Sains Malaysiana 52(12)(2023): 3589-3602
http://doi.org/10.17576/jsm-2023-5212-19

An Efficient Method of Identification of Influential Observations in Multiple Linear Regression and Its Application to Real Data

(Kaedah yang Cekap bagi Pengecaman Cerapan Berpengaruh dalam Model Regresi Linear Berganda dan Kegunaannya dalam Set Data Sebenar)

HABSHAH MIDI[1,*], HASAN TALIB HENDI[1], HASSAN URAIBI[2], JAYANTHI ARASAN[3] & SHELAN SAIED ISMAEEL[4]

[1]*Institute for Mathematical Research, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia*
[2]*Department of Statistics, University of Al-Qadisiyah, IRAQ*
[3]*Department of Mathematics & Statistics, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia*
[4]*Department of Mathematics, Faculty of Science, University of Zakho, Iraq*

Received: 20 June 2023/Accepted: 14 November 2023

ABSTRACT

Influential observations (IOs) are those observations which either alone or together with several other observations have detrimental effect on the computed values of various estimates. As such, it is very important to detect their presence. Several methods have been proposed to identify IOs that include the fast improvised influential distance (FIID). The FIID method has been shown to be more efficient than some existing methods. Nonetheless, the shortcoming of the FIID method is that, it is computationally not stable, still suffers from masking and swamping effects, time consuming issues and not using proper cut-off point. As a solution to this problem, a new robust version of influential distance method (RFIID) which is based on Reweighted Fast Consistent and High Breakdown (RFCH) estimator is proposed. The results of real data and Monte Carlo simulation study indicate that the RFIID able to correctly separate the IOs from the rest of data with the least computational running times, least swamping effect and no masking effect compared to the other methods in this study.

Keywords: Good leverage point; influential distance; influential observations; Reweighted Fast Consistent and High Breakdown (RFCH) estimator

ABSTRAK

Cerapan berpengaruh (IO) ditakrifkan sebagai cerapan sama ada bersendirian atau bersama dengan beberapa cerapan lain yang mempunyai kesan memudaratkan ke atas nilai kiraan pelbagai anggaran. Oleh itu, sangat penting untuk mengecam kehadiran cerapan berpengaruh. Beberapa kaedah telah dicadangkan untuk mengecam IO termasuk kaedah penambahbaikan jarak berpengaruh pantas (FIID). Kaedah FIID telah ditunjukkan lebih cekap dibandingkan dengan kaedah sedia ada. Walau bagaimanapun, kaedah FIID mempunyai kelemahan iaitu pengiraannya tidak stabil, masih mempunyai kesan penyorokan dan limpahan, isu masa pengiraan yang panjang dan tidak menggunakan titik genting yang betul. Kaedah teguh versi baharu bagi jarak berpengaruh yang berasaskan penganggar berpemberat konsisten pantas dan titik musnah tinggi (RFIID) dicadangkan untuk mengatasi masalah ini. Keputusan data sebenar dan kajian simulasi Monte Carlo menunjukkan RFIID berupaya untuk mengasingkan IO daripada keseluruhan data dengan masa pengiraan paling singkat, kesan limpahan paling kecil tanpa kesan penyorokan dibandingkan dengan kaedah lain dalam kajian ini.

Kata kunci: Cerapan berpengaruh; jarak berpengaruh; penganggar pantas tekal berpemberat dan titik musnah tinggi; titik tuasan baik

INTRODUCTION

Belsley, Kuh and Welsch (2004) noted that influential observations (IOs) are those observations which either alone or together with several other observations have detrimental effect on the computed values of various estimates. Such observations are responsible for misleading conclusions about the fitting of a multiple linear regression model. It is generally believe that IOs are

outlying observations in the *X*-space or *Y*-space. Outlying observations in the *X*-space is referred as high leverage points (HLPs). However, according to Chatterjee and Hadi (1986), IOs are not always HLPs and *vice versa*.

It is worth mentioning that the values of numerous estimates are much affected by IOs and the ordinary least squares (OLS) method no longer has the optimal property in their presence. Hampel et al. (2011) pointed out that there is no guarantee that IOs are absence in a real dataset. Hence, it is very imperative to detect them and to take them into account when interpreting the results of statistical analysis. Various methods are available in the literature for the detection of IOs, but the commonly used techniques are the Studentized residuals, Cook's distance and difference in fitted values (DFFITS). DFFITS was recommended by Welsch (1980) because it incorporates both the leverage and residual components. Rousseeuw and Leroy (1987) noted that the DFFITS is very successful in identifying single influential observation, but fails to detect multiple influential observations.

When establishing an approach to determine IOs, both the dependent and independent variables should be taken into consideration. According to Rahmatullah Imon (2002) and Rousseeuw and Leroy (1987), failing to do that may result in inaccurate detection of IOs and will lead to misleading interpretation. As such, Rahmatullah Imon (2005) combined both the group deleted leverage and residual components in the establishment of the generalized version of DFFITS, which is called the generalized distance and difference in fitted values (GDFF). Despite the fact that the GDFF can spot multiple IOs, it is still not very successful in detecting IOs as it still suffers from masking and swamping of IOs. The reason for this shortcoming is most likely owing to the GDFF's basic subset being insufficiently successful in separating the deletion and the remaining groups. Due to this deficiency, a new identification measure for IOs which is called influential distance (ID) is put forward by Nurunnabi, Nasser and Imon (2016). This method comprises of three main steps whereby group union method (GUM) is employed to detect suspected unusual observations in the first step. The GUM utilised the union of five different identifying methods such as the standardized studentized residual, standardized least median of squares (LMS) residuals, hat matrix, Cooks distance and difference in fits. High leverage points (HLPs) and vertical outliers (VOs) are identified in the second step, followed by calculating the ID using Mahalanobis distance (MD) in the final step. The performance of ID method is better than the GDFF in terms of IOs detection. Nonetheless the ID suffers from swamping and masking effect due to using inefficient detection methods in GUM which have been proven by Habshah, Norazan and Rahmatullah Imon (2009) to have such problems. Moreover, in their algorithm, they did not consider separating the observations into good and bad high leverage points (HLPs). Consequently, by ignoring this step, some of good observations are detected as suspected IOs (swamping) while some bad observations are not declared as IOs (masking). This will ultimately effect the final results of ID. Another weakness of ID is that it also suffers from long computational running times due to using GUM.

As a solution to this problem, Habshah, Muhammad and Ismaeel (2021) proposed a new version of ID (FIID) which has less computational running times. It is quite successful in detecting IOs with less swamping and masking effect. However, the FIID is still inefficient with regard to computation running times and detection of IOs because it is based on index set equality (ISE). It is now evident that ISE is unstable as its algorithm depends on the selected initial subset, h. Habshah et al. (2020) exemplified that the final estimator of location and scatter of (ISE) is equivalent to minimum covariance determinant (MCD) if the same initial subset is utilized, otherwise the results will be quite different. Moreover, the FIID used improper cut-off point, i.e., using (1-α) quantile of F distribution based on the assumption that the p-dimensional variables follow a multivariate normal distribution. However, in a real life problem, there is no guarantee that data would come from a multivariate normal distribution.

To overcome these weaknesses, we propose a new version of identifying IOs called the robust influential distance method which is based on the Reweighted Fast Consistent and High Breakdown estimator (RFCH), denoted as RFIID. The RFIID is the extension work of FIID. To reduce the computational running times and to improve the correct detection of HLPs with the least swamping effect, the RFCH is incorporated in the algorithm of RFIID in order to estimate the mean vector and variance covariance matrix of the predictor variables. It has been proven by Olive and Hawkin (2010) that the location and scatter estimators of RFCH were consistent estimators. Furthermore, since the distribution of RFIID is intractable, as per Habshah, Norazan and Rahmatullah Imon (2009), Rashid et al. (2022a), and Zahariah and Habshah (2023), we propose using a confidence bound type cut-off point for RFIID instead of using (1- α) quantile of F distribution as a cut-off point.

Hence, the objectives of this study were as follows: (1) to develop a new method of identification of influential observations by incorporating robust mahalanobis

distance (RMD) based on the RFCH estimators; (2) to establish a new algorithm of classifying observations into three groups including regular observations, good leverage points and influential observations; (3) to propose using a confidence bound type cut-off point for RFIID; (4) to assess the performance of the proposed RFIID method compared to ID and FIID by using simulation study; and (5) to apply the proposed method to real dataset, namely the Gunst and Mason data.

The rest of the paper is structured as follows. The Reweighted Fast Consistent and High breakdown (RFCH) estimator will be introduced in the next section. The proposed robust and fast improvised influential distance (RFIID) is described in the following section. The subsequent section will present the simulation study and real data examples. Finally, the conclusion of the study is presented in the last section.

REWEIGHTED FAST CONSISTENT AND HIGH BREAKDOWN (RFCH) ESTIMATORS

Mahalanobis Distance (MD) measures how far each point is from the centroid of all points for the independent variables. It can be used to determine whether or not a sample contain an outlier or whether or not a process is in control (Habshah & Shabbak 2011). Let present the i^{th} vector of independent variables as:

$$x_i' = (1, x_1, x_2, \dots, x_P) = (1, t_i),$$

where t_i is a p-dimensional row vector. The vector of the mean and the variance covariance matrix can be calculated, respectively as:

$$\bar{t} = 1/n \sum_{i=1}^{n} t_i$$

$$C = \left(\frac{1}{n-1}\right) \sum_{i=1}^{n} (t_i - \bar{t}\,)\,(t_i - \bar{t}\,)',$$

The MD for each observation is defined as follows:

$$MD_i = \sqrt{\left(t_i - T(X)\right)' C(X)^{-1} \left(t_i - T(X)\right)} \qquad (1)$$

$$i = 1,2, \dots, n,$$

where $T(X)$ is the mean vector ($\bar{t}$) and $C(X)$ is the variance covariance matrix (C). The MD is one of the most commonly used measures for the detection of HLPs. However, it is now evident that the MD is not very successful in the detection of HLPs because it is based on the classical mean vector, $T(X)$ and classical covariance matrix, $C(X)$ which is very sensitive to outliers. To remedy this problem, Rousseeuw and Leroy (1987) proposed substituting the classical mean vector, $T(X)$ and classical covariance matrix, $C(X)$ of MD_i with robust estimates such as the Minimum Volume Ellipsoid (MVE) and the Minimum Covariance Determinant (MCD) (Rousseeuw & Yohai 1984), and renamed the MD in Equation (1) as Robust Mahalanobis Distance (RMD). Other robust estimators of location and scatter can be employed to compute RMD. The Reweighted Fast Consistent and High breakdown (RFCH) estimator established by Olive and Hawkins (2010) has been proven to be a very fast estimators of location and scatter. The $\sqrt{n}$ consistent estimator of Devlin Ganadesikan Kettenring (DGK) (Devlin, Gnanadesikan & Kettenring 1981) and high breakdown Median Ball (MB) (Olive & Hawkins 2008) estimator are used as attractors in the algorithm of RFCH which can be presented in the following three stages:

STAGE 1: The Algorithm of DGK

Step 1 Compute the initial Mahalanobis Distance

$$MD_{i0,DGK} = \sqrt{\left(t_i - T_{0,Start}\right)' (C_{0,Start})^{-1} (t_i - T_{0,Start})}, \qquad (2)$$

$$i = 1,2, \dots, n.$$

where the initial points ($T_{0,Start}$, $C_{0,Start}$) is based on the computed values of p-dimensional row vector of location and the ($p \times p$) covariance matrix, ($T(X)$, $C(X)$) of the original data.

Step 2 The $MD_{i0,DGK}$ in Step 1 is then sorted in increasing order to compute its median, $MED = median(MD_{i0,DGK})$ Next, a new dataset which comprise of m observations are defined such that their Mahalanobis Distance values are less than MED:

$$\tilde{X}_{1,DGK} = \left\{X_{jl} : MD_{i0,DGK} \le MED\right\},\ j = 1,2, \dots, m; \qquad (3)$$

$$l = 1,2, \dots, p,$$

where p is the number of independent variables.

Step 3 Acquire the first attractors ($T_{1,DGK}$), $C_{1,DGK}$ by calculating the location and scatter estimators for the $\tilde{X}_{1,DGK}$ dataset.

Step 4 Compare $C_{1,DGK}$ with $C_{0,Start}$. Stop the process if the diagonal elements of $C_{1,DGK} = C_{0,Start}$, otherwise repeat

Steps 1 to 3 until convergence. The final location and scatter estimates ($T_{K,DGK}$,$C_{K,DGK}$) is obtained from the $\tilde{X}_{K,DGK}$, where K is final step at which convergence takes place.

STAGE 2: The Algorithm of MB

Step 1 Let an identity matrix, I_p be the scatter matrix, i.e., $C = I_p$. Next, calculate the Mahalanobis Distance:

$$MD_i = \sqrt{(t_i - Med(X))' (C)^{-1}(t_i - Med(X))}\ , \quad i = 1,2,\dots,n, \tag{4}$$

where $Med\ (X)$ = median (X).

Let define the cut-off point of MD_i, i.e., *Lcut* :

$$Lcut = \text{median}(MD_i) \tag{5}$$

where $Lcut \neq 0.5$. Next, establish a new dataset, $\tilde{X}_0$ which contain only m observations such that it's MD_i is less than or equal to the *Lcut,*

$$\tilde{X}_0 = \{X_{jl}: MD_i \leq Lcut\},\ j = 1,2,\dots,m;\ l = 1,2,\dots,p. \tag{6}$$

Step 2 Compute the initial Mahalanobis Distance.

$$MD_{i0,MB} = \sqrt{(t_i - T_{0,Start})'(C_{0,Start})^{-1}(t_i - T_{0,Start})}\ , \quad i = 1,2,\dots,n, \tag{7}$$

where $T_{0,Start}$ is the p- dimensional row vector of location and $C_{0,Start}$ is the $(p \times p)$ covariance matrix calculated based on the new dataset, $\tilde{X}_0$.

Establish a new dataset ($\tilde{X}_{1,MB}$) based on a new cut-off point:

$$\tilde{X}_{1,MB} = \{X_{jl}: MD_{0i,MB} \leq Lcut0\},\ j = 1,2,\dots,m;\ l = 1,2,\dots,p. \tag{8}$$

where $Lcut0$ = median $(MD_{i0,MB})$.

Step 3 Compute the attractor ($T_{1,MB}$,$C_{1,MB}$) based on the new dataset ($\tilde{X}_{1,MB}$) in Step 2.

Step 4 Stop the process if the diagonal elements of $C_{1,MB} = C_{0,Start}$. Otherwise, recompute the $MD_{1,MB}$ based on the attractor ($T_{1,MB}$,$C_{1,MB}$). Repeat Steps 2 to 3, until convergence. At convergence, the final attractor and the final remaining set is represented as ($T_{K,MB}$,$C_{K,MB}$) and $\tilde{X}_{K,MB}$, respectively.

STAGE 3: The Algorithm of the RFCH

In the first step of RFCH, an initial attractors (T_{FCH},C_{FCH}^{*}) are determined based on the final attractors of DGK and MB estimators.

Step 1 Define T_{FCH} and C_{FCH} as follows:

$$T_{FCH} = \begin{cases} T_{K,DGK} & if \quad \sqrt{|C_{K,DGK}|} < \sqrt{|C_{K,MB}|} \\ T_{K,MB} & Otherwise \end{cases} \tag{9}$$

$$C_{FCH} = \begin{cases} \frac{Med\left(MD_i(T_{K,DGK},C_{K,DGK})\right)}{x^2_{(p,0.5)}} \times C_{K,DGK}, & if \quad \sqrt{|C_{K,DGK}|} < \sqrt{|C_{K,MB}|} \\ \frac{Med\left(MD_i(T_{K,MB},C_{K,MB})\right)}{x^2_{(p,0.5)}} \times C_{K,MB}\ , & Otherwise \end{cases} \tag{10}$$

and

$$C^{*}_{FCH} = \frac{Med\left(MD_i(T_{FCH},C_{FCH})\right)}{x^2_{(p,0.5)}} \times C_{FCH}.$$

Here, $x^2_{(p,0.5)}$ denotes the chi-square distribution with p degrees of freedom and a significance level of 0.5. Olive and Hawkins (2010) noted (in Theorem 1), that the (T_{FCH},C_{FCH}^{*}) are consistent estimators.

Step 2 A new dataset, $\tilde{X}_{FCH}$ is constructed as follows:

$$\tilde{X}_{FCH} = \left\{X_{jl}: MD_i(T_{FCH}, C^{*}_{FCH}) \leq x^2{}_{(p,1-\alpha)}\right\}, \quad j = 1,2,\dots,m; l = 1,2,\dots,p \tag{11}$$

where MD_i (T_{FCH},C_{FCH}^{*}) represents the Mahalanobis Distance based on the location and scatter (T_{FCH},C_{FCH}^{*}) obtained from Step 1. Based on the new dataset, $\tilde{X}_{FCH}$, calculate the RFCH attractors, ($T_{1,RFCH}$,$C_{1,RFCH}$). As in Theorem 1 of Olive and Hawkins (2010), $C^{*}_{1,RFCH}$ can be written as follows

$$C^{*}_{1,RFCH} = \frac{Med\left(MD_i(T_{1,RFCH},C_{1,RFCH})\right)}{x^2_{(p,0.5)}} \times C_{1,RFCH}. \tag{12}$$

Again, compute the Mahalanobis Distance based on ($T_{1,RFCH}$,$C_{1,RFCH}$) and reconstruct a new dataset;

$$\tilde{X}_{2,RFCH} = \left\{X_{jl}: MD_i\left(T_{1,RFCH}, C^{*}_{1,RFCH}\right) \leq x^2{}_{(p,1-\alpha)}\right\}, \quad j = 1,2,\dots,m; l = 1,2,\dots,p. \tag{13}$$

Repeat the same process, and applying Theorem 1 of Olive and Hawkins (2010), compute ($T_{2,RFCH}$,$C_{2,RFCH}$) estimators based on the $\tilde{X}_{2,RFCH}$ dataset and define $C^{*}_{2,RFCH}$

as follows:

$$C^*_{2,RFCH} = \frac{Med\left(MD_i(T_{2,RFCH}, C_{2,RFCH})\right)}{x^2_{(p,0.5)}} \times C_{2,RFCH}. \quad (14)$$

Step 3 Repeat Steps 1 to 2, *K* times until convergence. Stop the process if the number of detected outliers or HLPs based on $MD_i(T_{K,RFCH}, C_{K,RFCH})$ is equal to the number detected by $MD_i(T_{K\text{-}1,RFCH}, C^*_{K\text{-}1,RFCH})$. Observations corresponding to MD_i,,. larger than $x^2_{(p,0.5)}$ are considered as HLPs. As proven by Olive and Hawkins (2010), upon convergence, $(T_{K,RFCH}, C_{K,RFCH})$ are High Breakdown (HB) $\sqrt{n}$ consistent estimators.

THE PROPOSED ROBUST AND FAST IMPROVISED INFLUENTIAL DISTANCE (RFIID)

Group deletion based-approach were employed in the algorithm of ID for the classification and detection of multiple HLPs, outliers, and IOs. At the outset, the GUM (union of the detected observation based on standardized studentized residual and/or standardized LMS residuals, leverage values or hat matrix, CDs, and DFFITS) was employed to separate the clean subset R with size (*n-d*) and suspected IOs of group D of size *d.* Subsequently, the remaining data points relative to the clean subset were tested for outlyingness. Afterwards, the Mahalanobis Distant (MD) based on the generalized studentized residual (GSR) and generalized leverage values (GLV) were utilized for calculating the ID. The weakness of ID has already been discussed in the preceding section where it suffers from long computation running times, masking and swamping effect due to employing unreliable methods of detection of unusual observations. As a solution to this problem, Habshah, Muhammad and Ismaeel (2021) proposed an alternative method of detection of IOs (FIID) by employing more effective algorithm to detect suspected IOs with less computer running time. In their algorithm of FIID, they employed robust mahalanobis distant (LMS-RMD_{ISE}) based on index set equality (ISE) which is proven to be more effective than the ID. Nonetheless, through our investigation, the FIID still suffers from swamping effect and computationally not stable due to using RMD-ISE as reported in Habshah, Talib, Jayanthi and Urabi (2020). Furthermore, the FIID used improper cut-off point, i.e., F distribution based on the assumption that the *p*-dimensional variables follow a multivariate normal distribution. However, in a real life problem, there is no guarantee that data would come from a multivariate normal distribution.

We anticipate that the computation running time and the percentage of correct detection of the FIID can be improved. As such, another improved version of FIID which is called Robust and Fast Improvised Influential Distance (RFIID) is put forward by incorporating the RMD based on Reweighted Fast Consistent and High breakdown (RFCH) estimators (Olive & Hawkins 2010). The suspected IOs are detected based on LMS-RMD_{RFCH}. The RFCH is very fast and computationally very stable. The attractive feature of this estimator is that its location and scatter estimators are $\sqrt{n}$ consistent estimators as proven by Olive and Hawkin (2010). Moreover, by classifying observations into six groups based on Mohammed, Habshah and Rahmatullah Imon (2015) algorithm with slide modifications, able to correctly separate the IOs (vertical outliers and bad HLPs) from the rest of the observations. Since the distribution of RFIID is intractable, following Habshah, Norazan and Rahmatullah Imon (2009), Rashid et al. (2022), and Zahariah and Habshah (2023), a confidence bound type cut-off point is employed.

The RFIID can be summarized as the following:

Step 1 The standardized least median of squares (LMS) is employed to detect suspected vertical outliers indicated as V set.

Step 2 Identify the suspected HLPs by using Robust Mahalanobis Distance based on RFCH

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

where the (T_{RFCH}) and (C_{RFCH}), is the location and scatter estimators of RFCH, respectively.

The cut-off point for RMD is given as:

$$CPRFID^*_i = \text{median}\ (RFID^*_i) + 3\text{MAD}(RFID^*_i),$$

where

$$MAD(RMD_i) = median\ (abs\ (RMD_i - median\ (RMD_i)))/0.6745.$$

We declare that any i^{th} case with $RMD_i > cut\_off$ point, is the suspected HLPs and include them in a deletion group, denoted as L Set.

Step 3 The suspected unusual observations, denoted as 'D' set is defined as the union of V and L sets, while the remaining or clean (n-d) observations is indicated as 'R' set. The breaking up of the X and Y matrices for both the clean 'R' group and the suspected IOs 'D' group are presented as the following:

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

Step 4 For the clean 'R' group, estimate the parameters of the linear model, $\hat{\beta}_R$ as follows

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

Then compute the residual, $r_{i(R)} = Y_i - X_i \hat{\beta}_R$.

Step 5 The Robust and fast generalized Studentized residual (RFGSR), indicated by Rfr_i^* is then computed for the detection of multiple outliers

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

where $\hat{\sigma}_R$ represent the residual standard error, and the diagonal element of the hat matrix is given by $\mathrm{h}_{ii(R)} = x_i^T (X_R^T X_R)^{-1} x_i$.

Step 6 Since the distribution of Rfr_i^* is intractable, following Habshah, Norazan and Rahmatullah Imon (2009) and Rashid et al. (2022), the confidence bound type cutoff point for Rfr_i^* is utilized,

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

where $\mathrm{MAD}(Rfr_i^*) = \mathrm{median}\{|\, Rfr_i^* - \mathrm{median}(Rfr_i^*)|\}/0.6745$.

An observation is then declared as an outlier if Rfr_i^* value is greater than the CP(Rfr_i^*)

Step 7 For the identification of multiple HLPs, the robust fast-generalized leverage values (RFGLV), indicated by $Rf\mathrm{h}_{ii}^*$ is calculated as follows;

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

Step 8 Similar to Step 6, the following cut-off point for $Rf\mathrm{h}_{ii}^*$ is used;

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

An observation is considered HLPs if $Rf\mathrm{h}_{ii}^*$ value is greater than the $\mathrm{CP}(Rf\mathrm{h}_{ii}^*)$

Step 9 Compute Mahalanobis distance for a two-column matrix and denote this matrix as φ, where RFGSR (Step 5) is in the first column and RFGLV (Step 7) is in the second column. This Mahalanobis distance is named the Robust and Fast-Influential Distance ($RFID_i^*$) and is given by:

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

where $\bar{\varphi}_R$ and $\Sigma_{\varphi_R}^{-1}$ are the mean and inverse covariance matrix of φ, respectively.

Since it is not easy to proof the distribution of $RFID_i^*$, confidence bound type cutoff point is again utilized as in Habshah, Norazan and Rahmatullah Imon (2009) and Rashid et al. (2022),

$$CPRFID_i^* = \mathrm{median}\,(RFID_i^*) + 3\mathrm{MAD}(RFID_i^*)$$

It is not sufficient to declare that any observation that correspond to $RFID_i^*$ that exceeds its cutoff points, $CPRFID_i^*$ is the suspected IOs as pointed out by Nurunnabi, Nasser and Imon (2016) in their decision for suspected IOs. Nurunnabi, Nasser and Imon (2016) then confirmed the suspected IOs by sketching a confidence bound on the ID's plot of GSR versus GLV and declared an observation is IO if it falls beyond its confidence bound. This approach is less efficient as it tends to declare more good observations as IOs and taking longer computational running time. This shortcoming has prompted us to establish an additional step in declaring an observation as IO. Similar to the approach of Mohammed, Habshah and Rahmatullah Imon (2015) and Rashid et al. (2021), the following rule is taken up to confirm the suspected IOs;

i. An observation is identified as RO if; $|\mathrm{RFIID}| \le \mathrm{CP}_{\mathrm{RFGSR}}$ and $|\mathrm{RFGLV}| \le \mathrm{CP}_{\mathrm{RFGLV}}$

ii. An observation is identified as GLO if; $|\mathrm{RFIID}| \le \mathrm{CP}_{\mathrm{RFGSR}}$ and $|\mathrm{RFGLV}| > \mathrm{CP}_{\mathrm{RFGLV}}$

iii. An observation is identified as IO if; $|\mathrm{RFIID}| > \mathrm{CP}_{\mathrm{RFGSR}}$ and $|\mathrm{RFGLV}| \le \mathrm{CP}_{\mathrm{RFGLV}}$

iv. An observation is identified as IO if; ; $|\mathrm{RFIID}| > \mathrm{CP}_{\mathrm{RFGSR}}$ and $|\mathrm{RFGLV}| > \mathrm{CP}_{\mathrm{RFGLV}}$

The detected IOs can be clearly presented in Figure 1.

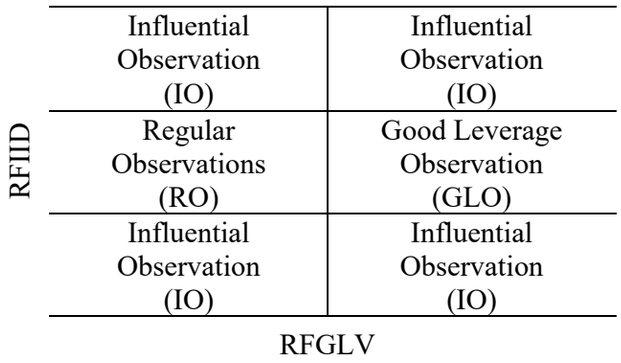

FIGURE 1. plot of (RFIID) against (RFGLV)

MONTE CARLO SIMULATION STUDY

In this section, we consider two different simulation studies.

SIMULATION 1

The first simulation study is performed to assess the performance of our proposed RFIID method and compare its performance with the existing ID and FIID methods. As per Mohammed, Habshah and Rahmatullah Imon (2015), we consider a general linear regression model with p explanatory variables given as:

$$Y_i = B_0 + B_1X_{i1} + B_2X_{i2}, \dots, B_pX_{ip} + e_i, \quad \mathrm{i} = 1,2, \dots, \mathrm{n}$$

where each of the regressors is generated from the uniform distribution (0,10), and e_i is generated from standard normal distribution with varying sample of sizes, n = 50,100,150,and 200. Various proportion of influential points (α = 0.05,0.10,0,15 and 0,20) and p = 3 and p = 5 are being considered. For p = 3 and p = 5, we set B_0 =1, B_1 =2, B_2 = 3, B_3 = 4 and B_0 = 1, B_1 = 2, B_2 = 3, B_3 = 4, B_4 = 5, B_5 = 6 as the true parameter values, respectively. The influential observations are generated such that the independent variables were generated from Uniform Distribution U(20,30) and the response variables were generated following the above linear equation. In each simulation run, there are 10,000 replications. Tables 1-2 exhibit the percentage of correct detection, masking and swamping of unusual observations (IOs) for p = 3 and =5 , respectively. A good method is one that has the highest percentage of correct detection of IOs and least percentages of swamping and masking. It can be seen from the tables that at 5% IOs, all the three methods successfully able to detect all the IOs irrespective of the sample sizes. However, the RFIID has the smallest percentage of swamping, followed by FIID and ID. The results of Tables (1-2) signify that the performance of FIID and ID decreases (percentage of detection of IOs decreases and percentage of swamping increases) as the percentage of IOs increases. On the other hand, the RFIID remains steady, with 100% correct detection of IOs with very least swamping rate compared to the ID and FIID methods. This results show the merit of the proposed RFIID method.

SIMULATION 2

The second simulation study has been conducted to investigate the computer running times of our proposed LMS-RMD$_{\mathrm{RFCH}}$ for the detection of suspected unusual observations and to compare its running times with the LMS-RMD$_{\mathrm{ISE}}$ and GUM method. The same simulation design as in the first simulation is carried out. Due to space constraint, we only present the results for p = 3 with various number of sample sizes, n = (20,40,60,80,100,150,200). The simulation was repeated 10,000 times.

The average of suspected unusual observations (SUO) detected and the running times for the three methods (LMS-RMD$_{\mathrm{RFCH}}$, GUM and LMS-RMD$_{\mathrm{ISE}}$) are presented in Table 3. It can be seen from Table 3 that the proposed LMS-RMD$_{\mathrm{RFCH}}$ outperformed the other two methods in term of having the least computer running time. The appealing performance of the LMS-RMD$_{\mathrm{RFCH}}$ can quickly be observed in Figure 2. It is also interesting to observe that the number of SUO detected by LMS-RMD$_{\mathrm{RFCH}}$ is the closest to the actual number of IOs, followed by LMS-RMD$_{\mathrm{RFCH}}$ and GUM method.

TABLE 1. Correct detection of influential observations, Masking and Swamping for simulation data, when *p=3*

α	n	Percentage of Correct detection			Percentage of Masking			Percentage of Swamping		
		ID	FIID	RFIID	ID	FIID	RFIID	ID	FIID	RFIID
5%	50	100	100	100	0	0	0	7.38	3.22	1.15
	100	100	100	100	0	0	0	5.07	2.60	1.16
	150	100	100	100	0	0	0	4.04	2.50	1.99
	200	100	100	100	0	0	0	6.03	2.29	1.71
10%	50	100	100	100	0	0	0	6.56	2.53	1.05
	100	94.57	100	100	5.43	0	0	7.01	3.24	2.09
	150	93.89	100	100	6.11	0	0	5.34	2.17	1.39
	200	93.50	100	100	6.50	0	0	9.02	2.85	1.75
15%	50	95.70	99.90	99.90	4.30	0.1	0.01	8.71	3.15	2.05
	100	96.06	97.60	100	3.94	2.4	0	10.68	3.35	1.94
	150	93.45	98.1	100	6.55	1.90	0	12.25	4.11	2.32
	200	91.52	98.19	100	8.48	1.81	0	13.87	4.97	2.95
20%	50	88.36	97.99	99.98	11.64	2.01	0.02	17.30	3.23	2.54
	100	90.36	98.89	100	9.64	1.11	0	19.75	3.05	2.80
	150	87.29	98.8	100	12.71	1.20	0	18.15	4.87	2.95
	200	89.60	98.1	100	10.40	1.90	0	20.66	5.79	3.39

TABLE 2. Correct detection of influential observations, Masking and Swamping for simulation data, when *p=5*

α	n	Percentage of Correct detection			Percentage of Masking			Percentage of Swamping		
		ID	FIID	RFIID	ID	FIID	RFIID	ID	FIID	RFIID
5%	50	100	100	100	0	0	0	8.48	3.21	2.85
	100	100	100	100	0	0	0	6.17	2.30	1.16
	150	100	100	100	0	0	0	7.30	3.20	2.09
	200	100	100	100	0	0	0	9.04	2.39	1.79
10%	50	100	100	100	0	0	0	5.60	2.33	1.80
	100	93.67	100	100	6.33	0	0	8.90	3.14	3.09
	150	92.79	100	100	7.21	0	0	9.43	2.77	1.89
	200	94.60	99.2	100	5.40	0.8	0	10.02	3.95	1.95
15%	50	91.6	98.9	99.70	8.40	1.10	0.30	9.71	2.92	2.95
	100	94.16	96.80	100	5.84	1.2	0	11.78	3.77	1.98
	150	90.35	97.98	100	9.65	2.02	0	12.34	4.91	3.06
	200	91.42	97.9	100	8.58	2.01	0	14.45	4.87	1.82
20%	50	87.16	98.91	99.50	12.84	1.09	0.50	19.20	5.73	2.84
	100	89.35	98.9	100	10.65	2.10	0	18.54	367	2.90
	150	86.19	98.1	100	13.81	1.90	0	20.30	4.97	3.79
	200	84.7	97.4	99.9	15.30	2.60	0.10	22.70	6.99	4.39

TABLE 3. Average of suspected unusual observations (SUO) detection and Running time of GUM, LMS-RMD$_{ISE}$ and LMS-RMD$_{RFCH}$ for IOs contamination

α	n	Actual No. of IOs	GUM		LMS-RMD$_{ISE}$		LMS-RMD$_{RFCH}$	
			Average No. of SUO	Running Times (second)	Average No. of SUO	Running Times (second)	Average No. of SUO	Running Times (second)
5%	20	1	2.410	145	2.332	101	2.114	75
	40	2	3.332	160	3.034	120	2.089	81
	60	3	3.223	193	3.203	140	3.125	100
	80	4	4.493	240	4.372	180	4.287	121
	100	5	5.326	301	5.248	240	5.089	140
	150	7	8.534	410	7.986	302	7.763	201
	200	10	10.210	590	10.205	390	10.143	215
10%	20	2	3.354	144	3.034	103	2.289	74
	40	4	4.543	159	4.362	121	4.258	82
	60	6	6.451	192	6.402	140	6.345	103
	80	8	8.451	245	8.392	180	8.378	120
	100	10	10.452	303	10.39	242	10.201	141
	150	15	16.012	415	15.987	301	15.89	202
	200	20	20.952	589	20.893	391	20.37	216
15%	20	3	3.491	146	3.382	101	3.298	75
	40	6	6.489	160	6.390	120	6.378	80
	60	9	9.437	194	9.429	140	9.401	100
	80	12	12.468	243	12.429	181	12.287	122
	100	15	15.368	300	15.208	241	15.182	140
	150	22	23.217	414	23.683	302	22.981	201
	200	30	31.904	591	30.987	390	30.410	214
20%	20	4	4.543	148	4.3982	100	4.013	74
	40	8	8.651	162	8.581	121	8.112	81
	60	12	12.968	198	12.629	142	12.212	103
	80	16	16.892	249	16.484	182	16.192	122
	100	20	20.765	306	20.791	241	20.105	143
	150	30	30.989	419	30.783	303	30.129	204
	200	40	41.671	595	41.583	393	40.551	215

REAL EXAMPLE

Real data set is used to further assess the performance of our proposed RFIID as compared to ID and FIID methods. The assessment of the method is based on the percentage change of the OLS estimates before and after deleting the IOs from the original data based on the ID, FIID and RFID. A good method is one that has the largest percentage change of an estimate. Another criterion used to evaluate the performance of a method is based on the SE of an estimates such that it has the least value of SE after the deletion of the IOs. The percentage of change of the OLS estimate is defined as follows:

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

where $\hat{\alpha}_{Original}$ and $\hat{\alpha}_{Proposed}$ are the OLS estimates for the original data with IOs and the remaining data after IOs have been deleted, respectively. |.| refers to the absolute value.

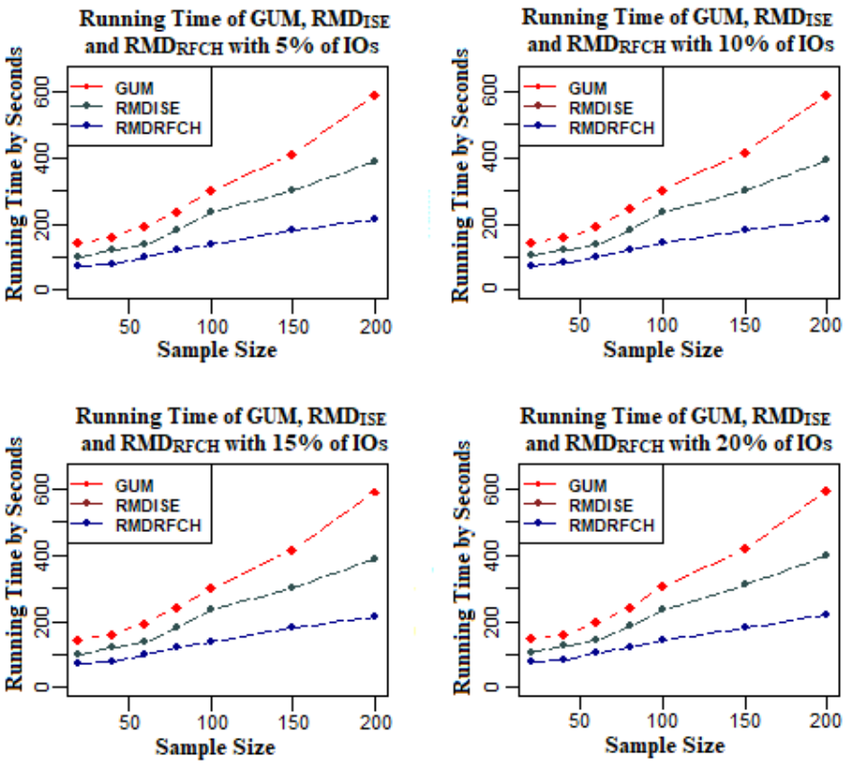

FIGURE 2. Running time GUM, LMS-RMD_{ISE} and LMS-RMD_{RFCH} for each sample size and contamination level for the case of IOs contamination

GUNST AND MASON DATA SET

The Gunst and Mason data set taken from Gunst and Mason (1980) is employed to show the merit of our RFIID method. This data set contains 49 observations, i.e., name of countries (Selected Demographic Characteristics of Countries of the World) and six independent variables (INFD, PHYS, DENS, AGDS, LIT, HIED) with response variable (GNP). The measures of diagnostic and their cut-off points (in parenthesis) are presented in Table 4. It can be observed from Table 4 that GUM and LMS-RMD_{ISE} spotted 18 and 17 observations as suspected unusual observation in the first stage of ID and FIID, respectively. On the other hand, the LMS-RMD_{RFCH} diagnosed 18 suspected IOs. From Figure 3(a) (Index vs Influence distance ID plot), it shows that ID confirmed only 13 cases as IOs. The FIID detects observations (4,7,25,33,42,46) as IOs and (cases 27,17,5,39) as good leverage points, as shown in Figure 3(b). It is interesting to note that the proposed RFIID technique in Figure 3(c) diagnosed observations (1,4,7,25,31,41,42,45,46) as IOs while observations (3,7,20,37) as good leverage values. The performance of the methods is further investigated based on the PCE and the SE of the estimates after deleting the IOs from the original data. The PCE and the SE of the OLS estimates are displayed in Table 5.

The results of Table 5 indicate that the RFIID is very successful in the detection of IOs because the PCEs-RFIID- based values for almost all parameter estimates are the highest, followed by the PCEs-FIID- based values and PCEs-ID- based values. Moreover, the SEs for the estimates based on RFIID method are the smallest. Table 5 results show the merit of RFIID where it correctly identifies IOs and makes the SEs of the estimates become smaller when IOs are removed from the original data set. Furthermore, the computational running time for the LMS-RMD_{RFCH} (RFIID) is the shortest (1.32 s), followed by the LMS-RMD_{ISE} (FIID) (2.14 s) and GUM (ID) (3.30 s). These results are in agreement with the results of the simulation study where the RFIID has the least computational running time.

TABLE 4. The measure of diagnostic for IOs (ID, FIID and RFIID) for Gunst and Mason data set

Ind	Identification of suspected unusual observations													
	Group Union Method (GUM)													
	Std.Stud res. |2.50|	Std.LMS res. |2.50|	hii (0.428)	CD (0.921)	DFF ITS |0.755|	RMD_{ISE} (11.883)	RMD_{RFCH} (12.598)	$ri^* \approx fri^*$ (5.12, -4.44)	$hii^* \approx fhii^*$ (0.590)	*ID* (2.856)	*FIID* (2.856)	$Rfri^*$ (4.781, -4.452)	$Rfhii^*$ (0.859)	*RFIID* (3.429)
1	0.369	**4.899**	0.064	0.001	0.097	2.319	2.411	4.428	0.269	2.304	2.167	4.940	0.287	**3.889**
2	-0.672	0.224	0.038	0.003	-0.133	1.271	1.746	0.213	0.079	0.888	0.898	0.215	0.075	0.804
3	-0.497	-2.830	0.181	0.008	-0.234	6.025	6.513	-0.576	0.334	**5.477**	1.160	-1.662	1.108	4.117
4	0.948	**3.997**	0.043	0.006	0.200	2.636	4.956	5.348	0.148	2.511	2.621	4.903	0.247	**3.838**
5	-0.270	-0.565	0.401	0.007	-0.221	9.968	**15.462**	-0.313	0.773	**5.093**	4.194	-0.140	0.434	0.868
6	-1.080	-1.347	0.024	0.004	-0.169	0.866	0.722	-1.485	0.046	1.583	1.595	-1.405	0.046	1.609
7	**2.916**	**10.33**	0.042	0.045	0.609	1.446	1.416	10.481	0.146	**5.329**	5.409	10.541	0.139	**8.330**
8	0.202	1.022	0.233	0.002	0.111	9.442	8.497	1.414	0.427	1.557	1.725	0.119	0.285	0.171
9	-0.464	-0.067	0.163	0.006	-0.205	6.477	8.740	0.012	0.198	0.272	0.310	-0.369	0.364	0.633
10	0.195	0.458	0.135	0.001	0.077	5.395	3.478	0.581	0.067	0.891	0.961	0.920	0.066	1.016
11	-0.389	0.336	0.040	0.001	-0.080	0.820	1.128	0.570	0.083	0.873	0.845	0.695	0.070	0.914
12	0.268	**3.047**	0.038	0.000	0.053	1.265	1.352	3.045	0.102	1.431	1.509	3.324	0.088	2.612
13	0.285	0.365	0.121	0.002	0.106	5.251	2.691	0.418	0.111	0.660	0.644	0.360	0.064	0.861
14	-0.477	0.902	0.066	0.002	-0.126	2.503	3.148	1.010	0.189	0.320	0.243	0.986	0.210	0.703
15	-0.184	2.232	0.039	0.000	-0.037	1.344	1.657	1.998	0.115	0.772	0.977	2.243	0.108	1.770
16	0.383	0.791	0.164	0.004	0.170	6.147	3.317	0.744	0.201	0.081	0.091	0.164	0.108	0.651
17	0.360	**-16.04**	**0.497**	0.019	0.358	**58.58**	**183.49**	1.501	0.986	**5.321**	5.751	-0.874	0.996	3.489
18	-0.498	-0.565	0.057	0.002	-0.122	2.031	2.160	-0.277	0.087	0.862	0.941	-0.133	0.086	0.799
19	-0.935	-0.565	0.046	0.006	-0.205	1.472	2.018	-0.964	0.142	0.894	0.943	-0.417	0.121	0.763
20	2.085	**3.211**	**0.559**	0.730	**2.348**	**14.27**	**12.60**	2.129	0.564	2.636	2.799	1.870	1.311	5.253
21	-0.760	-0.603	0.077	0.007	-0.219	2.439	3.279	-0.769	0.166	0.694	0.776	-0.386	0.165	0.585
22	-0.425	-0.159	0.027	0.001	-0.071	1.153	1.344	0.164	0.046	1.133	1.137	0.158	0.047	0.932

23	-0.108	-0.436	0.076	0.000	-0.031	3.042	6.186	-0.520	0.232	0.518	0.643	0.317	0.464	1.030
24	-2.177	**-3.769**	0.049	0.032	-0.495	3.969	**12.890**	-2.916	0.499	**3.227**	2.882	-2.782	0.488	2.485
25	**2.595**	**6.354**	0.084	0.078	**0.786**	2.561	2.733	8.106	0.231	**4.706**	**4.115**	7.929	0.240	**6.263**
26	-0.213	0.603	0.163	0.001	-0.094	6.521	6.847	0.580	0.230	0.639	0.230	-0.370	0.376	0.679
27	0.709	-0.397	**0.662**	0.143	**0.993**	**17.16**	**20.566**	0.623	0.770	**3.925**	4.154	1.254	0.796	2.777
28	0.272	-0.565	0.098	0.001	0.089	5.337	3.529	0.652	0.291	0.571	0.675	0.260	0.276	0.177
29	-0.142	0.276	0.087	0.000	-0.044	4.290	2.275	-0.180	0.119	0.721	0.715	-0.225	0.084	0.837
30	-1.064	0.338	0.105	0.019	-0.364	4.765	5.649	0.693	0.502	**6.344**	2.207	0.778	0.493	1.296
31	0.417	**4.911**	0.060	0.002	0.106	2.261	2.258	4.470	0.245	2.215	2.153	4.909	0.266	**3.851**
32	0.012	0.073	0.318	0.000	0.009	8.756	12.079	0.084	0.349	0.982	1.123	-0.115	0.709	2.132
33	1.010	**3.431**	0.087	0.014	0.311	3.194	7.718	5.614	0.593	**5.183**	**3.995**	2.763	0.401	2.297
34	-0.656	-0.565	0.064	0.004	-0.172	3.358	2.365	-0.926	0.081	1.191	1.192	-0.959	0.085	1.201
35	-0.413	0.233	0.082	0.002	-0.123	4.554	6.243	-0.004	0.118	0.695	0.669	0.359	0.264	0.211
36	-0.321	-0.237	0.071	0.001	-0.089	4.220	6.104	-0.387	0.107	0.851	0.857	-0.399	0.120	0.756
37	-1.960	-0.565	0.181	0.114	**-0.923**	6.138	8.256	-0.953	0.333	2.280	1.275	-2.013	1.066	4.019
38	-0.415	-0.565	0.069	0.002	-0.113	3.880	3.410	-0.297	0.076	0.959	1.021	-0.533	0.077	0.994
39	-0.093	**-17.19**	**0.617**	0.002	-0.118	**67.40**	**218.16**	1.452	0.989	**5.346**	5.775	-0.908	0.997	3.499
40	-0.952	-1.819	0.036	0.005	-0.183	1.356	1.791	-1.643	0.067	1.418	1.556	-1.578	0.074	1.645
41	1.443	**5.527**	0.059	0.018	0.363	1.981	2.875	6.368	0.206	**3.300**	**3.158**	6.219	0.226	**4.885**
42	1.692	**5.974**	0.043	0.018	0.360	1.471	1.458	6.696	0.106	**3.436**	**3.395**	6.613	0.086	**5.196**
43	-1.092	**-3.481**	0.131	0.026	-0.424	8.574	**19.353**	-0.797	0.415	1.556	1.732	-1.799	0.708	2.514
44	0.053	-0.565	0.075	0.000	0.015	3.383	6.462	0.752	0.201	0.416	0.096	-0.075	0.576	1.513
45	0.920	**4.094**	0.044	0.006	0.198	2.460	2.481	4.736	0.128	2.269	2.318	4.916	0.113	**3.843**
46	**2.739**	**13.941**	**0.491**	0.895	**2.689**	8.306	9.644	7.217	0.597	**4.285**	**4.667**	7.081	0.617	**5.987**
47	-0.857	-0.066	0.038	0.004	-0.170	1.428	1.306	-0.536	0.113	0.891	0.877	-0.227	0.113	0.710
48	0.115	1.976	0.039	0.000	0.023	1.682	2.888	2.633	0.090	1.271	1.363	2.414	0.104	1.903
49	-0.867	-0.503	0.117	0.014	-0.316	6.168	6.549	-0.804	0.180	0.938	0.770	-1.283	0.156	1.240

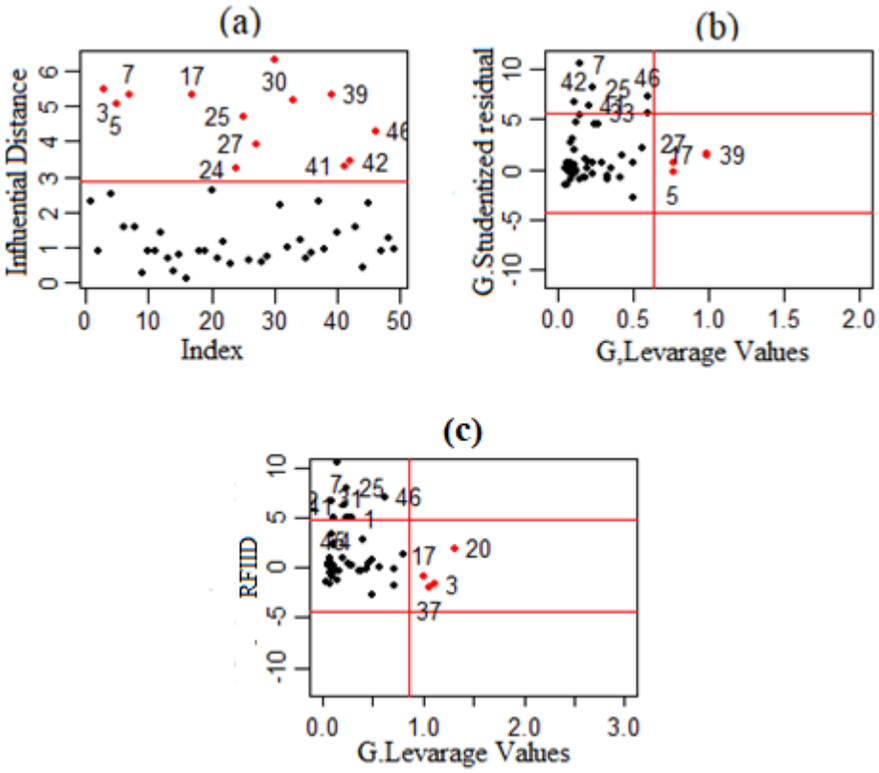

FIGURE 3. a) ID, b) FIID and c) RFIID plots for Gunst and Mason data set

TABLE 5. The values of PCE for RFIID, ID and FIID based on OLS for Gunst and Mason data set

Variables	Original Data	Removed IOs by ID [13 observations]		Removed IOs by FIID [(6 Observation)]		Removed IOs by RFIID [(9 Observations)]	
	Estimation (S.E)	Estimation (S.E)	PCE	Estimation (S.E)	PCE	Estimation (S.E)	PCE
Constant	116.455 (344.051)	83.071 (244.985)	28.66 (28.80)	101.478 (217.577)	12.86 (36.76)	**133.0 (154.30)**	**14.20 (55.15)**
INFD	-3.367* (1.973)	-2.274. (1.329)	32.46 (32.64)	-2.183. (1.263)	35.16 (35.98)	**-1.663** (0.189)**	**50.60 (90.42)**
PHYS	-0.004 (0.068)	-0.003 (0.056)	25 (17.64)	-0.011 (0.043)	175 (36.76)	**-0.014 (0.030)**	**250 (55.88)**
DENS	-0.197 (0.373)	0.728 (0.687)	469.54 (84.18)	0.033 (0.243)	116.75 (39)	**-0.009 (0.173)**	**104.56 (53.61)**
AGNS	0.003 (0.010)	-0.033 (0.033)	1200 (230)	-0.001 (0.007)	133.33 (30)	**-0.0003 (0.005)**	**110 (50)**
LIT	5.638 (3.354)	4.595. (2.511)	18.49 (25.13)	4.593* (2.211)	18.53 (34)	**4.040** (1.576)**	**28.34 (53)**
HIED	0.683** (0.167)	0.557* (0.176)	18.44 (5.38)	0.484** (0.155)	29.13 7.18	**0.030*** (0.149)**	**95.60 (10.77)**
Residuals Std.Errors		220		224.4			**55.38**
F-Statistics	9.803**	9.474***	3.47	9.172**		**8.783*****	**10.40**

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

Conclusions

In this paper, we propose a Robust and Fast Improvised Influential Distance RFIID method for detecting influential observations in multiple linear regression. The existing ID method is capable of detecting IOs at 5% of contamination but it suffers from swamping effects. As the percentage of IOs increases to more than 5%, the performance of ID prone to decrease with higher masking and swamping effects. The performance of FIID is also not very encouraging as it inclines to suffer from swamping effect when the percentage of contamination is at least 10%. Moreover, it is not computationally stable. The performance of our proposed RFIID has been comprehensively examined by employing Monte Carlo simulation study and real data set. The results signify that the RFIID is very successful in the detection of IOs with imperceptible swamping effects, irrespective of the number of independent variables, sample sizes, and percentage of contaminations. Another attractive feature of RFIID is that it is computationally stable and its computational running time is much quicker than the ID and FIID.

References

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. 1986. Influential observations, high leverage points, and outliers in linear regression. *Statistical Science* 1(3): 379-393.

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

Gunst, R.F. & Mason, R.L. 1980. *Regression Analysis and Its Application: A Data Oriented Approach*. New York: Marcel Dekker.

Habshah, M. & Shabbak, A. 2011. Robust multivariate control charts to detect small shifts in mean. *Mathematical Problems in Engineering* 2011: 923463. doi: 10.1155/2011/923463

Habshah, M., Muhammad, S. & Ismaeel, S.S. 2021. Fast improvised influential distance for the identification of influential observations in multiple linear regression. *Sains Malaysiana* 50(7): 2085-2094.

Habshah, M., Talib, H., Jayanthi, A. & Uraibi, H.S. 2020. Fast and robust diagnostic technique for the detection of high leverage points. *Journal of Science and Technology* 28(4): 1203-1220.

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.

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.

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: 279472. doi.org/10.1155/2015/279472

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

Olive, D.J. & Hawkins, D.M. 2010. *Robust Multivariate Location and Dispersion*. Preprint, www. Math. Siu. Edu/olive/preprints. Htm

Olive, D.J. & Hawkins, D.M. 2008. *High Breakdown Multivariate Estimators*. https://www.researchgate.net/profile/David_Olive2/publication/240737720_High_Breakdown_Multivariate_Estimators/links/0a85e53234b7db7f90000000.pdf

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

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

Rashid, A.M., Midi, H., Dhnn, W. & Arasan, J. 2021. An efficient estimation and classification methods for high dimensional data using robust iteratively reweighted SIMPLS algorithm based on Nu-Support vector regression. *IEEE Access* 9: 45955-45967.

Rashid, A.M., Midi, H., Dhnn, W. & Arasan, J. 2022. Detection of outliers in high-dimensional data using Nu-Support vector regression. *Journal of Applied Statistics* 49(10): 2550-2569.

Rousseeuw, P. & Leroy, A.M. 1987. *Robust Regression and Outlier Detection*. New York: Wiley Series in Probability and Mathematical Statistics.

Rousseeuw, P. & Yohai, V. 1984. Robust regression by means of S-estimators. In *Robust and Nonlinear Time Series Analysis*. New York: Springer.

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

Zahariah, S. & Midi, H. 2023. Minimum regularized covariance determinant and principal component analysis-based method for the identification of high leverage points in high dimensional sparse data. *Journal of Applied Statistics* 50(13): 2817-2835.

*Corresponding author; email: habshah@upm.edu.my