Sains Malaysiana 50(7)(2021): 2085-2094
http://doi.org/10.17576/jsm-2021-5007-22

Fast Improvised Influential Distance for the Identification of Influential Observations in Multiple Linear Regression

(Penambahbaikan Pantas Jarak Pengaruh bagi Pengecaman Cerapan Berpengaruh dalam Regresi Linear Berganda)

HABSHAH MIDI*, MUHAMMAD SANI, SHELAN SAIED ISMAEEL & JAYANTHI ARASAN

ABSTRACT

Influential observations (IO) are those observations that are responsible for misleading conclusions about the fitting of a multiple linear regression model. The existing IO identification methods such as influential distance (ID) is not very successful in detecting IO. It is suspected that the ID employed inefficient method with long computational running time for the identification of the suspected IO at the initial step. Moreover, this method declares good leverage observations as IO, resulting in misleading conclusion. In this paper, we proposed fast improvised influential distance (FIID) that can successfully identify IO, good leverage observations, and regular observations with shorter computational running time. Monte Carlo simulation study and real data examples show that the FIID correctly identify genuine IO in multiple linear regression model with no masking and a negligible swamping rate.

Keywords: Bad leverage point; good leverage point; influential distance; influential observations

ABSTRAK

Cerapan berpengaruh (IO) adalah cerapan yang bertanggungjawab ke atas kesimpulan yang mengelirukan bagi penyesuaian model regresi linear berganda. Kaedah pengecaman IO sedia ada seperti jarak berpengaruh (ID) tidak begitu berjaya untuk mengesan IO. Kami mengesyaki bahawa ID menggunakan kaedah yang kurang cekap dengan masa pengiraan yang panjang pada langkah awal bagi pengecaman cerapan IO. Tambahan pula, kaedah ini menunjukkan cerapan tuasan baik sebagai IO yang mengelirukan keputusan kajian. Dalam kertas ini, kami mencadangkan penambahbaikan jarak berpengaruh pantas (FIID) yang boleh mengecam IO, cerapan tuasan yang baik dan cerapan biasa dengan jayanya dengan masa pengiraan yang pantas. Kajian Monte Carlo simulasi dan contoh data sebenar menunjukkan bahawa FIID mengecam IO dalam model linear regresi berganda dengan betul tanpa penyorokan dan kadar limpahan yang sangat kecil.

Kata kunci: Cerapan berpengaruh; jarak berpengaruh; titik tuasan buruk; titik tuasan tinggi baik

INTRODUCTION

The existence of IO is inevitable in real data sets (Hampel et al. 2011). In the presence of IO the ordinary least squares (OLS) estimates become bias and loose the property of best linear unbiased estimates (BLUE). Belsley et al. (2004) stated that IO are those observations which either alone or together with several other observations have larger impact on the computed values of various estimates. Chatterjee and Hadi (1986) highlighted that high leverage points (outlying observations in X direction) are not always influential, and influential observations are not necessarily high leverage points. Since influential observations have a great influence on the values of various estimates, it is therefore very crucial to identify them and to take them into consideration when interpreting the results. Rousseeuw and Leroy (1987) noted that both the dependent and independent variables need to be considered when developing technique for identification of IOs. Ignoring or considering only one of them, may fail to identify multiple IOs (Rahmatullah Imon 2002; Rousseeuw & Leroy 1987).

Studentized residuals, Cook's distance and difference in fitted values (DFFITS) are the commonly used methods for identifying influential observations. Welsch (1980) recommended DFFITS as it combines both the leverage and the residual components. Even though DFFITS is successful in detecting single influential observation, it is not effective enough when a group of influential observations are present in a data (Rousseeuw & Leroy 1987).

Rahmatullah Imon (2005) developed the generalized version of DFFITS, denoted by GDFF which combined both the group deleted leverage and residual components. Although the GDFF can detect multiple IOs, it is not effective enough in identifying the exact number of IOs. It has a tendency of detecting lesser IOs as it should be and produce several masking of IOs. This is probably due to the determination of the initial basic subset of the GDFF which is not adequately effective in classifying the deletion and the remaining groups. Rahmatullah Imon's GDFF diagnostic measure uses the generalized potentials (Hawkins et al. 1984).

Recently, Nurunnabi et al. (2016) proposed new identification measure for IOs named influential distance (ID) based on group detection technique for the identification of multiple IOs. The technique has three major stages. The first stage identifies the suspected unusual observations using a method that we call group union method (GUM), the second stage identifies high leverage points (HLPs) and vertical outliers (VOs), and the third stage computes the ID using Mahalanobis distance (MD). The ID method is very good for the identification of IOs. However, the shortcoming of ID is that in the first stage it employed the union of five different detection methods (standardized studentized residual, standardized least median of squares (LMS) residuals, hat matrix, Cooks distance and difference in fits) for the identification of the suspected unusual observation. Some of these detection methods have been reported to have high rate of masking and swamping (Habshah et al. 2009). According to Hadi (1992), the choice of the initial suspected unusual observations is very important as it may lead to correct detection of the final IOs. Moreover, the computation of GUM method takes a lot of computer times. In addition, the ID method declared good leverage points as IOs and also has high rate of masking and swamping. Therefore, these shortcomings motivated us to propose a new version of ID which has fast computer running time and able to successfully differentiate between the regular observations, good leverage points and IOs.

Next section introduced the influential distance methods. The proposed fast improvised influential distance is described in the following section. The simulation study and real data examples are presented in the subsequent section. Last section presents the conclusion of the study.

INFLUENTIAL DISTANCE

Nurunnabi et al. (2016) proposed a technique for identifying multiple IOs called influential distance denoted by ID. The technique uses group deletion based-approach which is designed for the classification and identification of multiple HLPs, outliers, and IOs. The approach of group deletion is to extract clean subset of the data which is free from unusual observations and then test the outlyingness of the remaining data points relative to the clean subset. The way of determining the clean subset R with size (*n-d*) is by obtaining the deletion group D (group of suspected unusual cases of size *d*) based on GUM (union of the detected observation based on standardized studentized residual and/or standardized LMS residuals, leverage values or hat matrix, CDs, and DFFITS). For more details of these methods, one can refer to Atkinson (1986), Atkinson and Riani (2000), Belsley et al. (2004), Chatterjee and Hadi (2006), Cook (1998), Hadi and Simonoff (1993), Nurunnabi et al. (2016) and Rahmatullah Imon (2005).

The ID is then computed using MD based on the generalized studentized residual (GSR) and generalized leverage values (GLV). The algorithm for the computation of ID is summarized as follows:

Step 1 Obtain the group of suspected unusual observations to be deleted using GUM and index them by 'D', and index the remaining (*n-d*) observations by 'R'. Partition matrices of *X* and *Y* for both clean 'R' group and suspected unusual 'D' group as follows:

$$X = \begin{bmatrix} X_R \\ X_D \end{bmatrix}, \quad Y = \begin{bmatrix} Y_R \\ Y_D \end{bmatrix},$$

Step 2 Fit the linear model of the clean 'R' group and obtain the parameter estimate $\hat{\beta}_R$ given as,

$$\hat{\beta}_R = (X_R^T X_R)^{-1} X_R^T Y_R$$

Step 3 Compute the GSR (r_i^*) for the identification of multiple outliers as follows:

$$r_i^* = \begin{cases} \dfrac{r_{i(R)}}{\hat{\sigma}_{R-i}\sqrt{1-h_{ii(R)}}} & for \quad i \in R, \\ \dfrac{r_{i(R)}}{\hat{\sigma}_{R}\sqrt{1+h_{ii(R)}}} & for \quad i \in D, \end{cases}$$

where $r_{i(R)} = Y_i - X_i\, \hat{\beta}_R$, $h_{ii(R)} = x_i^T\, (X_R^T X_R\,)^{-1}\, x_i$ and $\hat{\sigma}_R$ is the standard error of the residual. The cutoff point for GSR is given by,

$$CP(r_i^*) = \text{median}(r_i^*) \pm 3\text{MAD}(r_i^*)$$

where $\text{MAD}(r_i^*) = median\{|r_i^* - median(r_i^*\,)|\}/0.6745$. Any observation corresponds to r_i^* that exceed the $CP(r_i^*)$ is considered as outlier.

Step 4 Compute the GLV (h_{ii}^*) for the identification of multiple HLPs as follows:

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

The cutoff point for GLV is given by,

$$CP\,(h_{ii}^*) = \text{median}\,(h_{ii}^*) + 3\text{MAD}(h_{ii}^*)$$

For any h_{ii}^* observation that exceed the $CP(h_{ii}^*)$ is considered as HLPs.

Step 5 Use Steps (3) and (4) to compute the ID based on MD for a two-column matrix of GSR in the first column and GLV in the second column. This matrix is denoted by *G*. The ID is now defined as,

$$ID_i = \sqrt{(G_i - \bar{G}_R)^T \Sigma_{G_R}^{-1} (G_i - \bar{G}_R)}, \qquad i = 1,2,\dots,n,$$

where $\bar{G}_R$ and $\Sigma_{G_R}^{-1}$ are the mean and inverse covariance matrix of the *R* group of *G* matrix, respectively. Nurunnabi et al. (2016) proposed a cutoff point for *ID* given by,

$$CPID = \sqrt{\frac{(n-1)p}{(n-p)} F_{\alpha(p,n-p),}} \qquad i = 1,2,\dots,n,$$

where p = 2 (number of variable in *G* matrix); n is the number of observations; and α is the level of significance (we use $\alpha = 0.025$).

Step 6 Plot a graph of Index versus ID. Any value of ID exceeds the CPID is declared as IO.

Step 7 The IO identified in Step (6) is shown clearly in the 'LRI' plot by plotting the ID in the GSR-GLV plot.

PROPOSED FAST IMPROVISED INFLUENTIAL DISTANCE

In this section, we proposed an improvised diagnostic method for identifying and classifying multiple IOs named fast improvised influential distance denoted by FIID, by adopting the ID method of Nurunnabi et al. (2016). As already mentioned, the GUM of Nurunnabi et al. (2016) that involves five different outlying observations detection methods for the identification of suspected unusual observations take a lot of computer times. Moreover, the ID has high rate of swamping and masking effect and also not efficient in identifying IOs because good high leverage points also identified as IOs. As such, we attempt to improvise their method by using more efficient methods to identify suspected unusual observations with less computer running time and also establish reliable method of classification of IOs. In doing so, we replace the GUM with LMS-RMD$_{\text{ISE}}$ (the union of standardized LMS and robust Mahalanobis distance (RMD) based on index set equality (ISE)). The RMD$_{\text{ISE}}$ is constructed using ISE of Lim and Habshah (2016) which is expected not only capable of reducing the effect of masking and swamping but also to have lesser computer running time. The ISE is reported to have very fast computer running time due to the simplicity of its algorithm. It is important to note that the ID method identifies IOs with longer computational running time and fail to separate the good leverage from IOs. It has been reported that some unusual observations may be influential but not in a negative way (Chatterjee & Hadi 1986; Mohammed et al. 2015). Therefore, identifying the good leverage is very important as they have no effect, instead they may contribute to the precision of the estimate. The good leverage point should not be considered as IOs. We employed the idea of Mohammed et al. (2015) for HLPs classification. Our proposed FIID only considered bad HLPs and vertical outliers as IOs. Firstly, we have to establish a classification plot which consists of six portions indicating regular observations (ROs), good leverage observations (GLOs) and IOs.

The algorithm for the detection and classification of observations into ROs, GLOs and IOs is summarized as follows:

Step 1 The suspected vertical outliers are detected using the LMS denoted as S set.

Step 2 Employ the RMD$_{\text{ISE}}$ to identify HLPs denoted as H set.

Step 3 The union of H set and S set will be considered as the group of suspected unusual observations and index them by 'D', and index the remaining (*n-d*) observations by 'R'. Partition matrices of *X* and *Y* for both clean 'R' group and suspected unusual 'D' group as follows:

$$X = \begin{bmatrix} X_R \\ X_D \end{bmatrix}, \quad Y = \begin{bmatrix} Y_R \\ Y_D \end{bmatrix},$$

Step 4 Fit the linear model of the clean R group and obtain the parameter estimate $\hat{\beta}_R$ given as,

$$\hat{\beta}_R = (X_R^T X_R)^{-1} X_R^T Y_R$$

Step 5 Compute the fast generalized studentized residual (FGSR) denoted by fr_i^* for the identification of multiple outliers as follows:

$$fr_i^* = \begin{cases} \dfrac{r_{i(R)}}{\hat{\sigma}_{R-i}\sqrt{1 - \mathrm{h}_{ii(R)}}} & for \quad i \in R, \\ \dfrac{r_{i(R)}}{\hat{\sigma}_R\sqrt{1 + \mathrm{h}_{ii(R)}}} & for \quad i \in D, \end{cases}$$

where $r_{i(R)} = Y_i - X_i\,\hat{\beta}_{(R)}$, $h_{ii(R)} = x_i^T (X_R^T X_R)^{-1} x_i$ and $\hat{\sigma}_R$ is the standard error of the residual. The cutoff point for FGSR is given by,

$$CP(fr_i^*) = \mathrm{median}(fr_i^*) \pm 3\mathrm{MAD}(fr_i^*)$$

where MAD(fr_i^*) = median{| fr_i^* - median (fr_i^*)|}/0.6745. Any observation fr_i^* that exceed the $CP(fr_i^{*)}$ is considered as outlier.

Step 6 Compute the fast generalized leverage values (FGLV) denoted by fh_{ii}^* for the identification of multiple HLPs as follows:

$$f\mathrm{h}_{ii}^* = \begin{cases} \dfrac{h_{ii(R)}}{1 - h_{ii(R)}} & for \quad i \in R, \\ \dfrac{\mathrm{h}_{ii(R)}}{1 + \mathrm{h}_{ii(R)}} & for \quad i \in D, \end{cases}$$

The cutoff point for fh_{ii}^* is given by,

$$CP(fh_{ii}^*) = \mathrm{median}(fh_{ii}^*) + 3\mathrm{MAD}(fh_{ii}^*)$$

For any fh_{ii}^* observations that exceed the $CP(fh_{ii}^*)$ is considered as HLPs.

Step 7 The fast-influential distance (FID_i^*) is formulated by formulating MD for a two-column matrix of FGSR in the first column and FGLV in the second column. This matrix is denoted by φ. The FID_i^* is now defined as,

$$FID_i^* = \sqrt{(\varphi_i - \bar{\varphi}_R)^T \Sigma_{\varphi_R}^{-1} (\varphi_i - \bar{\varphi}_R)}, \qquad i = 1,2,\dots,n,$$

where $\bar{\varphi}_R$ and $\Sigma_{\varphi_R}^{-1}$ are the mean and inverse covariance matrix of the *R* group of φ matrix, respectively. The cutoff point for FID_i^* is given by,

$$CPFID_i^* = \sqrt{\frac{(n-1)p}{(n-p)}} F_{\alpha(p,n-p),} \qquad i = 1,2,\dots,n,$$

where p = 2 (number of variable in φ matrix); n is the number of observations; and α is the level of significance (we use $\alpha = 0.025$). It is important to note that Nurunnabi et al. (2016) procedure declared any observation that correspond to ID larger than CPID as IO. The IO is confirmed by plotting ID on the plot of FGSR versus FGLV and sketch the confidence bound of each given by $\mathrm{CP}_{\mathrm{FGSR}}$ and $\mathrm{CP}_{\mathrm{FGLV}}$, respectively. Any observation falls outside the confidence bound declared as IO. By doing so, some good observations also declared as HLPs. This is the main weakness of Nurunnabi et al. (2016) detection of IOs in addition of taking longer computational running time. Therefore, relying on the ID or FID will lead to inaccurate identification of IOs. As such, we adopt the procedure given by Mohammed et al. (2015), and classify observations as follows:

An observation is declared as RO if; $|FGSR| \leq CP_{FGSR}$ and $|FGLV| \leq CP_{FGLV}$

An observation is declared GLO if; $|FGSR| \leq CP_{FGSR}$ and $|FGLV| > CP_{FGLV}$

An observation is declared as IO if; $|FGSR| > CP_{FGSR}$ and $|FGLV| \leq CP_{FGLV}$

An observation is declared as IO if; $|FGSR| > CP_{FGSR}$ and $|FGLV| > CP_{FGLV}$

With this additional step, IO is correctly identified whereby good leverage observations are not considered as IO, in contrast to Nurunnabi et al. (2016) procedure.

Subsequently, the proposed FIID is formulated in Figure 1, whereby the IOs are clearly separated from the regular and good leverage points.

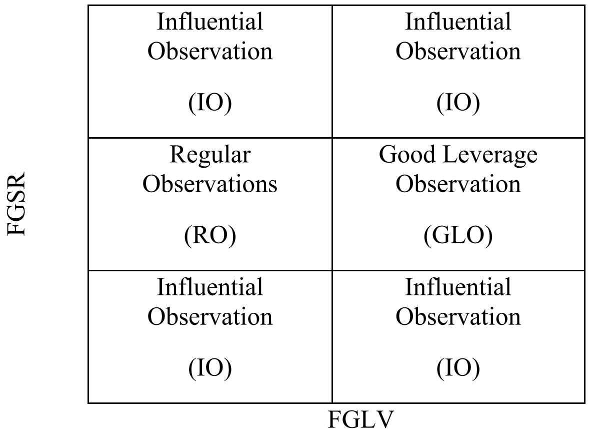
Influential Observation (IO)
Influential Observation (IO)
Regular Observations (RO)
Good Leverage Observation (GLO)
Influential Observation (IO)
Influential Observation (IO)
FGSR
FGLV

FIGURE 1. FGLV against Fast Generalized Studentized Residuals (FGSR)

SIMULATION STUDY RESULTS AND VERIFICATION

Monte Carlo simulation study is carried out to evaluate the performance of the proposed method.

First simulation study: In this section, we used a simulation study to assess our new proposed (LMS-RMD$_{\text{ISE}}$) and existing (GUM) methods. Following Lim and Habshah (2016), we generate a data with difference sizes, n = 20, 40, 60, 80, 100, and 200. A linear regression model with three independent variables (x_1, x_2 and x_3) are generated by uniform distribution U(0,10). The error term is generated from standard normal distribution. The response variable is generated as $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$ with the true coefficient $\beta = (1, 2, 3, 4)$. The simulation was replicated 5,000 times.

First contamination (HLPs): We contaminated the data by introducing HLPs at three different percentage levels (α = 0.05, 0.10 and 0.15). To generate HLPs in the dataset, the first 100α% observations of the regular data in x_1, x_2 and x_3 are generated by uniform distribution U(20,30) at different contamination levels for different samples.

Second contamination (VOs): The data was contaminated by introducing VOs at three different percentage levels (α = 0.05, 0.10 and 0.15). The VOs are generated in such a way that the first 100α% observations of the regular data in are replaced by uniform distribution U(20,30) at different contamination levels for different samples.

Tables 1 and 2 presents the simulation result of the number of suspected unusual observations (SUO) detected and running times to complete the simulation of 5,000 runs by GUM and LMS-RMD$_{\text{ISE}}$ for both HLPs and VOs contamination, respectively.

It is clearly seen that the number of SUO detected by LMS-RMD$_{\text{ISE}}$ and GUM are closed to each other. However, the SUO of LMS-RMD$_{\text{ISE}}$ is much closer to the actual number of HLPs compared to GUM. Moreover, the running time of LMS-RMD$_{\text{ISE}}$ is shorter than the GUM due to the excellent performance of ISE. Figure 2 presents the running time of GUM and LMS-RMD$_{\text{ISE}}$ for the case of HLPs contamination (Table 1). It is clearly seen that the LMS-RMD$_{\text{ISE}}$ has lower computer running time compared to GUM which also indicate that the algorithm of LMS-RMD$_{\text{ISE}}$ is more efficient than that of GUM.

TABLE 1. Running time and suspected unusual observations (SUO) detection of GUM and LMS-RMD$_{\text{ISE}}$ for HLPs contamination

Simple size	Actual no. of HLPs	GUM		LMS-RMD$_{\text{ISE}}$		Percentage reduction in running times
		Average Est. No. of SUO (Approx No.)	Running times (s)	Average Est. No. of SUO (Approx No.)	Running times (s)	
			5%			
20	1	2.6202 (3)	285	2.0916 (2)	10	96.49
40	2	3.5524 (4)	234	2.7052 (3)	13	94.44
60	3	3.5180 (4)	320	3.2474 (3)	15	95.31
80	4	4.4892 (4)	414	4.2312 (4)	17	95.89
100	5	6.0116 (6)	516	5.1162 (5)	19	96.32
200	10	11.0020 (11)	1074	10.1472 (10)	29	97.30
			10%			
20	2	2.4568 (2)	287	2.7834 (3)	10	96.52
40	4	4.6644 (5)	234	4.3740 (4)	12	94.87
60	6	6.5508 (7)	318	6.3832 (6)	15	95.28
80	8	8.6950 (9)	413	8.2306 (8)	17	95.88
100	10	10.7108 (11)	514	10.1140 (10)	19	96.30
200	20	21.0806 (21)	1067	20.1856 (20)	28	97.38
			15%			
20	3	3.6112 (4)	287	3.3014 (3)	10	96.52
40	6	6.6738 (7)	234	6.2536 (6)	12	94.87
60	9	9.5098 (10)	319	9.1632 (9)	14	95.61
80	12	12.5736 (13)	413	12.1424 (12)	16	96.13
100	15	15.4786 (15)	513	15.1160 (15)	18	96.49
200	30	31.1480 (31)	1068	30.0710 (30)	28	97.38

TABLE 2. Running time and suspected unusual observations (SUO) detection of GUM and LMS-RMD$_{ISE}$ for VOs contamination

Simple size	Actual no. of VOs	GUM		LMS-RMD$_{ISE}$		Percentage reduction in running times
		Average Est. No. of SUO (Actual No.)	Running times (s)	Average Est. No. of SUO (Actual No.)	Running times (s)	
			5%			
20	1	2.4209 (2)	330	2.2962 (2)	19	94.24
40	2	3.3244 (3)	234	2.3026 (2)	22	90.60
60	3	3.4827 (3)	317	3.3741 (3)	24	92.43
80	4	4.4243 (4)	416	4.3297 (4)	26	93.75
100	5	5.2621 (5)	515	5.2465 (5)	28	94.56
200	10	10.1203 (10)	1103	10.1372 (10)	37	96.65
			10%			
20	2	2.3543 (2)	331	2.2341 (2)	19	94.26
40	4	4.4437 (4)	234	4.3742 (4)	22	90.60
60	6	6.5811 (7)	318	6.4325 (6)	25	92.14
80	8	8.4150 (8)	417	8.3056 (8)	27	93.53
100	10	10.4875 (10)	516	10.2102 (10)	29	94.38
200	20	21.5212 (21)	1103	20.4506 (20)	38	96.55
			15%			
20	3	3.4018(3)	331	3.3707 (3)	19	94.26
40	6	6.4792(6)	235	6.3924 (6)	21	91.06
60	9	9.4008 (9)	319	9.3814 (9)	25	92.16
80	12	12.3692 (12)	417	12.2423 (12)	27	93.53
100	15	15.2752 (15)	516	15.1898 (15)	29	94.38
200	30	30.3956 (30)	1104	30.4106 (30)	40	96.38

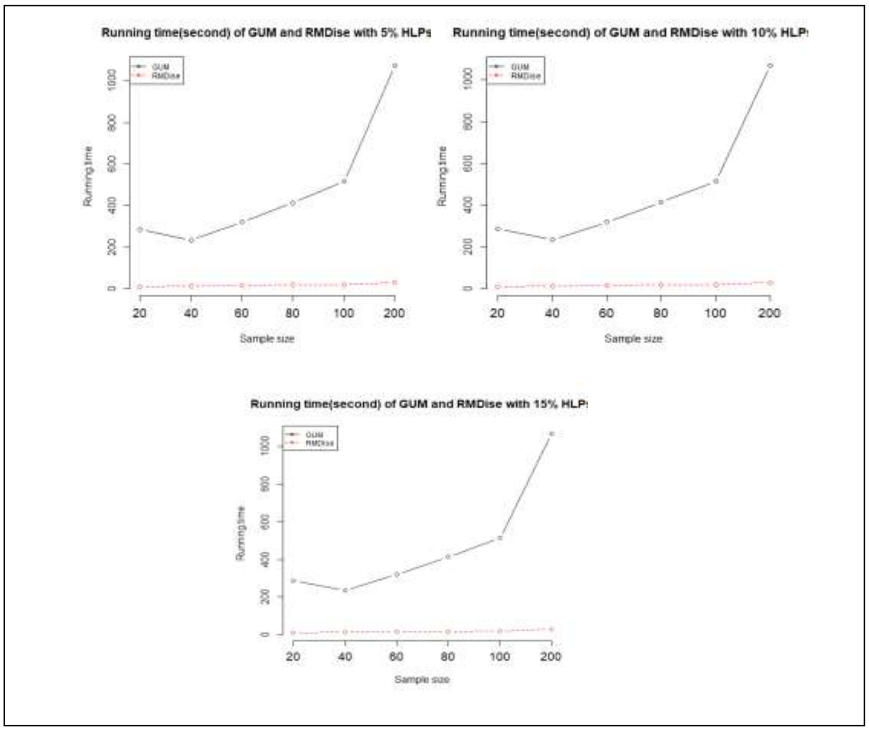

FIGURE 2. Running time GUM and LMS-RMD$_{ISE}$ for each sample size and contamination level for the case of HLPs contamination

Second simulation study: We used a Monte Carlo simulation study to assess the performance of our proposed IOs identification method, FIID and the existing ID method. The evaluation of these schemes is based on the rate of correct detection of IOs and the rate of masking and swamping effects. The best method is the one that has higher percentage of correct detection of IO's with smaller rate of masking and swamping. Following the designed experiment of Mohammed et al. (2015), the explanatory variables are generated randomly from a uniform distribution with mean zero and variance one. The response variable is generated from $y = X\beta + \varepsilon$ with $p = 3$ (number of explanatory variables). The true coefficient $\beta = (1,1,1,1)$ and $\varepsilon \sim N(0,1)$. In each experiment, different size of samples (n = 50, 100, 150 and 200) and different percentage of influential observations (α = 0.05, 0.10, 0.15 and 0.20) are considered. The influential observations are created at the position of the first 100α% observation for both X and y variables. To generate the influential observations, the first observation in each variable is kept fixed at 5 and the consecutive values are generated by multiplying the values index, j, by 5. The simulation was replicated 5,000 times.

Table 3 present the percentage of correct detection of IOs and masking and swamping rates for all possible combinations of n and α. As already explained, masking is declaring outliers as inliers and swamping is declaring inliers as outliers. It is interesting to observe that the proposed FIID method consistently display higher rate of correct detection of IOs with a smaller swamping and masking rates regardless of the size of n and α. The results of the study show that the FIID method performs better than ID method in correctly identification of IOs in multiple linear regression.

TABLE 3. Percentage of correct identification of influential observations, masking and swamping for simulation data

Cont. level	n	% Correct detection		% Masking		% Swamping	
		ID	FIID	ID	FIID	ID	FIID
5%	50	100	100	0	0.02	9.28	1.70
	100	100	100	0	0	6.04	0.64
	150	100	100	0	0	5.02	0.40
	200	100	100	0	0	6.05	1.26
10%	50	100	100	0	0	5.52	0.35
	100	87.56	100	12.44	0	10.06	1.92
	150	94.30	100	5.70	0	8.45	1.41
	200	82.36	100	17.64	0	11.01	2.30
15%	50	96.50	100	3.50	0	9.61	0.45
	100	76.02	100	23.98	0	15.78	3.82
	150	60.40	100	39.60	0	13.45	2.71
	200	53.21	100	46.79	0	11.67	2.65
20%	50	83.72	100	16.28	0	27.20	5.32
	100	56.32	100	43.68	0	21.75	7.28
	150	47.42	100	52.58	0	18.45	5.51
	200	43.56	100	56.44	0	22.86	4.64

NUMERICAL EXAMPLE

We use air craft data set taken from Gray (1985) to compare the performance of the proposed FIID method with the existing ID method. This data set contains 23 number of observations with 4 predictor variables (aspect ratio, lift to drag ratio, weight of the plane, and maximal thrust) and the cost being the response variable. Table 4 exhibits diagnostic measures and their cut-off points (in parenthesis). In the initial step both GUM and LMS-RMD$_{\text{ISE}}$ identified eight observations (10, 11, 14, 16, 17, 18, 19 and 22) as suspected unusual. Finally, the ID identifies observations 14, 16, 17, 18, and 22 as IOs as shown in Figure 3 (Index vs ID plot). However, the FIID identifies observation 16 and 22 as IOs and, 17 as good leverage observation which can be clearly seen in Figure 3 (GLV vs GSR plot). In order to justify which methods correctly identify the IOs, we applied the OLS method to the original data and the remaining data after omitting the IOs for both ID and FIID. The parameter estimates and standard error (S.E.) for each variable were computed. A good identification method is one which corresponds to the highest total percentage changes for various estimates. The percentage of change in estimator (PCE) is computed as,

$$\text{PCE} = \left|\frac{\hat{\alpha}_{Proposed} - \hat{\alpha}_{Original}}{\hat{\alpha}_{Original}}\right| \times 100\%$$

where $\hat{\alpha}_{Original}$ is the OLS parameter estimates of the original data; $\hat{\alpha}_{Proposed}$ is the OLS estimates of the remaining data after IOs are removed; and |.| is the absolute value.

Beside using PCE as a criterion of a good method, we also look at the standard errors of the estimates after removing the suspected IO. The method which has the smaller SE of the estimates after deleting the suspected IO can be considered a good method.

Table 5 shows the OLS parameter estimates and the standard errors of estimates (in parenthesis) for original and remaining data. It can be observed from this table that most parameter estimates associated with FIID identification method have the highest percentage of changed compared to ID method. Moreover, the standard error of estimates after removing the IO by FIID method is smaller than the ID. These results justified that the FIID method correctly identified the IOs, because the IOs are responsible for this changing in the results.

TABLE 4. Diagnostic measure for IOs (ID and FIID) for Aircraft data

Ind	Identification of suspected unusual observations									
	Group Union Method (GUM)									
	Std.Stud res. |2.50|	Std.LMS res. |2.50|	hii (0.522)	CD (0.870)	DF FITS |0.834|	RMDISE (5.531)	$r_i^* \approx fr_i^*$ (5.965, -5.296)	$h_{ii}^* \approx fh_{ii}^*$ (0.546)	ID_i (3.043)	*FIID*
1	0.500	-0.534	0.181	0.014	0.231	1.538	-0.456	0.240	1.437	0.993
2	0.929	0.019	0.150	0.037	0.383	1.706	0.306	0.229	1.247	1.252
3	0.764	1.332	0.147	0.024	0.309	1.503	1.531	0.163	0.531	1.901
4	-0.312	0.253	0.150	0.004	-0.126	1.893	-0.517	0.053	1.151	1.516
5	0.276	-0.092	0.084	0.002	0.077	1.273	-0.259	0.097	0.701	0.591
6	-0.113	0.491	0.227	0.001	-0.061	3.013	0.400	0.140	0.165	2.102
7	-0.424	-0.534	0.135	0.007	-0.160	2.070	-0.932	0.105	0.927	0.533
8	-0.301	-0.432	0.114	0.003	-0.102	2.516	-1.099	0.046	1.344	0.953
9	-0.024	2.395	0.242	0.000	-0.014	1.975	1.651	0.119	0.874	0.597
10	0.495	4.785	0.081	0.005	0.136	3.002	4.646	0.163	2.455	2.156
11	0.188	4.619	0.150	0.002	0.076	1.778	4.311	0.234	2.135	0.760
12	-1.128	-0.057	0.051	0.015	-0.245	1.920	-0.907	0.173	1.065	2.203
13	0.308	1.235	0.061	0.001	0.072	1.085	1.032	0.069	1.134	0.354
14	0.234	-5.809	0.875	0.636	1.568	19.120	-5.274	0.106	3.493	2.572
15	0.246	0.534	0.063	0.001	0.058	2.650	0.334	0.063	1.040	1.163
16	0.748	9.203	0.126	0.019	0.273	5.723	7.421	0.313	3.946	5.062
17	-2.069	-0.534	0.190	0.257	-1.124	5.537	-1.769	0.550	5.530	5.160
18	-0.497	4.299	0.099	0.006	-0.155	2.283	4.187	0.508	4.053	2.444
19	-0.925	3.715	0.100	0.022	-0.293	5.226	2.379	0.307	1.780	6.821
20	0.861	-0.374	0.152	0.033	0.358	5.344	-0.292	0.070	0.953	1.576
21	-1.758	1.684	0.120	0.099	-0.665	4.647	0.475	0.341	2.308	6.851
22	2.911	24.847	0.445	2.536	5.352	8.310	14.995	0.631	8.678	12.157
23	-0.912	-0.534	0.055	0.011	-0.204	2.118	-1.374	0.121	1.156	1.507

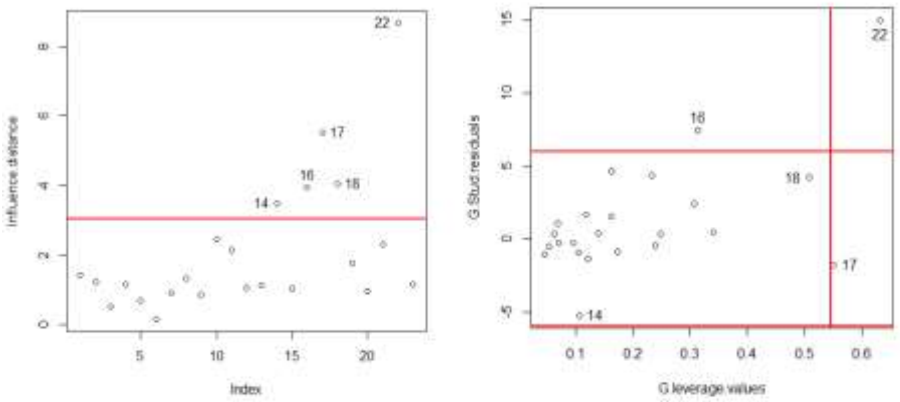

FIGURE 3. ID and FIID plots for Aircraft data set

TABLE 5. PCE values for ID and FIID based on OLS for Aircraft data set

Variables	Original data	Remaining data			
		Removed IOs by ID [cases (14,16,17,18,22)]		Removed IOs by FIID [cases (16,22)]	
	Estimation (S.E)	Estimation (S.E)	PCE	Estimation (S.E)	PCE
Aspect ratio	-3.853*** (1.763)	-2.752** (0.979)	28.58	-3.049*** (0.919)	20.90
Lift to Drag ratio	2.488* (1.187)	3.182* (1.770)	27.89	1.210* (0.649)	51.37
Weight of Plane	0.003*** (0.0005)	0.0012*** (0.0004)	60.00	0.001*** (0.0004)	66.74
Maximal Thrust	-0.002*** (0.0005)	-0.0006** (0.0003)	70.00	-0.001*** (0.0003)	50.00
Constant	-3.791 (10.116)	4.979 (6.180)	231.34	9.501 (5.578)	350.60 (44.93)
No. of obs.	23	18		21	
Residual Std. Error	8.406	4.424	47.37	4.349	48.26
Df	18	13		16	
F Statistics	34.17***	19.77***	42.14	18.95***	44.54

Note: * p < 0.1; ** p < 0.05; *** p < 0.01

Conclusion

In this paper, we proposed a method for identifying multiple influential observations and good leverage values in multiple linear regression named Fast Improvised Influential Distance (FIID). The proposed FIID is very successful in correctly identifying the influential observation and reducing the effects of swamping and masking compared to the existing method (ID) in this study. Another advantage of FIID is that it is very fast as it employs LMS-RMD$_{ISE}$ at the initial step for detecting the suspected unusual observations. The LMS-RMD$_{ISE}$ has a very fast computation time compared to GUM used by ID method due to its simplicity and computational ease. Therefore, the algorithm of FIID is simple and faster than that of ID. Hence, it is highly recommended to use the new proposed FIID method in the identification of influential observation in multiple linear regression.

ACKNOWLEDGEMENTS

This article was partially supported by the Fundamental Research Grant Scheme (FRGS) under the Ministry of Higher Education, Malaysia with project number FRGS/1/2019/STG06/UPM/01/1. The authors declare that there is no conflict of interests regarding the publication of this paper.

REFERENCES

Atkinson, A.C. 1988. Masking unmasked. *Biometrika* 73(3): 533-541.

Atkinson, A.C. & Riani, M. 2000. *Robust Diagnostic Regression Analysis*. New York: Springer-Verlag.

Belsley, D., Kuh, E. & Welsch, R. 2004. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Chatterjee, S. & Hadi, A.S. 2006. *Regression Analysis by Example*. 4th ed. Hoboken, New Jersey: John Wiley & Sons, Inc.

Chatterjee, S. & Hadi, A.S. 1986. Influential observations, high leverage points, and outliers in regression. *Statistical Science* 1(3): 379-393.

Cook, R.D. 1998. *Regression Graphic: Ideas for Studying Regression through Graphics*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Gray, J.B. 1985. Graphics for regression diagnostics. In *American Statistical Association Proceedings of Statistical Computing Section*. American Statistical Association. pp. 102-107.

Habshah, M., Norazan, M.R. & Rahmatullah Imon, A.H.M. 2009. The performance of diagnostic-robust generalized potentials for the identification of multiple high leverage points in linear regression. *Journal of Applied Statistics* 36(5): 507-520.

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

Hadi, A.S. & Simonoff, J. 1993. Procedure for the identification of outliers in linear models. *Journal of the American Statistics Association* 88(424): 1264-1272.

Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. & Stahel, W.A. 2011. *Robust Statistics: The Approach based on Influence Functions*. Hoboken, Ney Jersey: John Wiley & Sons, Inc.

Hawkins, D.M., Bradu, D. & Kass, G.V. 1984. Location of several outliers in multiple regression data using elemental sets. *Technometrics* 26(3): 197-208.

Lim, H.A. & Habshah, M. 2016. Diagnostic robust generalized potential based on index set equality (DRGP(ISE)) for the identification of high leverage points in linear models. *Computational Statistics* 31(3): 859-877.

Mohammed, A., Habshah, M. & Rahmatullah Imon, A.H.M. 2015. A new robust diagnostic plot for classifying good and bad high leverage points in a multiple linear regression model. *Mathematical Problems in Engineering* 2015: Article ID. 279472.

Nurunnabi, A.A.M., Nasser, M. & Rahmatullah Imon, A.H.M. 2016. Identification of multiple outliers, high leverage points and influential observations in linear regression. *Journal of Applied Statistics* 43(3): 509-525.

Rahmatullah Imon, A.H.M. 2005. Identifying multiple influential observations in linear regression. *Journal of Applied Statistics* 32(9): 929-946.

Rahmatullah Imon, A.H.M. 2002. Identifying multiple high leverage points in linear regression. *Journal of Statistical Studies* 3: 207-218.

Rousseeuw, P.J. & Leroy, A.M. 1987. *Robust Regression and Outlier Detection*. Wiley series in probability and mathematical statistics. Hoboken, New Jersey: John Wiley & Sons, Inc.

Welsch, R.E. 1980. Regression sensitivity analysis and bounded-influence estimation. In *Evaluation of Econometric Models,* edited by Kemnta, J. & Ramsey, J.B. New York: Academic Press, Inc. pp. 153-167.

Habshah Midi* & Jayanthi Arasan
Department of Mathematics
Faculty of Science and Institute for Mathematical Research
Universiti Putra Malaysia
43400 UPM Serdang, Selangor Darul Ehsan
Malaysia

Muhammad Sani
Department of Mathematical Sciences
Federal University Dutsin-Ma
Katsina State
Nigeria

Shelan Saied Ismaeel
Department of Mathematics
Faculty of Science
University of Zakho, Zakho
Iraq

*Corresponding author; email: habshahmidi@gmail.com

Received: 5 February 2020
Accepted: 19 November 2020