Pak. J. Statist.
2015 Vol. 31(5), 461-472

THE PERFORMANCE OF ROBUST-DIAGNOSTIC F IN THE IDENTIFICATION OF MULTIPLE HIGH LEVERAGE POINTS

Habshah Midi[1,2] and **Nor Mazlina Abu Bakar**[2,3,§]

[1] Institute of Mathematical Research, Universiti Putra Malaysia
Serdang, Malaysia.
[2] Faculty of Science, Universiti Putra Malaysia, Serdang, Malaysia
[3] Universiti Sultan Zainal Abidin, Terengganu, Malaysia
[§] Corresponding author Email: normazlina@gmail.com

ABSTRACT

High leverage points have undue effects on the Least Square estimates. They are responsible for misleading conclusions in regression and multicollinearity problems. Hence, it is imperative to detect high leverage points and use robust estimators to estimate the parameters of a regression model, so as to arrive at valid conclusions. Several well-known methods have failed to detect multiple high leverage points correctly because of the swamping and/or masking effects. The Diagnostic Robust Generalized Potential (DRGP), is an appealing alternative method that successfully detects high leverage points correctly. However, for small percentages of high leverage points, it has the tendency to identify few low leverage points to be points of high leverage. In this paper, an attempt is made to correctly identify real high leverage point by reducing swamping effects. We propose a method we call Robust Diagnostic-F (RDF), in which robust approach is employed to detect the suspected high leverage points. Then, *F* statistics that relates the change in data covariance structure is used to confirm the suspicion. The performance of RDF is evaluated through real data and simulations. Comparisons are also made with existing methods.

KEY WORDS

Robust; Diagnostic; Outliers; High leverage points; MM estimator.

1. INTRODUCTION

Outliers are considered as inconsistent data or data that behaves differently from the majority of the dataset (Barnett and Lewis, 1994 and Rousseeuw and Van Zomeren, 1990). In regression analysis, observations are labeled as outliers when the fitted regression equation failed to accommodate them. Observations that are located further away from the majority of explanatory variables are called *x*-outlier or popularly known as high leverage points (HLPs). The outliers are usually indicated as observations with excessively large residuals (Imon, 2005). According to Hampel et al. (1986), the existence of 1 – 10% outliers in a routine data is rather a rule than exception. It is expected that a small percentage of outliers may occur in real data sets (Hampel, 1971 and Hampel, 2001). They may arise from many different sources such as incorrect measurement, data entry errors, mechanical faults, human errors or natural phenomenal causes (Maronna et al., 2006). Their occurrences often provide useful information to the data set. Hodge and Austin (2004) provided an exhaustive list on examples of information provided by the occurrence of outliers in a dataset. The existence of outliers may give an indication that machine needs calibration, manufacturing lines detect faulty production of cracked beams or an indication of faults in motors, generators or space instruments. Very more often, the quality of a dataset can deteriorate significantly in the presence of HLPs. Statistical analysis may become unreliable and biased results may be produced under the influence of HLPs. They are responsible for misleading conclusions in regression and also multicollinearity problems (Habshah et al., 2009). Therefore, it is critical to identify HLPs points so that corrective measures can be taken up in order to improve statistical results and analysis.

Many techniques have been introduced to detect HLPs. Masking and swamping are two common problems involved in detecting them. The two effects become more crucial under the influence of multiple HLPs. Masking occurs when inliers are falsely identified as outliers and swamping occurs when inliers are declared as outliers. Non-robust techniques such as Mahalanobis Distance $\left(MD_i\right)$ is known to suffer from masking effect. To remedy the masking problem, Robust Mahalanobis Distance (RMD_i) is introduced to detect HLPs. Unfortunately, the RMD_i is later found to suffer from swamping effect, in which too many points are declared as outliers (Hardin and Rocke, 2005). In an attempt to reduce swamping and masking, a combined use of robust-diagnostic methods is introduced by Habshah et al. (2009). They introduced Diagnostic Robust Generalized Potential or known as DRGP which is a combination of RMD_i and Generalized Potential (GP). The initial stage of DRGP involves dividing the dataset into two subsets; a clean dataset and a suspicious dataset by using RMD_i . Each suspicious data is later checked with the GP to confirm the true outliers. As a result, DRGP is found to successfully identify the true outliers and largely reduce the swamping effect of RMD. Hadi (1992) mentioned that the key factor to a successful identification of real outliers lies in finding the correct initial basic subset. In DRGP, the swamping property of RMD_i provides a new advantage in obtaining a real clean initial data subset. Once a clean data subset is obtained, any outlier entering the data can be detected by another diagnostic method. The results obtained by DRGP (Habshah et al., 2009) seem to be very encouraging, in which the method is proven to successfully detect genuine HLPs. However, the method tends to swamp a few low leverage points for a small percentage of HLPs. This has inspired us to develop another robust-diagnostic method called Robust-Diagnostic F(RDF). The main aim of RDF is to even more reduce the swamping effect; thus, improve the rate of correct detection of genuine HLPs.

The next section describes a few existing outlier detection methods. The proposed procedure which involves the combine use of RMD and a diagnostic measure, F which is proposed by Djauhari (2010) is described in Section 3. Extensive simulations are carried out at different data contamination levels and sample sizes. Results of simulations and numerical examples are reported in Section 4. Rates of detection, rates of swamping and masking for three different outlier detection methods are also measured. The measures are checked in order to investigate the behavior of the proposed method. The performances of RDF in detecting multiple HLPs are studied and compared with the classical and robust detection methods. Discussions are made in the final section.

2. EXISTING OUTLIER DETECTION METHODS

Mahalanobis Distance

Rousseeuw and Leroy (1987) suggested using the well-known Mahalanobis distances as measures of leverages. Let X be an $n \times p$ matrix representing a random sample of size n from a p-dimensional population. Mahalanobis Distance $\left(MD_i\right)$ of the i-th observation is defined as:

$$MD_i = \sqrt{\left(X_i -)\right) C^{-1} \left(X_i - T\right)'} \qquad \text{for } i = 1, \ldots, n \tag{1}$$

T and C are the classical measures of location and shape respectively represented by the arithmetic mean, $\bar{X}$ and the sample covariance matrix, S. MD_i measures the distance of each data point from the centre of mass of the data points based on covariances and variances of each variable. Observations with small MD_i values provide information that the data points lie within the centre of mass of the data points.

The cut-off point is usually taken at the 97.5th percentile of χ_p^2 distribution. Any observation with MD_i values greater than the cutoff point may give an indication of outlyingness. Unfortunately, MD_i are largely affected by multiple high leverages due to the non-robust property of arithmetic mean, $\bar{X}$ and the sample covariance matrix S. In the existence of outliers, the value of arithmetic mean can be inflated or deflated significantly. The arithmetic mean can no longer represent the centre of mass for the contaminated data. The estimate of location is wrongly estimated where the value is largely pulled towards the outliers. As a result, masking may occur where outliers are falsely identified as inliers. Thus, MD_i is largely affected by the presence of outliers or high leverage points.

Robust Mahalanobis Distance

In order to resolve the problem of masking, robust location and scatter estimates such as Minimum Volume Ellipsod (MVE) and Minimum Covariance Determinant (MCD) are used to obtain the Robust Mahalanobis Distance, RMD_i (Rousseeuw and Van Zomeren, 1990). Both MVE and MCD are resistant to outliers and provide the robust location and scatter estimates. Both measures can stand up to 50% outliers which is the maximum breakdown point that can be achieved by any robust estimate.

As the name suggested, the MVE searches the best subset for the minimum volume ellipsoid formed by the data subset. The search aims to allocate the majority of data, $h = \left(n + p + 1\right)/2$ into the ellipsoid and hence provide the robust scatter and location estimates. On the other hand, the MCD searches the best subset which gives the lowest covariance determinant value. In the extensive search, dataset with the minimum determinant is expected to be free of outliers. Both MVE and MCD require a handy amount of time in order to search for the best subset. However, the existence of cheap and fast computing has greatly enhanced the development of these robust estimators (Salibian-Barrera and Yohai, 2006).

Rousseeuw and Leroy (1987) suggested a cut-off point for RMD_i as $\chi^2_{\mathrm{p},0.975}$. This is the same cut-off point as MD_i, where any value greater than the cut-off point may be declared as outliers. This cut-off value comes from the assumption that the *p*-dimensional variables follow a multivariate normal distribution. Nevertheless, in a real life problem there is no guarantee that data would come from a multivariate normal distribution. Another disadvantage of the usual cut-off point is that it depends only on the dimension of the regressors, but does not take any account of the number of observations. To overcome these shortcomings, Imon (2002) suggested a cut-off value for the robust Mahalanobis distances as:

$$\text{Median}\left(RMD_i\right)+3\left(MD_i\right) \text{ for } i=1,2,...,n \tag{2}$$

Problem arises when RMD with MVE (or MCD) are found to be too robust and have a tendency to swamp some inliers (Imon, 2005).

Diagnostic Robust Generalized Potential (DRGP)

In order to reduce the swamping problem of RMD_i , the complementary use of both robust and diagnostic measures has been proposed. From the work of Fung (1993), the complementary use of robust-diagnostic method is found to give some satisfactory results. This is further supported by Habshah et al. (2009) who proposed Diagnostic Robust Generalized Potential (DRGP). DRGP is a combination of RMD_i and the Generalized Potential (GP) which effectively lessen the swamping effect. DRGP involves three important steps:

Step 1:

The deleted data subset is identified by calculating RMD_i with a cutoff point as in Equation (2). Data are now be classified into two different subsets. Any data which exceeds the cutoff point will be grouped into the deleted data subset, D and the clean data in the remaining data subset, R. The swamping property of RMD_i ensures the remaining data subset to be free of outliers.

Step 2:

The GPs for each data point are evaluated. In order to understand GP, let us consider two data subsets; a remaining data subset, R and a deleted data subset, D. Hence, R contains $(n-d)$ points after $d<(n-d)$ points in D are deleted. Without loss of generality, assume that these observations are the last of d rows of X and Y so that the weight matrix $W=X\left(X^TX\right)^{-1}X^T$ can be partitioned as:

$$W=\begin{bmatrix} U_R & V \\ V^T & U_D \end{bmatrix} \tag{3}$$

where $U_R = X_R \left(X^T X \right)^{-1} X_R^T$ and $U_D = X_D \left(X^T X \right)^{-1} X_D^T$ are symmetric matrices of order $(n-d)$ and d respectively and $V = X_R \left(X^T X \right)^{-1} X_D^T$ is $(n-d)\times d$ matrix. The subset of deleted cases indexed by D , can be defined as:

$$w_{ii}^{(-D)} = x_i^T \left(X_R^T X_R \right)^{-1} x_i \quad \text{for} \quad i = 1,2,...,\text{n} \tag{4}$$

From Equation (4), $w_{ii}^{(-D)}$ is the i-th diagonal elements of $X \left(X_R^T X_R \right)^{-1} X^{T'}$ matrix. When the size of R is $(n-1)$ and $D = i$, then it follows that:

$$w_{ii}^{(-D)} = x_i^T \left(X_{(i)}^T X_{(i)} \right)^{-1} x_i = p_{ii} \tag{5}$$

which shows that $w_{ii}^{(-D)}$ as a natural extension of p_{ii} . Any *i*-th case which is removed from the remaining subset R and joins the deletion subset D can be written as

$$w_{ii}^{-(D+i)} = x_i^T \left(X_R^T X_R \right)^{-1} x_i + \frac{\left(x_i^T \left(X_R^T X_R \right)^{-1} x_i \right)^2}{1 - x_i^T \left(X_R^T X_R \right)^{-1} x_i} = \frac{w_{ii}^{-(D)}}{1 - W_{ii}^{-(D)}} \tag{6}$$

GP can then be defined as:

$$p_{ii}^* = \begin{cases} \dfrac{w_{ii}^{(-D)}}{1 - w_{ii}^{(-D)}} & \mathit{for}\ i \in R \\ w_{ii}^{(-D)} & \mathit{for}\ i \in D \end{cases} \tag{7}$$

where D is any arbitrary deleted set of points. The cutoff point of p_{ii} is taken as:

$$\mathit{median}\left(p_{ii} \right) + \mathit{cMAD}\left(p_{ii} \right) \tag{8}$$

where $c = 3$ for this study. In this step, p_{ii} or GP are computed using Equation (7) to check the potential of each member in the set. Any p_{ii} which exceeds the cutoff point will be tagged as outliers.

Step 3:

Update the initial subset. In this step, any inlier detected in Step 2 will be added into the basic initial subset and form a new basic subset. Then the procedure is repeated until every data in the deleted group is checked.

DRGP is found to be successful in reducing the amount of swamping by RMD_i . Nevertheless, the rate of swamping and the rate of correct detection of HLPs by DRGP are still low and need major improvements.

3. THE PROPOSED METHOD – ROBUST DIAGNOSTIC F(RDF)

Robust Diagnostic F , or RDF is proposed not only to reduce the amount of swamping, but also to increase the rate of correct number of outliers detected. RDF also adopts the combine use of robust and diagnostics measures and consists of three major steps:

First Step:

The RMD_i is used to identify the initial basic subset. Based on the RMD_i values, the dataset can be categorized into two subsets – initial basic subset and suspicious subset. The initial basic subset will only contain clean data and the suspicious subset may contain potential outliers due to the swamping property of RMD_i . As suggested by Imon (2002) the cutoff value for RMD_i is taken as Equation (2).

Second Step:

A diagnostic procedure is performed whereby each member of the suspicious subset with the lowest RMD_i value is tested with F statistic as proposed by Djauhari (2010). Consider $X_1, X_2,, X_n, X_{n+1}$ to be a random sample from p-variate normal distribution with covariance matrix Σ and let:

$$SS_k = \sum_{i=1}^{k}\left(X_i - \bar{X}_k\right)\left(X_i - \bar{X}_k\right)^T \quad \text{with} \quad \bar{X} = \frac{1}{k}\sum_{i=1}^{k} X_i \tag{9}$$

where SS_k is the scatter matrix from the clean initial subset when $k = n$ and from the suspicious subset when $k = n+1$ If $D = SS_{n+1} - SS_n$ then $F = \sqrt{Tr(D)}$ represents the effect of X_{n+1} on the scatter matrix of the initial subset. It is understood that any outlier will immediately change the sample variance structure and be detected by the F statistics. Thus, it can be tested whether the addition of another data point can change the covariance structure of the initial subset. From Djauhari (2010) when Σ is unknown, the distribution of F can be approximated by $c\chi_r^2$ where:

$$c = \frac{Tr\left(S_n^2\right)}{Tr\left(S_n\right)} \quad \text{and} \quad r = \frac{\left\{Tr\left(S_n\right)\right\}^2}{Tr\left(S^2_n\right)} \tag{10}$$

S_n is the sample covariance structure of the remaining dataset. Thus, the cutoff point is taken as the $(1-\alpha)$-th quantile of $c\chi_r^2$, where $\alpha = 0.025$. Any data point which exceeds the cutoff value is considered as an outlier.

Third Step:

Update the initial subset. In this step, any inlier detected in Step 2 will be added into the basic initial subset and form a new basic subset. Then the procedure is repeated until every data in the suspicious group is checked.

4. RESULTS

Simulation Results

A simulation study is carried out to compare the performances of our proposed RDF with some existing methods, namely the MD_i , RMD_i and DRGP. The performances of these methods are evaluated based on the average and standard deviations of the number of high leverages points detected. The percentages of correct detection and percentage of masking and swamping are also observed to compare their performances. The experiments are conducted at different sample sizes (n = 20, 40, 100 and 200) and p = 2 and 3. Data are generated from multivariate normal $N\left(0,1\right)$ and further contaminated at 5% or 10% contamination level. The contaminated data are generated from multivariate normal N (5,1) and N (10,1); representing low distance and high distance outliers respectively. Each simulation run comprises of 10,000 replications.

Table 1
Mean and Standard Deviation for the Number of Detected Outliers for in for Multivariate Normally Distributed Data with 5% and 10% Contamination of Outliers Generated from N(5,1) and N(10,1)

Cont. level	*n*	No. of outliers	Low distance outliers				High distance outliers			
			MD	RMD	DRGP	RDF	MD	RMD	DRGP	RDF
p=2										
5%	20	1	1(0.3)	2(1.0)	2(0.9)	1(0.5)	1(0.3)	2(1.0)	2(0.9)	1(0.5)
	40	2	2(0.5)	3(0.9)	3(0.9)	2(0.5)	2(0.4)	3(0.9)	3(0.9)	2(0.5)
	100	5	6(0.8)	6(1.0)	6(0.9)	5(0.6)	6(0.7)	6(1.0)	6(0.9)	5(0.6)
	200	10	11(1.1)	11(1.1)	11(1.1)	10(0.8)	11(1.0)	11(1.1)	11(1.1)	11(0.8)
10%	20	2	2(0.6)	2(0.9)	2(0.8)	2(0.5)	2(0.3)	2(0.8)	2(0.8)	2(0.4)
	40	4	3(0.8)	4(0.7)	4(0.7)	4(0.5)	4(0.5)	4(0.7)	4(0.7)	4(0.5)
	100	10	8(1.3)	10(0.7)	10(0.7)	10(0.5)	10(0.9)	10(0.7)	10(0.7)	10(0.5)
	200	20	16(1.9)	21(0.8)	21(0.8)	20(0.7)	21(1.3)	21(0.8)	21(0.8)	20(0.7)
p=3										
5%	20	1	1(0.3)	2(1.6)	2(1.3)	1(0.3)	1(0.3)	2(1.6)	2(1.3)	1(0.3)
	40	2	2(0.5)	3(1.1)	3(0.9)	2(0.4)	2(0.5)	3(1.1)	3(0.9)	2(0.4)
	100	5	6(0.9)	6(1.0)	6(0.9)	5(0.6)	6(0.9)	6(1.0)	6(0.9)	5(0.6)
	200	10	12(1.3)	11(1.0)	11(1.0)	10(0.7)	12(1.3)	11(1.0)	11(1.0)	10(0.7)
10%	20	2	1(0.6)	3(1.2)	3(1.0)	2(0.3)	2(0.6)	3(1.2)	3(1.0)	2(0.3)
	40	4	3(0.9)	4(0.8)	4(0.7)	4(0.4)	4(0.8)	4(0.8)	4(0.7)	4(0.4)
	100	10	8(1.5)	10(0.7)	10(0.7)	10(0.5)	10(1.3)	10(0.7)	10(0.7)	10(0.5)
	200	20	15(2.0)	20(0.7)	20(0.7)	20(0.6)	19(1.8)	20(0.7)	20(0.7)	20(0.6)

Table 1 exhibits the average number of outliers detected; the standard deviations are in parenthesis. The results clearly indicate that the RDF has a superior ability to detect the correct number of outliers compared to MD_i , RMD_i and DRGP regardless of the contamination level and distance of outliers. As can be expected, MD_i tends to detect a lower number of HLPs at 10% contamination because of masking effects. On the other hand, at 5% contamination level, the RMD_i and $DRGP$ detect few low leverage points to be points of high leverages due to the swamping effects. Nonetheless the standard deviations for rate of correct detection by DRGP are slightly lesser than that of RMD_i . It can be observed from Table 1 that the RDF correctly identifies the exact number of outliers with the least standard deviations.

The effects of masking and swamping are summarized in Table 2 and 3 at different levels of contamination and distance of outliers. Masking can be clearly observed in MD_i and swamping is largely observed in RMD_i and DRGP. However, the swamping effect is slightly reduced in DRGP compares to RMD_i . For RDF, it is particularly remarkable to observe the ability of RDF in reducing swamping especially for small data sets. The performance of RDF is also found to be consistent and stable at different contamination levels and number of p. The RDF also outperforms other methods, including DRGP in detecting correct number of outliers. The rate of correct detection is high even when the outliers are placed at a close distance to the good data. It is also observed that the rate of correct detection for RDF differs significantly with DRGP in small samples. The ability is gained since in the second step of RDF, we merely calculate the variance structure of the data. Thus, RDF is able to identify the structural change only when real outlier is introduced into the clean data set.

Table 2
Percentages of Correctly Identified Outliers, Masking and Swamping for Low Distance Outliers

Cont. level	*n*	**% Correct Detection**				**% Masking**				**% Swamping**			
		MD	**RMD**	**DRGP**	**RDF**	**MD**	**RMD**	**DRGP**	**RDF**	**MD**	**RMD**	**DRGP**	**RDF**
p=2													
5%	20	91.1	59.3	63.2	88.7	0.1	0.5	0.5	0.1	8.8	40.3	36.3	11.2
	40	74.3	62.6	63.2	83.6	1.4	0.3	0.3	0.1	24.3	37.1	36.5	16.3
	100	40.7	54.7	55.5	79.2	3.2	0.1	0.1	0.2	56.1	45.3	44.3	20.6
	200	18.9	40.8	41.3	61.0	2.6	0.1	0.0	0.1	78.6	59.1	58.7	38.9
10%	20	47.7	67.7	70.8	88.9	49.1	1.1	1.1	0.3	3.2	31.2	28.1	10.8
	40	27.3	74.2	74.6	83.9	68.7	0.7	0.7	0.1	4.0	25.2	24.8	16.0
	100	10.9	70.0	70.6	81.7	85.8	0.4	0.4	0.4	3.3	29.6	29.0	17.9
	200	2.4	59.9	60.8	68.4	96.8	0.4	0.3	0.5	0.8	39.7	38.9	31.1
p=3													
5%	20	92.8	35.7	53.0	92.4	0.0	0.2	0.3	0.0	7.2	64.2	46.7	7.6
	40	73.0	54.4	59.3	85.8	0.7	0.0	0.0	0.0	26.3	45.6	40.7	14.2
	100	34.1	55.3	58.7	81.9	1.3	0.0	0.0	0.0	64.6	44.7	41.3	18.1
	200	11.4	46.3	46.7	65.6	1.0	0.0	0.0	0.0	87.6	53.7	53.3	34.4
10%	20	25.8	49.7	63.5	93.3	73.0	0.2	0.2	0.0	1.2	50.1	36.3	6.7
	40	15.8	67.3	70.8	86.7	81.7	0.1	0.1	0.0	2.5	32.6	29.2	13.3
	100	10.3	70.9	72.5	85.4	84.5	0.0	0.0	0.0	5.2	29.1	27.5	14.6
	200	1.2	65.7	66.0	73.5	98.3	0.0	0.0	0.0	0.5	34.3	34.0	26.5

Table 3
Percentages of Correctly Identified Outliers, Masking and Swamping High Distance Outliers

Cont. level	n	% Correct Detection				% Masking				% Swamping			
		MD	RMD	DRGP	RDF	MD	RMD	DRGP	RDF	MD	RMD	DRGP	RDF
p=2													
5%	20	92.5	59.6	63.7	89.1	0.0	0.0	0.0	0.0	7.5	40.4	36.3	10.9
	40	21.0	62.9	63.5	83.7	78.3	0.0	0.0	0.0	0.8	37.1	36.5	16.3
	100	46.3	54.8	55.6	78.6	0.0	0.0	0.0	0.0	53.7	45.2	44.4	21.4
	200	20.4	40.8	41.3	61.5	0.0	0.0	0.0	0.0	79.6	59.2	58.7	38.5
10%	20	88.2	69.0	72.4	89.5	3.9	0.0	0.0	0.0	7.9	31.0	27.6	10.5
	40	72.2	74.7	75.1	83.9	8.9	0.0	0.0	0.0	18.9	25.3	24.9	16.1
	100	44.5	69.7	71.0	81.8	13.0	0.0	0.0	0.0	42.5	30.3	29.0	18.2
	200	27.6	60.9	59.7	69.0	14.0	0.0	0.0	0.0	58.5	39.1	40.3	31.0
p=3													
5%	20	93.2	35.8	53.3	92.4	0.0	0.0	0.0	0.0	6.8	64.2	46.7	7.6
	40	75.2	54.3	59.3	85.9	0.0	0.0	0.0	0.0	24.9	45.7	40.7	14.1
	100	36.3	55.3	58.7	81.8	0.0	0.0	0.0	0.0	63.7	44.7	41.3	18.2
	200	11.8	46.2	46.7	66.1	0.0	0.0	0.0	0.0	88.2	53.8	53.3	33.9
10%	20	57.1	49.9	63.7	93.3	39.3	0.0	0.0	0.0	3.6	50.2	36.3	6.7
	40	44.8	67.4	70.8	86.8	43.4	0.0	0.0	0.0	11.9	32.6	29.2	13.2
	100	30.0	70.7	72.4	85.3	45.8	0.0	0.0	0.0	24.2	29.3	27.6	14.8
	200	21.2	65.6	66.0	73.9	54.0	0.0	0.0	0.0	24.8	34.4	34.0	26.1

Numerical Examples

We consider three well-known data sets to illustrate the performance of RDF and compare its performance with existing methods to identify multiple HLPs. The datasets are aircraft data, stackloss data and Hawkin-Bradu-Kass (HBK) data which consist of data with sample sizes of $n=23$, $n=21$ and $n=75$; respectively. Outliers are known to exist prominently in stackloss and HBK data (Hawkins et al., 1984). The results are reported in Table 4 which summarizes the abilities of RDF, DRGP, MD_i and RMD_i to identify HLPs in all three datasets.

Table 4
High Leverage Points Detected by Different Methods in Three Well-Known Datasets

Data	Sample size, n	No. of variables	Number of High Leverage Points Detected/Datapoints			
			MD	RMD	DRGP	RDF
Aircraft Data	23	4	2 (14,22)	3 (14,20,22)	1 (22)	1 (22)
Stackloss Data	21	3	1 (21)	8 (1,…,4,13,14,20,21)	4 (1,2,3,21)	3 (1,2,3)
HBK Data	75	3	2 (12,14)	14 (1, 2,…,14)	14 (1, 2,…,14)	14 (1, 2,…,14)

RDF has the ability to detect even small changes in the structure of the dataset. Thus, any aberrant data introduced in a clean dataset is detected as HLPs by RDF whereas MD_i and RMD_i are prone to masking and swamping. DRGP is also found to be subjected to swamping even though the effect is largely reduced. The RDF detects the same number of HLPs as DRGP for aircraft and HBK data.

The parameter estimates and collinearity diagnostics for aircraft and HBK data with and without HLPs are presented in Tables 5 and 6. Here, we introduced a robust MM estimator which is resistant to outliers, highly efficient and has high breakdown point[18]. Let us first focus to Table 5 to see the effect of HLPs on the LS and MM parameter estimates. It is important to note that a good estimator is the one that has parameter estimates and standard errors which are fairly closed to the LS estimates in the absence of HLPs. We can see that the standard errors of the LS estimates in the presence of HLPs are larger than those without HLPs. The HLPs altered not only the LS estimates and its standard errors but also the signs of β_2, β_3 for HBK data and β_4 for aircraft data. Consequently, misleading conclusions are obtained when the commonly LS method is used to analyze contaminated data. It is interesting to observe that the robust MM estimator is not easily affected by HLPs and its estimates are reasonably close to the LS estimates without HLPs.

Table 6 shows the collinearity diagnostics for Aircraft and HBK data under the influence of outliers and also when outliers are removed from the data. For aircraft data, we observe that a single outlier causes Pearson’s Correlation coefficients and variance inflation factors (VIF) to be inflated/deflated at some degree. More influence can be observed in HBK data where all values of correlation coefficients and VIFs are hugely inflated. This situation is referred as high leverage collinearity-enhancing observations; those HLPs that induce multicollinearity pattern of a data (Yohai, 1987). From these two tables, we can clearly see the damaging effect of HLPs on the LS estimates which need to be rectified. The alarming influence of HLPs indicates the critical need of the HLPs to be detected before further analysis.

Table 5
Least Squares (LS) and MM Estimates with Standard Errors (in Parenthesis) for Aircraft and Hawkin-Bradu-Kass (HBK) Data

Data	Coef.	With High Leverage Points		Without High Leverage Point
		LS	MM	LS
Aircraft Data (original data has 14 HLP)	β_1	-3.853(1.763)	-3.049(1.003)	-3.353(1.109)
	β_2	2.488(1.187)	1.210(0.708)	1.525(0.645)
	β_3	0.004(0.001)	0.001(0.001)	0.002(0.001)
	β_4	0.002(0.001)	-0.001(0.001)	-0.001(0.001)
HBK Data (original data has 1 HLP)	β_1	0.240(0.263)	0.081(0.073)	0.062(0.069)
	β_2	-0.335(0.155)	0.040(0.044)	0.012(0.068)
	β_3	0.383(0.129)	-0.052(0.040)	-0.107(0.071)

Table 6
Collinearity Diagnostics for Selected Datasets to Show the Influence of Outliers

Data	Variable	Pearson's Correlation Coef.		VIF	
		With Outliers	Without Outliers	With Outliers	Without Outliers
Aircraft Data	1	r_{12} = – 0.151	– 0.024	1.927	2.476
	2	r_{23} =0.336	0.050	1.431	1.511
	3	r_{24} =0.463	0.337	6.501	3.419
	4	r_{34} =0.914	0.807	8.433	5.638
HBK Data	1	r_{12} =0.946	0.044	13.432	1.012
	2	r_{23} =0.979	0.127	23.853	1.017
	3	r_{13} =0.962	0.107	33.432	1.027

5. CONCLUSIONS

In the existence of HLPs, the LS estimator produces sub-optimal or even invalid inferential statements and inaccurate predictions. In this situation, MM estimator is recommended because it gives more efficient estimates. This is the reason why it is important to identify HLPs as they are responsible for the misleading inferences about the fitting of the regression model. By correctly detecting those observations, may help statistics practitioners to use appropriate statistical methods to analyze their data. In this regard, we have proposed a robust diagnostic measure, RDF to identify multiple HLPs. Its performance is compared to MD_i and robust outlier detection methods such as RMD_i and DRGP. The overall results indicate that MD_i is not efficient at all; it suffers from masking effects. The RMD_i and the *DRGP* tend to swamp few low leverage points. Nevertheless, the swamping rate of the DRGP is lesser than the RMD_i . The numerical examples and simulation study signify that the RDF offers a substantial improvement over the other existing methods. The RDF successfully identify high leverage points with the lowest rate of swamping. We relate the effectiveness of this method to the key factor- F statistic that detects the change in data covariance structure upon the entrance of outlying values into the data set. In order to obtain reliable inferences, the robust MM estimator is recommended when HLPs occurs in the data.

REFERENCES

1. Bagheri, A., Habshah, M. and 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.
2. Barnett, V. and Lewis, T. (1994). *Outliers in Statistical Data*, John Wiley & Sons, Third Edition.
3. Djauhari, M. (2010). A Multivariate Process Variability Monitoring Based on Individual Observations. *Modern Applied Science*, 4(10). P91.
4. Fung, W. (1993). Unmasking Outliers and Leverage Points: A Confirmation. *Journal of the American Statistical Association*, 88, 515-519.
5. Habshah, M., Norazan, M.R. and 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, 507-520.
6. Hadi, A.S. (1992). Identifying Multiple Outliers in Multivariate Data. *Journal of the Royal Statistical Society B*, 54, 761-771.
7. Hampel, F.R. (1971). A General Qualitative Definition of Robustness. *The Annals of Mathematical Statistics*, 42(6), 1887-1896.
8. Hampel, F.R. (2001). Robust Statistics: A Brief Introduction and Overview. *Research Report 24*, Seminar for Statistics.
9. Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel W.A. (1986). *Robust Statistics: The Approach Based on Influence Functions*, Wiley, New York.
10. Hardin, J. and Rocke, D.M. (2005). The Distribution of Robust Distances. *Journal of Computational and Graphical Statistics,* 14, 928-946.
11. Hawkins, D.M., Bradu, D. and Kass, G.V. (1984). Location of Several Outliers in Multiple Regression Data Using Elemental Sets. *Technometrics*, 26(3), 197-208.
12. Hodge, V.J. and Austin, J. (2004). A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review*, 22(2), 85-126.
13. Imon, A.H.M.R. (2002). Identifying Multiple High Leverage Points in Linear Regression. *Journal of Statistical Studies,* 3, 207-218.
14. Imon, A.H.M.R. (2005). Identifying Multiple Influential Observations in Linear Regression. *Journal of Applied Statistics*, 32, 929-946.
15. Maronna, R.A., Martin, R.D. and Yohai, V.J. (2006). *Robust Statistics: Theory and Methods,* John Wiley, New York.
16. Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust Regression and Outlier Detection*, John Wiley and Sons, New York.
17. Rousseeuw, P.J. and Van Zomeren, B.C. (1990). Unmasking Multivariate Outliers and Leverage Points. *Journal of the American Statistical Association,* 85, 633-639.
18. Salibian-Barrera, M. and Yohai, V.J. (2006). A Fast Algorithm for S-Regression Estimates. *Journal of Computational and Graphical Statistics*, 15(2), 414-427.
19. Yohai, V.J. (1987). High Breakdown-Point and High Efficiency Estimates for Regression. *The Annals of Statistics*, 15, 642-65.