USA, 1985.


Advances in
Operations Research

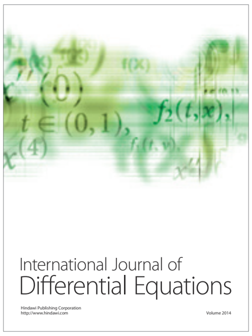

Advances in
Decision Sciences


Journal of
Applied Mathematics

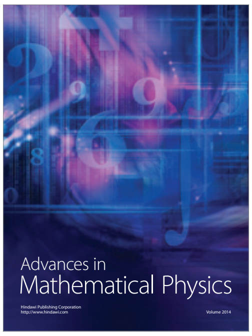

Algebra

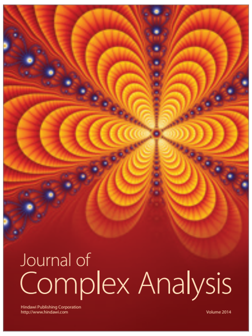

Journal of
Probability and Statistics

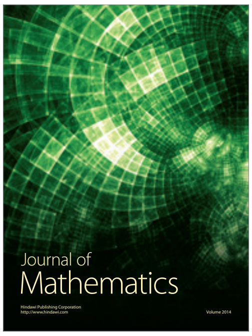
The Scientific
World Journal

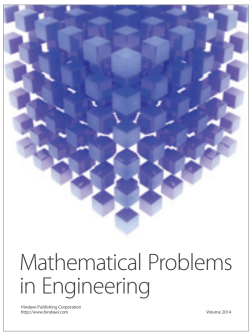

International Journal of
Differential Equations

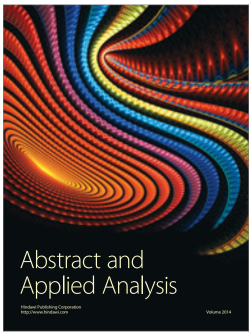
International Journal of
Combinatorics

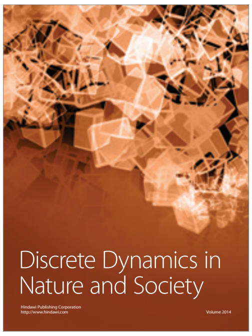

Advances in
Mathematical Physics

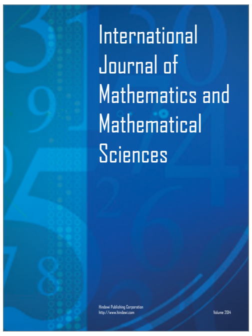

Journal of
Complex Analysis


Journal of
Mathematics


Mathematical Problems
in Engineering

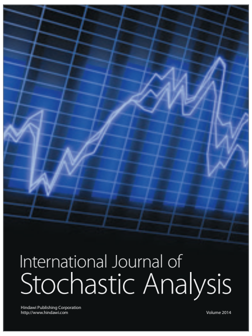

Abstract and
Applied Analysis


Discrete Dynamics in
Nature and Society


International
Journal of
Mathematics and
Mathematical
Sciences

Journal of
Discrete Mathematics

Journal of
Function Spaces


International Journal of
Stochastic Analysis

Journal of
Optimization