Pertanika J. Sci. & Technol. 28 (4): 1203 - 1220 (2020)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Fast and Robust Diagnostic Technique for the Detection of High Leverage Points

Habshah Midi[1,2]*, Hasan Talib Hendi[1], Jayanthi Arasan[2] and Hassan Uraibi[3]

[1]Institute for Mathematical Research, Universiti Putra Malaysia, 43400 UPM, Serdang, Selangor, Malaysia
[2]Department of Mathematics, Faculty of Science, Universiti Putra Malaysia, 43400 UPM, Serdang, Selangor, Malaysia
[3]Department of Statistics, University of Al-Qadisiyah, 88 -Al-Qadisiyah -Al-Diwaniyah, Iraq

ABSTRACT

High Leverage Points (HLPs) are outlying observations in the X -directions. It is very imperative to detect HLPs because the computed values of various estimates are affected by their presence. It is now evident that Diagnostic Robust Generalized Potential which is based on the Minimum Volume Ellipsoid (DRGP(MVE)) is capable of detecting multiple HLPs. However, it takes very long computational running times. Another diagnostic measure which is based on Index Set Equality denoted as DRGP(ISE) is put forward with the main aim of reducing its running time. Nonetheless, it is computationally not stable and still suffers from masking and swamping effects. Hence, in this paper, we propose another version of diagnostic measure which is based on $\sqrt{n}$ Reweighted Fast Consistent and High Breakdown (RFCH) estimators. We call this measure Diagnostic Robust Generalized Potential based on RFCH and it is denoted by DRGP(RFCH). The results of simulation study and real data indicate that our proposed method outperformed the other two methods in term of having the least computing time, highest percentage of correct detection of HLPs and smallest percentage of swamping and masking effects compared to the DRGP(MVE) and DRGP (ISE).

Keywords: Diagnostic robust generalized potentials, high leverage points, mahalanobis distance, outliers

ARTICLE INFO

Article history:
Received: 19 April 2020
Accepted: 27 July 2020
Published: 21 October 2020

DOI: https://doi.org/10.47836/pjst.28.4.05

E-mail addresses:
habshah@upm.edu.my (Habshah Midi)
h.applied.t88@gmail.com (Hasan Talib Hendi)
jayanthi@upm.edu.my (Jayanthi Arasan)
hssn.sami1@gmail.com (Hassan Uraibi)
*Corresponding author

ISSN: 0128-7680
e-ISSN: 2231-8526

INTRODUCTION

The Ordinary Least Squares (OLS) is the most popular technique in regression analysis because of tradition and ease of computation. Moreover, the OLS is easy to use as it is available in most of the statistical software like SPSS, SAS and MINITAB. The OLS technique has many attractive features under normality assumption of regression errors, not only in parameters estimation but also in testing of hypothesis. However, many are not aware that the one immediate consequence of the presence of outliers especially outlying observations in the X- direction which is call High Leverage Points (HLPs) may cause apparent non-normality (Huber, 1973). Since most of the statistical analysis are based on normality assumption, the violation of this assumption may lead to invalid inferential statements and inaccurate predictions. Evidences are now available in the literatures that the presence of HLPs have an adverse effect on the computed values of various estimates (Rousseeuw, 1985; Imon & Khan, 2003; Midi et al., 2009; Riazoshams et al., 2010; Bagheri et al., 2012). As such it is very crucial to search for a very effective method of detecting HLPs. The HLPs can easily be spotted from a plot of response variable against the predictor variable for simple linear regression model. Nonetheless it is hard to identify multiple HLPs for more than one independent variable due to swamping and masking effect (Peña & Yohai, 1995).

There are many papers that deal with the diagnostic tools for the identification of HLPs (Rousseeuw, 1985; Rousseeuw & Driessen, 1999; Midi et al., 2009). Midi et al. (2009) had shown that their method was very successful for the detection of HLPs compared to hat matrix approach of (Hoaglin & Welsh, 1978) and Hadis' potential (Hadi,1992). Even though some of them are able to correctly identify multiple HLPs, their running times are very long due to using Minimum Volume Ellipsoid (MVE) or Minimum Covariance Determinant (MCD) for obtaining the final estimator of location and scatter. Lim and Midi (2016) exemplified that Index Set Equality (ISE) had tremendously sped up the computation of location and scatter estimator, even much faster than fast MCD (Rousseeuw & Driessen, 1999). The only shortcoming of this method is that it is not very stable because its computation depends on the selected initial subset, h. According to Salleh (2013), the final estimator of location and scatter of (ISE) is equivalent to MCD if the same initial subset is utilized, otherwise the results will be quite different. In order to obtain more efficient and much faster location and scatter estimators, we propose employing Reweighted Fast Consistent and High Breakdown (RFCH) estimators (Olive & Hawkins, 2010; Alkenani & Yu, 2013). By employing the results of Lopuhaa (1999), Olive and Hawkin (2010) proved that the location and scatter estimators of RFCH were $\sqrt{n}$ consistent estimators. For this reason, Uraibi et al. (2017) developed a robust forward selection method by formulating a correlation matrix based on RFCH estimators which produced very appealing results. Hence in this paper, we propose using the $\sqrt{n}$ RFCH consistent estimators of location and scatter in the establishment of Robust Mahalanobis Distance (RMD).

The paper is organized as follows. Section 2 discusses the importance of detection of high leverage points. Section 3 reviews a few methods of detection of high leverage points. The proposed Diagnostic Robust Generalized Potential based on RFCH estimators is presented in Section 4. Section 5 discusses the results of the simulation and numerical example. The concluding remarks are given in Section 6.

MATERIALS AND METHODS

Real Data to Show Why it is Very Important to Detect HLPs

As already mentioned in the preceding section, various estimates can be affected by HLPs. That is why it is very important to first check their existence before making any inferences to avoid misleading conclusion. In this section, we want to show that HLPs can cause multicollinearity and heteroscedasticity by using real examples. Let us first focus on Hawkins Bradu Kass Data (Hawkins et al., 1984). This artificial data set consists of 75 observations and 3 independent variables. Hawkins et al. (1984) claimed that this data set had 14 HLPs. Bagheri et al. (2012) exemplified that the first 14 observations (included in the original data) as displayed in Table 1, caused multicollinearity evident by showing maximum value of Variance Inflation Factor (VIF=33.342 for $X3$) greater than 10. On the contrary, no multicollinearity was observed in their absence.

Table 1
Multicollinearity Diagnostics (VIF) for Hawkin Bradu Kass Data

Status	$X1$	$X2$	$X3$
Original Data	13.432	23.853	33.432
Without observations 1-14	1.012	1.017	1.027

Education Expenditure Data taken from Chatterjee and Hadi (2006) will be our second example to illustrate that HLPs can caused heteroscedasticity. Many authors frequently used this data (n = 50 observations with three independent variables) to deal with heteroscedasticity (Chatterjee & Hadi, 2006; Imon, 2002; Midi et al., 2014). This data consists of 50 observations where per capita income on education project for 1975 is the dependent variable and three explanatory variables namely per capita income in 1973, number of residents per thousand under 18 years of age, and number of residents per thousand under 18 years of age in 1974. According to Midi et al. (2014), observation 49, i.e. Alaska (AK) is HLP and influences the heteroscedasticity pattern of the data. We investigated this data and apply the White test (WT) which is a Lagrange Multiplier (LM) test statistic proposed by White (1980) to test the presence of heteroscedasticity in linear regression model. The white test is defined as $LM = nR^2$, where R^2 is the coefficient of multiple determination and n is size of sample. The LM test statistics is distributed

as Chi-Squared χ_p^2 , where p is the number of predictors. Table 2 exhibits the values of *LM* test statistics with their corresponding *p*-values. We can see from Table 2 that in the absence of HLP the WT shows no heteroscedasticity but in their presence the WT shows heteroscedasticity.

Table 2
Heteroscedasticity Diagnostics (White test)

Status	$LM = nR^2$	*p*-values
Without AK (HLP)	5.7978	0.1219
With AK (HLP)	22.7817	4.48e-05

We have seen the effect that the HLPs had on the heteroscedasticity and multicollinearity pattern of a data and it is crucial to detect them before any further analysis to be carried out. This is the reason why we need to find the more reliable method for detecting their existence.

Review of Some Methods of Identifications of HLPs

In this section, some methods of identification of HLPs are reviewed.

Diagnostics Robust Generalized Potential (DRGP)

Mahalanobis (1936) defined Mahalanobis Distance (MD) as a measure of deviation of a data point from its center. Let us write the i^{th} vector of predictor 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 mean vector and the variance covariance matrix are calculated as:

$$\bar{t} = {}^1\!/_n \sum_{i=1}^n t_i \text{ and } C = \left(\frac{1}{n-1}\right) \sum_{i=1}^n (t_i - \bar{t}\,)\,(t_i - \bar{t}\,)', \text{ respectively.}$$

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

$$MD_i = \sqrt{\left(t_i - T(X)\right)' C(X)^{-1} \left(t_i - T(X)\right)} \qquad i = 1,2, \dots, n, \qquad [1]$$

where *T(X)* is the mean vector $(\bar{t}\,)$ and *C(X)* is the variance covariance matrix *(C)*. Rousseeuw and Leroy (1987) suggested using Robust Mahalanobis Distance (RMD) as a diagnostic tool for detection of HLPs by replacing the classical mean vector *T(X)*, and classical covariance matrix, *C(X)* of MD_i in Equation 1 by robust estimators such as Minimum Volume Ellipsoid (MVE) or Minimum Covariance Determinant (MCD)

(Rousseeuw & Yohai, 1984), because the former estimators are not robust. They considered observation as HLPs if its corresponding (RMD value) exceeds the cutoff points $\sqrt{x^2_{p,\alpha}}$. Midi et al. (2009) noted that the RMD was not very successful in the identification of HLPs and established Diagnostics Robust Generalized Potential (DRGP) whereby its algorithm consisted of two steps. The suspected HLPs were detected using RMD based on MVE and on the second steps, generalized potentials, denoted as (p_{ii}), were employed to confirm the suspected HLPs. Since the distribution of generalized potentials was intractable, they suggested a confidence bound of cutoff points as follows:

$$cutoff.p_{ii} = median\,(p_{ii}) + 3 * MAD(p_{ii}),$$

where

$$MAD(p_{ii}) = median\{|\,p_{ii} - median(p_{ii})|\}/0.6745.$$

Even though the DRGP is very successful in identifying HLPs, its running time is very slow due to using MVE in the first step. Lim and Midi (2016) improvised the DRGP to speed up the computation of location and scatter estimator by using Index Set Equality. They showed that the DRGP based on ISE was much faster than the DRGP based on MVE.

Index Set Equality (ISE)

Salleh (2013) established Index Set Equality (ISE) where it is an innovation from fast MCD. The following steps illustrate the computation of ISE.

Step 1. Choose arbitrarily *h* observations from a dataset to be included in the subsample denoted as H_{old}, where $h = \frac{n+p+1}{2}$ and *p* is the number of independent variables (Rousseeuw & Driessen, 1999).

Let $I_{old} = \{\pi^{old}_{(1)}, \pi^{old}_{(2)}, \ldots, \pi^{old}_{(h)}\}$ be the index set for H_{old}.

Step 2. Compute the *p*-dimensional mean vector $\bar{T}_{H_{old}}$ and the (*p x p*) covariance matrix of C_{old} from the subset H_{old}.

Step 3. Compute the squared Mahalanobis Distance for each observation, as

$$d^2_{old}(i) = (t_i - \bar{T}_{Hold})'\, C^{-1}_{Hold}(t_i - \bar{T}_{Hold}) \quad \text{for } i = 1, 2, \ldots, n.$$

Step 4. Arrange $d^2{}_{old}(i)$ in increasing order,

$$d^2_{old}(\pi(1)) \le d^2_{old}(\pi(2)) \le \cdots \le d^2_{old}(\pi(n)),$$

where π is permutation equal to {1, 2, ..., n}.

Step 5. The first *h* items that correspond to the smallest $d^2_{old}(i)$ will be placed in set $H_{New} = \{t_{\pi(1)}, t_{\pi(2)}, \ldots, t_{\pi(h)}\}$. Then list the new Index Set, as

$$I_{New} = \{\pi^{New}_{(1)}, \pi^{New}_{(2)}, \ldots, \pi^{New}_{(h)}\}.$$

Step 6. If $I_{New} = I_{old}$. Stop the process, then the location vector $H_{old} = H_{New}$ and covariance matrix $C_{H_{old}} = C_{H_{New}}$, if $I_{New} \neq I_{old}$ let $H_{old} = H_{New}$, then recompute $\bar{T}_{H_{new}}$ and $C_{H_{New}}$, repeat steps 3 to 6, until $I_{New} = I_{old}$. Then the corresponding $\bar{T}_{H_{new}}$ and are the location and scatter estimates for ISE.

Reweighted Fast Consistent and High Breakdown (RFCH)

Olive and Hawkins (2010) developed Reweighted Fast Consistent and High breakdown (RFCH) estimators of location and scatter which was faster than the fast MCD developed by Rousseeuw and Driessen (1999). The attractive feature of RFCH technique is that not only its computation is very fast which is even faster than Fast MCD (Zhang et al., 2012), but it is $\sqrt{n}$ consistent estimators. The RFCH utilizes the $\sqrt{n}$ consistent DGK (Devlin et al., 1981) estimator and high breakdown Median Ball (MB) (Olive & Hawkins, 2008) estimators as attractors. The RFCH algorithms can be summarized as follows:

The DGK Algorithm Steps

Step 1. Compute the *p*-dimensional row vector of location and (*pxp*) the $(T(X), C(X))$ covariance matrix, () of the original data and use it as the initial or starting point $(T_{0,Start}, C_{0,Start})$, for calculating the initial Mahalanobis Distance (Equation 2).

$$MD_{i0,DGK} = \sqrt{(t_i - T_{0,Start})'(C_{0,Start})^{-1}(t_i - T_{0,Start})}, \quad i = 1, 2, ..., n. \qquad [2]$$

Step 2. Sort the $MD_{i0,DGK}$ in increasing order. Then calculate its median, $MED = median(MD_{i0,DGK})$. The observation corresponding to the Mahalanobis Distance less than the median will be in the remaining half dataset (*m* observations), defined as Equation 3

$$\tilde{X}_{1,DGK} = \{X_{jl}: MD_{i0,DGK} \leq MED\},\ j = 1,2, \dots, k,\ l = 1,2, \dots, m, \qquad [3]$$

where *k* is the number of predictor variables.

Step 3. Consider $C_{0,DGK} = C_{0,Start}$ where $C_{0,Start}$ is the original dataset's scatter matrix, then recompute the location and scatter estimators for the $\tilde{X}_{1,DGK}$ dataset to obtain the first attractors $(T_{1,DGK}, C_{1,DGK})$.

Step 4. Stop the process if the diagonal elements of $C_{1,DGK} = C_{0,Start}$,otherwise repeat Steps 1 to 3 until convergence where at convergence the final location and scatter estimates $(T_{K,DGK}, C_{K,DGK})$ is acquired from the $\tilde{X}_{K,DGK}$, where *K* is final step at which convergence takes place.

The MB Algorithm Steps

Step 1. Let an identity matrix be the scatter matrix, denoted as $C = I_p$. Then compute Mahalanobis Distance based on the median vector, median (X) and C as Equation 4:

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

where $Med(X) = \text{median } (X)$.

Let the median of MD_i be the cut-off point, which is denoted by $Lcut$ (Equation 5),

$Lcut$=median (MD_i), [5]

where $Lcut \neq 0.5$. Determine the $\tilde{X}_0$ for half of the dataset *(m)* whose MD_i is less than or equal to the *Lcut*, such that (Equation 6)

$$\tilde{X}_0 = \{X_{jl} : MD_i \leq Lcut\}, j = 1,2, \dots, k, l = 1,2, \dots, m. \quad [6]$$

Step 2. Compute the *p*-dimensional row vector of location and the *(pxp)* covariance matrix of scatter estimators of the $\tilde{X}_0$ and use it as the initial or starting point $(T_{0,Start}, C_{0,Start})$, for calculating the initial Mahalanobis Distance (Equation 7).

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

determined the remaining half dataset by using new cut-off point as Equation 8:

$$\tilde{X}_{1,MB} = \{X_{jl} : MD_{0i,MB} \leq Lcut0\}, j = 1,2, \dots, k, \ \ l = 1,2, \dots, m, \quad [8]$$

where $Lcut0 = \text{median } (MD_{0i,MB})$.

Step 3. Based on the $\tilde{X}_{1,MB}$, calculate the attractor $(T_{1,,MB}, C_{1,MB})$.

Step 4. If the diagonal elements of $C_{1,MB} = C_{0,Start}$ stop the process, otherwise recalculate the $MD_{1,MB}$ based on attractor $(T_{1,MB}, C_{1,MB})$ and iterate the Steps 2 to 3, until the convergence is achieved at final attractor $(T_{K,MB}, C_{K,MB})$ and final remaining set $\tilde{X}_{K,MB}$.

The RFCH Algorithm Steps

The RFCH consists of three steps where in the first step the Fast Consistent and High breakdown (FCH) attractors of Olive and Hawkins (2010) is determined based on the final attractors of DGK and MB estimators that adhere the following rules:

Step 1. The T_{FCH} and C_{FCH} are determined as Equation 9:

$$T_{FCH} = \begin{Bmatrix} T_{K,DGK} \ \ if \ \ \sqrt{|C_{K,DGK}|} < \sqrt{|C_{K,MB}|} \\ T_{K,MB} \qquad\qquad Otherwise \end{Bmatrix}, \quad [9]$$

And Equation 10

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

where $x^2_{(p,0.5)}$ is chi-square distribution with p degrees of freedom and significance level 0.5. The (T_{FCH}, C^*_{FCH}) are the consistent estimators of the FCH attractors according to Theorem 1 of Olive and Hawkins (2010),

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

Step 2. Construct a new set of data, $\tilde{X}_{FCH}$ by using the following Equation 11,

$$\tilde{X}_{FCH} = \{X_{jl}: MD_i(T_{FCH}, C^*_{FCH}) \leq x^2{}_{(p,1-\alpha)}\}, \quad [11]$$

$j = 1, 2, \ldots, k, l = 1, 2, \ldots, m,$

where $MD_i(T_{FCH}, C^*_{FCH})$ is the Mahalanobis Distance based on the location and scatter of FCH estimators in Step 1. Then compute the location and scatter estimators for the $\tilde{X}_{FCH}$ dataset to obtain the RFCH attractors, $(T_{1,RFCH}, C_{1,RFCH})$. Again, following Theorem 1 of Olive and Hawkins (2010),
$C^*_{1,RFCH}$ is defined as Equation 12

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

Subsequently the Mahalanobis Distance based on is computed and a new set of data is constructed using the following Equation 13;

$$\tilde{X}_{2,RFCH} = \{X_{jl}: MD_i(T_{1,RFCH}, C^*_{1,RFCH}) \leq x^2{}_{(p,1-\alpha)}\}, \quad [13]$$

$$j = 1,2, \ldots, k, l = 1,2, \ldots, m.$$

Following the same process, $(T_{2,RFCH}, C_{2,RFCH})$ estimators are calculated based on the $\tilde{X}_{2,RFCH}$ dataset. Afterwards, $C^*_{2,RFCH}$ is defined as in Equation 14 by applying Theorem 1 of Olive and Hawkins (2010),

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

Step 3. Step 1 to 2 is repeated K-times until convergence. Convergence is achieved if the number of detected outliers or HLPs is the same for $MD_i(T_{K,RFCH}, C_{K,RFCH})$ and $MD_i(T_{K-1,RFCH}, C^*_{K-1,RFCH})$.

As stated by Olive and Hawkins (2010), on convergence, the final estimators of RFCH, i.e. $(T_{K,RFCH}, C_{K,RFCH})$ are High Breakdown (HB) $\sqrt{n}$ consistent estimators (see Olive and Hawkins (2010) for description of $\sqrt{n}$ consistent estimator).

The Proposed Diagnostic Robust Generalized Potential based on Reweighted Fast Consistent and High Breakdown Estimators (DRGP(RFCH))

Midi et al. (2009) proposed Diagnostic Robust Generalized Potential based on Minimum Volume Ellipsoid (DRGP(MVE)) for detecting HLPs. The DRGP algorithm comprises two steps where in the first step, Robust Mahalanobis Distance (RMD) based on MVE is used to detect the suspected HLPs and on the second step, the generalized potential is used to confirm whether or not the suspected HLPs is a genuine HLPs. Although the DRGP is proven to be very successful in detecting HLPs, its computation running time is very slow since it uses the location and scatter estimators obtained from the MVE. As such, Lim and Midi (2016) proposed another diagnostic method, Diagnostic Robust Generalized Potential based on Index Set Equality (DRGP(ISE) to identify HLPs by incorporating the location and scatter estimators based on ISE. However, through our investigation, the DRGP (ISE) is not very stable and we anticipate that it still suffers from small percentage of swamping and masking effect. We also expect that the running time of the DRGP (ISE) can be improved. In this regard, we attempt to improvise the existing DRGP by integrating the location and scatter estimators obtained from the Reweighted Fast Consistent and High breakdown (RFCH) estimators (Olive & Hawkins, 2010). The attractive feature of this estimator is that it is High Breakdown $\sqrt{n}$ consistent estimator as noted by Olive and Hawkins (2010). Our improvised DRGP is denoted as DRGP (RFCH).

The Proposed DRGP (RFCH) Technique is Summarized as Follows

Step 1. Identify the suspected HLPs by using RMD_i for each i^{th} observation based on RFCH (Equation 15)

$$RMD_i = \sqrt{(t_i - T_{RFCH})'(C_{RFCH})^{-1}(t_i - T_{RFCH})} \ , i = 1,2, \dots, n. \quad [15]$$

Step 2. As per Midi et al. (2009) the cut-off point is defined as follows;

$$cut\_off = median\ (RMD_i) + 3 * MAD(RMD_i),$$

where

$$M(A(D(RMD_i(= \ (\ (m(edian\ abs\ RMD_i - \ median\ \ RMD_i/0.6745\ .)$$

We declare that any i^{th} case with Robust $RMD_i > cut\_off$ point, is the suspected HLPs and include them in a deletion group, denoted as *D* Group, while the rest of the observations are kept in the *R* group.

Step 3. Following Midi et al. (2009), we employ generalized potential, P_{ii}, to confirm the suspected HLPs whether are not they still can be considered as HLPs (Equation 16).

$$p_{ii} = \begin{cases} z_{ii}^{(-D)} & for\ i\ \in D \\ \dfrac{z_{ii}^{(-D)}}{1 - z_{ii}^{(-D)}} & for\ i\ \in R \end{cases} \quad [16]$$

Where (Equation 17)

$$z_{ii}{}^{(-D)} = x_i^T (X_R^T X_R)^{-1} x_i, \quad i = 1,2, \dots, n\ . \quad [17]$$

Step 4. Compute the cut-off point p_{ii}, for i.e. cut-off $p_{ii} = Median(p_{ii}) + 3Q_n(p_{ii})$. Rousseeuw and Croux (1993) defined $Q_n = C\{|x_i - x_j|; i < j\}_{(k)}$ as a pairwise order statistic of whole distance where $k = \binom{h}{2} \approx \binom{n}{2} \approx \binom{n}{2}/4$ and $h = \left[\frac{n}{2}\right] + 1$. Rousseeuw and Croux (1993) noted that to make Qn a consistent estimator for Gaussian data the value of C should be chosen equals to 2.2219.

We declare that all members of the D group as HLPs if P_{ii} > cut-off P_{ii}, otherwise place those observations back into the estimation subset R sequentially begin with the least p_{ii} value.

RESULTS AND DISCUSSION

Monte Carlo Simulation Study

Monte Carlo simulation study was carried out to assess the performance of our proposed DRGP(RFCH) compared with DRGP(MVE) and DRGP(ISE). As per Lim and Midi (2016), we consider a general linear regression model with p explanatory variables as Equation 18

$$Y_i = B_0 + B_1 X_{i1} + B_2 X_{i2}, \dots, B_p X_{ip} + e_i, \quad \mathrm{i} = 1,2, \dots, \mathrm{n}, \quad [18]$$

where each of the explanatory variable is generated from Uniform Distribution (0, 10), e_i is generated from standard normal distribution with varying sample of sizes, n = 20, 40, 60, 80, 100 and 200. We consider various proportion of High Leverage Points (α = 0.05, 0.10 and 0.15) and p = 2and p = 4. For p = 2, we set B_0= 1, B_1 = 2 and B_2 = 3 as the true parameter values and set B_0 = 1, B_1 = 2, B_2 = 3, B_3 = 4 and B_4 = 5 as the true parameter values for p = 4. The HLPs are created by replacing the first 100 α% observations of the original good data for p = 2 and p = 4 with values of X_1 and X_2 and values of X_1, X_2, X_3, and X_4, respectively, generated from Uniform Distribution U (15, 20), without changing their y values. The simulation was repeated 10,000 times.

Table 3 and 4 exhibit the percentage of correct detection, masking, and swamping of HLPs for p = 2 and p = 4. It can be observed from Table 3 and 4 that for p =2, p = 4, n = 20 and at 5% HLPs, the performance of the three methods are fairly closed, but DRGP(ISE)

Table 3

Percentage of Correct detection, Masking and Swamping, p = 2

% of HLPs		% of Correct detection			% of Masking			% of Swamping		
	n	DRGP (MVE)	DRGP (ISE)	DRGP (RFCH)	DRGP (MVE)	DRGP (ISE)	DRGP (RFCH)	DRGP (MVE)	DRGP (ISE)	DRGP (RFCH)
5	20	100	100	100	0	0	0	7.73	6.875	7.726
	40	100	100	100	0	0	0	1.419	1.527	1.193
	60	100	100	100	0	0	0	0.685	0.824	0.617
	80	100	100	100	0	0	0	0.426	0.539	0.389
	100	100	100	100	0	0	0	0.289	0.379	0.281
	200	100	100	100	0	0	0	0.085	0.104	0.084
10	20	99.98	99.89	99.99	0.02	0.11	0.01	3.318	3.145	2.854
	40	100	99.99	100	0	0.01	0	0.847	1.048	0.731
	60	100	100	100	0	0	0	0.372	0.572	0.369
	80	100	100	100	0	0	0	0.206	0.368	0.199
	100	100	100	100	0	0	0	0.135	0.29	0.132
	200	100	100	100	0	0	0	0.028	0.1	0.024
15	20	99.86	99.55	99.85	0.14	0.45	0.15	2.222	2.242	1.935
	40	100	99.84	100	0	0.16	0	0.505	0.685	0.451
	60	100	99.9	100	0	0.1	0	0.187	0.408	0.176
	80	100	99.98	100	0	0.02	0	0.095	0.256	0.079
	100	100	99.95	100	0	0.05	0	0.048	0.209	0.048
	200	100	100	100	0	0	0	0.006	0.071	0.004

Table 4

Percentage of Correct detection, Masking and Swamping, p = 4

% of HLPs		% of Correct detection			% of Masking			% of Swamping		
	n	DRGP (MVE)	DRGP (ISE)	DRGP (RFCH)	DRGP (MVE)	DRGP (ISE)	DRGP (RFCH)	DRGP (MVE)	DRGP (ISE)	DRGP (RFCH)
5	20	100	100	100	0	0	0	7.397	5.348	6.810
	40	100	100	100	0	0	0	2.195	1.821	1.544
	60	100	100	100	0	0	0	1.118	1.011	0.995
	80	100	100	100	0	0	0	0.720	0.700	0.675
	100	100	100	100	0	0	0	0.531	0.534	0.506
	200	100	100	100	0	0	0	0.262	0.271	0.257
10	20	99.99	99.24	99.99	0.01	0.76	0.01	4.789	3.690	4.582
	40	100	100	100	0	0	0	1.362	1.477	1.188
	60	100	100	100	0	0	0	0.623	0.590	0.59
	80	100	100	100	0	0	0	0.401	0.382	0.374
	100	100	100	100	0	0	0	0.274	0.273	0.283
	200	100	100	100	0	0	0	0.117	0.117	0.115
15	20	99.84	89.16	99.92	0.16	10.84	0.08	3.092	2.553	3.306
	40	100	99.71	100	0	0.29	0	0.788	0.707	0.651
	60	100	99.99	100	0	0.01	0	0.337	0.309	0.295
	80	100	100	100	0	0	0	0.201	0.19	0.185
	100	100	100	100	0	0	0	0.132	0.124	0.120
	200	100	100	100	0	0	0	0.039	0.038	0.025

Table 5

Running time by (Seconds) and Average of HLPs Detection of DRGP(MVE), DRGP(ISE) and DRGP(RFCH).

			DRGP(MVE)		DRGP(ISE)		DRGP(RFCH)	
%HLPs	Sample Size	Actual HLPs	Average of HLPs	Running time	Average of HLPs	Running time	Average of HLPs	Running time
5	20	1	2.456	60.81	2.375	40.61	2.545	13.25
	40	2	2.676	100.6	2.611	48.28	2.477	13.34
	60	3	3.411	129.2	3.494	54.48	3.370	13.35
	80	4	4.341	152.9	4.431	60.45	4.311	13.44
	100	5	5.289	178.6	5.379	65.21	5.281	13.51
	200	10	10.17	310.2	10.252	91.28	10.168	13.78
10	20	2	2.663	60.79	2.629	40.45	2.571	13.28
	40	4	4.339	100	4.419	47.51	4.292	13.35
	60	6	6.223	126.3	6.343	53.70	6.221	13.34
	80	8	8.164	152.5	8.290	59.58	8.159	13.39
	100	10	10.135	178.7	10.29	64.33	10.132	13.39
	200	20	20.056	308.2	20.20	89.47	20.048	13.41
15	20	3	3.445	60.72	3.45	40.34	3.387	13.22
	40	6	6.202	100.2	6.274	47.84	6.200	13.25
	60	9	9.112	126.8	9.245	52.95	9.106	13.48
	80	12	12.076	152.4	12.205	58.56	12.062	13.49
	100	15	15.048	178.1	15.209	63.05	15.048	13.58
	200	30	30.01	307.01	30.143	86.95	30.009	13.72

is slightly better than the other two methods in terms of having the smallest swamping effect. However, other than at 5% of HLPs and $n = 20$, the DRGP(RFCH) outperforms other methods regardless of sample size and percentage of HLPs, followed by DRGP(MVE) and DRGP(ISE). In this situation, the DRGP(RFCH) consistently having the highest percentage of correct detection of HLPs and the least percentage of swamping and masking effects. Similar conclusion can be made for $p = 4$ where again the DRGP(RFCH) shows the best result followed by DRGP(MVE) and DRGP(ISE). Let us now focus on the computer running time for our DRGP(RFCH) compared to DRGP(ISE) and DRGP(MVE) as displayed in Table 5 and Figure 1. Table 5 presents the computer running times in seconds, the average HLPs detected by DRGP(RFCH), DRGP(ISE) and DRGP(MVE) and the actual number of HLPs planted in the dataset. At 5% of HLPs, $n = 20$, DRGP(RFCH) slightly over detected the HLPs because the average of HLPs detected is slightly larger than the actual HLPs planted in the data. Nonetheless, the computer running time for the DRGP(RFCH) is much smaller than the DRGP(ISE) and DRGP(MVE). On other scenarios, the average of HLPs detected by DRGP(RFCH) consistently the nearest to the actual HLPs planted in the data, followed by the DRGP(MVE) and DRGP(ISE). It is also interesting to see that the computer running times for the DRGP(RFCH) was consistently the least, followed by the DRGP(ISE) and DRGP(MVE). The results are depicted in Figure 1 for clear and quick visualization. The results for $p = 4$ and greater than $p = 4$ are consistent and not reported here due to space limitation.

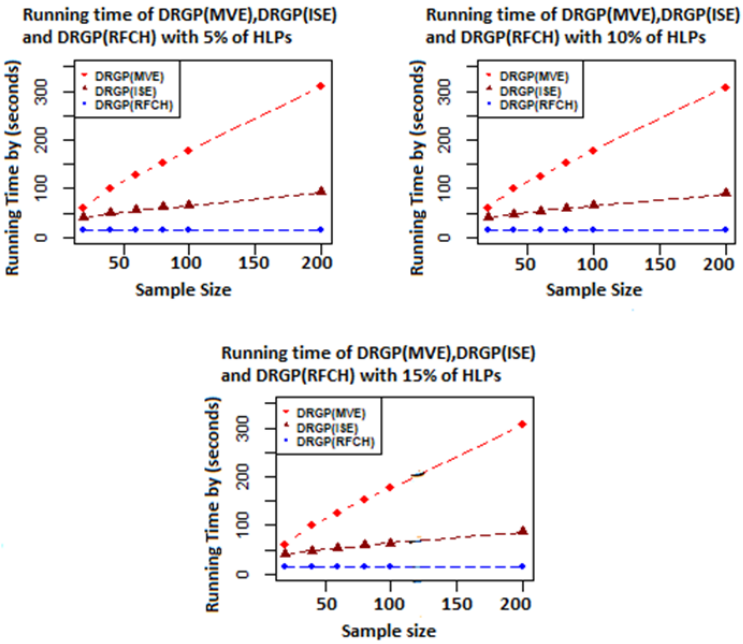

Figure 1. The running time for DRGP(MVE), DRGP(ISE) and DRGP(RFCH), at various sample size

Real Example

A real Hand Grip Strength dataset (Hossain et al., 2012) was used to evaluate the performance of our proposed DRGP(RFCH) method. In this study, we considered a sample of size 196 men, comprising of healthy staff, medical students and visitors of University of Malaya Medical Center between January and April. Four explanatory variables (Age, Height, Weight and BMI) were considered in this study and the dependent variable is the right-hand grip strength. The DRGP(RFCH), DRGP(ISE) and DRGP(MVE) were then applied to the data. The number of HLPs detected by each method is displayed in Figure 2. It is interesting to see from the graph of Figure 2(a) and 2(c) that both DRGP(RFCH) and DRGP(MVE), having the same cut-off points, detected the same observations as HLPs (cases 24, 45, 91, 107, 137, 140, 183). On the other hand, as expected the DRGP(ISE) with cut-off point 0.0615 does not detect the same number of observations. It detects only six observations as HLPs (cases 24, 45, 91, 107, 137, 140, 183) where its masked case 107. The value of DRGP(ISE) which corresponds to case 107 is less than the cut-off point 0.0615. The results of real data are consistent with the results of simulation study where the DRGP(ISE) suffers from swamping and masking effect.

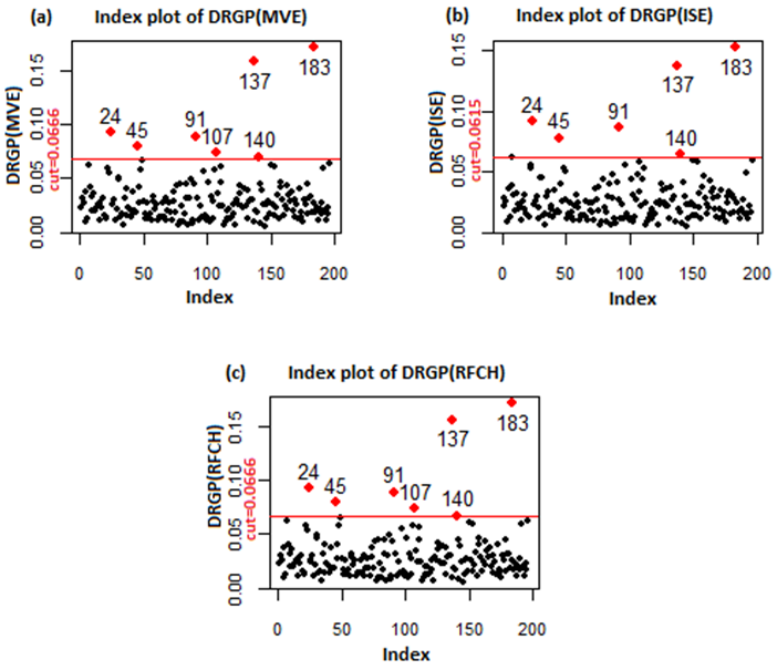

Figure 2. The number of detected HLPs by DRGP(MVE), DRGP(ISE) and DRGP(RFCH)

CONCLUSION

The main aim of this paper is to propose another diagnostic method of detecting HLPs that we call DRGP(RFCH). The existing DRGP(MVE) is quite successful in identifying HLPs but its running time is very slow. The DRGP(ISE) running time is much faster than the DRGP(MVE). However, the DRGP(ISE) was not computationally stable and still possessed masking and swamping effect. Contrarily, the propose DRGP(RFCH) is very successful in detecting HLPs with negligible swamping effect. Moreover, it is based on $\sqrt{n}$ RFCH consistent estimators of location and scatter. The numerical study also signifies that the DRGP(RFCH) needs much lesser computer running time and computationally very stable in the sense of having consistent estimated values. The results of this study appear to recommend that the DRGP(RFCH) may give the most appealing diagnostic method for the identifying HLPs in multiple linear regression model.

ACKNOWLEDGMENTS

This article was partially supported by the Fundamental Research Grant Scheme (FRGS) under Ministry of Education with project number FRGS/1/2019/STG06/UPM/01/1

REFERENCES

Alkenani, A., & Yu, K. (2013). A comparative study for robust canonical correlation methods. *Journal of Statistical Computation and Simulation*, *83*(4), 692-720.

Bagheri, A., Midi, H., & Imon, A. H. M. R. (2012). A novel collinearity-influential observation diagnostic measure based on a group deletion approach. *Communications in Statistics-Simulation and Computation, 41*(8), 1379-1396.

Chatterjee, S., & Hadi, A. S. (2006). *Regression analysis by example*. New Jersey, USA: John Willey & Sons.

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.

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

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.

Hoaglin, D. C., & Welsch, R. E. (1978). The hat matrix in regression and ANOVA. *The American Statistician, 32*(1), 17-22.

Hossain, M. G., Zyroul, R., Pereira, B. P., & Kamarul, T. (2012). Multiple regression analysis of factors influencing dominant hand grip strength in an adult Malaysian population. *Journal of Hand Surgery (European Volume), 37*(1), 65-70.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *The Annals of Statistics, 1*(5), 799-821.

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

Imon, A. H. M. R., & Khan, M. A. I. (2003). A solution to the problem of multicollinearity caused by the presence of multiple high leverage points. *International Journal of Statistical Sciences*, *2*, 37-50.

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

Lopuhaa, H. P. (1999). Asymptotics of reweighted estimators of multivariate location and scatter. *Annals of Statistics*, *27*(5), 1638-1665.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *National Institute of Science of India, 2*(1), 49-55.

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

Midi, H., Rana, S., & Imon, A. H. M. (2014). Tow-step robust estimator in heteroscedastic regression model in the presence of outliers. *Economic Computation and Economic Cybernetics Studies and Research, 48*(3), 255-272.

Olive, D. J., & Hawkins, D. M. (2008). *High breakdown multivariate estimators*. Retrieved September 5, 2019, from https://www.researchgate.net/profile/David_Olive2/publication/240737720_High_Breakdown_Multivariate_Estimators/links/0a85e53234b7db7f90000000.pdf.

Olive, D. J., & Hawkins, D. M. (2010). *Robust multivariate location and dispersion*. Retrieved September 5, 2019, from https://www.researchgate.net/profile/David_Olive2/publication/228434748_Robust_multivariate_location_and_dispersion/links/02bfe51015be5c88ca000000.pdf.

Peña, D., & Yohai, V. J. (1995). The detection of influential subsets in linear regression by using an influence matrix. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 145-156.

Riazoshams, H., Midi, H., & Sharipov, O. S. (2010). The performance of robust two-stage estimator in nonlinear regression with autocorrelated error. *Communications in Statistics-Simulation and Computation*, *39*(6), 1251-1268.

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications, 37*(8), 283-297.

Rousseeuw, P. J., & Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, *88*(424), 1273-1283.

Rousseeuw, P. J., & Driessen, K. V. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics, 41*(3), 212-223.

Rousseeuw, P., & Leroy, A. M. (1987). *Robust regression and outlier detection*. New York, USA: Wiley Series in Probability and Mathematical Statistics.

Rousseeuw, P., & Yohai, V. (1984). Robust regression by means of S-estimators. In J. Franke, W. Härdle & D. Martin (Eds.), *Robust and nonlinear time series analysis* (pp. 256-272). New York, NY: Springer.

Salleh, R. (2013). *A robust estimation method of location and scale with application in monitoring process variability* (Doctoral dissertation). Universiti Teknologi Malaysia, Malaysia.

Uraibi, H. S., Midi, H., & Rana, S. (2017). Selective overview of forward selection in terms of robust correlations. *Communications in Statistics-Simulation and Computation, 46*(7), 5479-5503.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, *48*, 817-838.

Zhang, J., Olive, D., & Ye, P. (2012). Robust covariance matrix estimation with canonical correlation analysis. *International Journal of Statistics and Probability, 1*(2), 119-136.