Journal of Mathematics and Statistics 5 (2):97-106, 2009
ISSN 1549-3644

Two-Step Robust Diagnostic Method for Identification of Multiple High Leverage Points

[1]Arezoo Bagheri, [1]Habshah Midi and [2]A.H.M. Rahmatullah Imon
[1]Laboratory of Applied and Computational Statistics, Institute for Mathematical Research,
University Putra Malaysia, 43400 Serdang, Selangor, Malaysia
[2]Department of Mathematical Sciences, Ball State University Muncie, IN 47306, USA

Abstract: Problem statement: High leverage points are extreme outliers in the X-direction. In regression analysis, the detection of these leverage points becomes important due to their arbitrary large effects on the estimations as well as multicollinearity problems. Mahalanobis Distance (MD) has been used as a diagnostic tool for identification of outliers in multivariate analysis where it finds the distance between normal and abnormal groups of the data. Since the computation of MD relies on non-robust classical estimations, the classical MD can hardly detect outliers accurately. As an alternative, Robust MD (RMD) methods such as Minimum Covariance Determinant (MCD) and Minimum Volume Ellipsoid (MVE) estimators had been used to identify the existence of high leverage points in the data set. However, these methods tended to swamp some low leverage points even though they can identify high leverage points correctly. Since, the detection of leverage points is one of the most important issues in regression analysis, it is imperative to introduce a novel detection method of high leverage points. **Approach:** In this study, we proposed a relatively new two-step method for detection of high leverage points by utilizing the RMD (MVE) and RMD (MCD) in the first step to identify the suspected outlier points. Then, in the second step the MD was used based on the mean and covariance of the clean data set. We called this method two-step Robust Diagnostic Mahalanobis Distance (RDMD^{TS}) which could identify high leverage points correctly and also swamps less low leverage points. **Results:** The merit of the newly proposed method was investigated extensively by real data sets and Monte Carlo Simulations study. The results of this study indicated that, for small sample sizes, the best detection method is (RDMD^{TS}) (MVE)-mad while there was not much difference between (RDMD^{TS}) (MVE)-mad and (RDMD^{TS}) (MCD)-mad for large sample sizes. **Conclusion/Recommendations:** In order to swamp less low leverage as high leverage point, the proposed robust diagnostic methods, (RDMD^{TS}) (MVE)-mad and (RDMDTS) (MCD)-mad were recommended.

Key words: Multicollinearity, high leverage points, robust mahalanobis distance, distance-distance plot

INTRODUCTION

Outliers are observations which break the pattern shown by the majority of the data set. They can be classified in the following categories: (1) Good leverage points: Observations which follow the same regression line as the other data in the data set although they fall far from the majority of the explanatory variables (2) Bad leverage points: Observations not only deviate from the same regression line as the other data in the data set but also fall far from the majority of explanatory variables, (3) Vertical Outliers or high y residual outliers: Observations which are not leverage points but have high response variables residuals[19]. Generally, those leverages that are far from the rest of the other x variables are high leverage points. It is now evident that outliers have some destructive effects on regression fitted line. Rousseeuw and Van Zomeren[25] pointed out that high leverages can affect the estimated slope of the regression line in Ordinary Least Squares (OLS), thus may cause more serious problems than other outliers which might only affect the estimated intercept term. Moreover, their presence in regression models may make some low leverage as high leverage and vice versa. These two concepts are called masking and swamping in linear regression[23]. Furthermore, the

Corresponding Author: Arezoo Bagheri, Laboratory of Applied and Computational Statistics,
Institute for Mathematical Research, University Putra Malaysia, 43400 Serdang, Selangor, Malaysia

range of explanatory variables increases when they exist in regression analysis. Thus, the multiple coefficient determination statistics (R^2) which is a well-known and popular measure of goodness-of-fit in the regression models will increase even by any changes of a single x variable[26]. In addition, high leverages may be the prime source of collinearity-influential observations whose presence can make collinearity and can destroy the existing collinearity pattern among the x variables[7]. In this respect, the identification of high leverage points to prevent their effect on linear regression becomes necessary.

Outlier detection is one of the most important tasks in data analysis. The outliers describe the abnormal data behavior, i.e., data which are deviating from the natural data variability. Various methods for detecting outliers have been studied[1,2,5,7,8,18,21,25]. One way to identify possible multivariate outliers is to calculate a distance from each point to a center of the data. An outlier would then be a point with a distance larger than some predetermined value. For a p-dimensional multivariate sample x_i (i = 1,.., n), the Mahalanobis Distance (MD) is defined as:

$$MD_i = \sqrt{(X - T(X))'C(X)^{-1}(X - T(X))} \text{ for } i = 1,...,n \qquad (1)$$

Where:

T(X) = The estimated multivariate location which is usually the multivariate arithmetic mean

C(X) = The estimated covariance matrix which is usually the sample covariance matrix

The distribution of the MD with both the true location and shape parameters and the conventional location and shape parameters is well known[5]. If there are only a few outliers, large values of MD, indicate that the point x_i is an outlier[2]. Any value of which the MD exceeds the cutoff $\sqrt{\chi^2_{p,0.975}}$ is considered as outliers where p is the number of explanatory variables[16]. Data sets with multiple outliers are subject to problems of masking and swamping[20]. Masking occurs when a group of outlying points skews the mean and covariance estimates toward these points and the resulting distance of the outlying point from the mean is small. While, swamping occurs when a group of outlying points skews the mean and covariance estimates toward these points and away from other inlying points and the resulting distance from the inlying points to the mean is large. Mahalanobis Distance is known to suffer from masking problems[24]. Mahalanobis Distances give a one-dimensional measure of how far a point is from a location with respect to a shape. Utilizing MD, we can find the points that are unusually far away from a location and call those points outlying. A large body of diagnostic tools is available in the literature for detection of high leverage points in linear regression[4,11,12,27]. Mahalanobis Distance (MD) is one of these well-known multivariate methods for detecting high leverage points as well. Although it is a reliable diagnostic tool for detecting high leverage points, it suffers from masking problem. Most of the classical diagnostic methods fail to identify the multiple high leverage points due to their masking effects[14]. Problems of masking can be resolved by using robust estimates of shape and location, which by definition are less affected by outliers. Outlying points are less likely to enter into the calculation of the robust procedures, so they will not be able to influence the parameters used in the MD. The inlying points, which all come from the underlying distribution, will completely determine the estimate of the location and shape of the data. Several robust estimators of multivariate location and scatter have been proposed, such as Maronna’s pioneering paper on multivariate M-estimation [17], the Minimum Volume Ellipsoid (MVE) and the Minimum Covariance Determinant (MCD) estimators by Rousseeuw[22]. For a thorough overview of robust multivariate estimation, one can refer to the article by Maronna and Yohai[18].

The Minimum Covariance Determinant (MCD) method of Rousseeuw[22] aims to find h observations out of n whose covariance matrix C has the lowest determinant. In the Minimum Volume Estimator (MVE), proposed by Rousseeuw[22], an ellipsoid of the smallest volume with a subset of p objects (non-contaminated data) is constructed. In one of the proposed iterative algorithms, n+1 object is selected iteratively at random in each of iterations and their mean and covariance are determined. Then, the ellipsoid containing exactly p data objects is found by deflating or expanding the data covariance. The steps of the algorithm are repeated until the subset of p objects yielding the smallest volume of the covariance ellipsoid is found.

Finally the robust MD distance can be written as:

$$RMD_i = \sqrt{(X - T_R(X))'C_R(X)^{-1}(X - T_R(X))} \text{ for } i = 1,...,n \qquad (2)$$

where, T_R (X) and C_R (X) are robust location and shape estimate such as MCD or MVE. By using a robust location and shape estimate in the RMD, outlying points will not skew the estimates and can be identified as outliers by large values of the RMD. Unfortunately, using robust estimates gives RMDs with unknown distributional properties[25]. The use of $\sqrt{\chi^2_{p,0.975}}$ quantile as cutoff point for RMD will prone to declare some good and low leverage as high leverage point-sand often lead to identifying too many points as outliers[25]. To develop robust multivariate estimators, Rousseeuw and Leroy[23] first proposed to detect outliers by RMD and then find the estimates by using the reweighted least squares regression when the weight function is a hard rejection function. Specifically, the latter proposal consists of discarding those observations whose RMD exceeds a certain fix threshold value. Previously, the MVE was commonly used as initial estimator for these procedures. In the context of linear regression, many estimators have been proposed that aim to reconcile high efficiency and robustness. Typically, these methods are also two-stage procedures[6,10,15,22,28,29].

Let us consider a k variables regression model as:

$$Y = X\beta + \in \tag{3}$$

The weight matrix $W = X\ (X^T\ X)^{-1}\ X^T$ is the orthogonal projector matrix onto the model space, or hat matrix which is traditionally used as a measure of leverage points in regression analysis. If a diagonal entry W_{ii} of W is large, changing y_i will move the fitted surface appreciably towards the altered value. Therefore, W_{ii} is said to measure the leverage of the observation y_i. Different cutoff points exist in the literature for the hat matrix to find high leverage points such as twice-the- mean-rule (2 k/n) by[11], thrice-the-mean-rule (3 k/n)[27] when k and n are the number of variables and observations respectively and three interval range of Huber[12] (observations with $0.2 < W_{ii} < 0.5$ are risky to consider in analysis and those with $W_{ii} \geq 0.5$ should be avoided when W_{ii} is diagonal elements of hat matrix).

The hat matrix may fail to identify the high leverage points because of the effect of high leverage points in leverage structure[7]. Hadi[7] introduced another diagnostic tool as follows:

$$p_{ii} = \frac{w_{ii}}{1 - w_{ii}} \tag{4}$$

where, $w_{ii} = x_i^T(X^T\ X)^{-1}\ x_i$ is the diagonal element of W and the i-th diagonal potential p_{ii} can be defined as:

$$p_{ii} = x_i^T (X_{(i)}^T X_{(i)})^{-1} x_i$$

where, $X_{(i)}$ is the data matrix X without the i-th row. He proposed a cutoff point for potential values (p_{ii}) as Median (p_{ii}) +c Mad (p_{ii}) where Mad = median $|p_{ii}$-median$(p_{ii})|/0.6745$ and c can be taken as constant values of 2 or 3. Observations exceeding Hadi's cutoff point is considered as high leverage points. But this method also can't detect all of the high leverage points.

Imon[13] introduced another diagnostic tool as generalized potentials for the whole data set as follows: Let consider that D is deleted group from data set, those which suspected as outliers (the choice of this deletion group is very important since the omission of this group determines the weights for the whole data set). R is the remaining set after deleting d<(n-k) therefore it contains (n-d) cases. If we assume that the suspected data are the last d rows of X and Y so the weight matrix $W = X\ (X^T\ X)^{-1}\ X^T$ can be written as:

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

where, $U_R = X_R(X^T\ X)^{-1}\ X_R^T$ and $U_D = X_D(X^T\ X)^{-1}\ X_D^T$ are symmetric matrices of order (n-d) and d respectively. $V = X_R(X^T\ X)^{-1}\ X_D^T$ is an (n-d)×d matrix. Now we can define:

$$w_{ii}^{(-D)} = x^T{}_i\,(X_R{}^T\ X)^{-1} x_i\ , \text{ for } i = 1,2,\ldots n$$

where, $w_{ii}^{(-D)}$ is the i-th diagonal element of $X(X_R{}^T\ X_R)^{-1}\ X^T$ matrix.

Then Imon[14] introduced generalized potentials for all members in a data set which are defined as:

$$\begin{aligned} p_{ii}^* &= \frac{w_{ii}^{(-D)}}{1 - w_{ii}^{(-D)}} && \text{for } i \in R \\ &= w_{ii}^{(-D)} && \text{for } i \in D \end{aligned} \tag{5}$$

We should notice that there isn't any finite upper bound for pii* 's and the derivation of the theoretical distribution of them are not easy. He introduced the same cutoff point as potential values Median (p_{ii}^*) + c Mad (p_{ii}^*) for the generalized potential as well.

Habshah *et al.*[6] developed a new method for determining outlying points in multivariate data set by combining the RMD (MVE) method for detecting the suspected group (D group) in generalized potential method which is proposed by[14]. This method which is called DRGP (MVE) is also a two-step method for high leverage point detection. In their methods, the mad cutoff point has been used in the first and second steps.

However, this method can identify more swamped low leverage points. According to Werner [28], "A successful method of identifying outliers in all multivariate situations would be ideal, but is unrealistic". By "successful", he means both highly sensitive, the ability to detect genuine outliers and highly specific, the ability not to swamp regular points as outliers. Therefore a practical and efficient robust detection method of high leverage points (outliers in X-direction) is the method which is sensitive to detect genuine high leverage points and specific, thus it swamps less low leverage as high leverage.

MATERIALS AND METHODS

In this study, we propose a two-step diagnostic tool for detecting multiple high leverage points which can detect less swamped low leverages. In order to improve DRGP (MVE) performance proposed by[6], we follow the idea of Rousseeuw and Leroy[23] in developing robust multivariate estimators and propose a relatively new method for high leverage points identification which is called two-steps Robust Diagnostic Mahalanobis Distance ($RDMD^{TS}$). In the first step, the RMD (MCD) or RMD (MVE) method is used to detect the suspected outlier group which will be deleted from the data set resulting in the clean data for the next step. In the second step, we apply the MD for the entire data set that based on the mean and covariance matrix of the clean data set which was obtained from the first step. Therefore, Two-Steps Robust Diagnostic Mahalanobis Distance ($RDMD^{TS}$) is written as follows:

$$(RDMD)_i^{TS} = \sqrt{(X - T_0(X))'C_0(X)^{-1}(X - T_0(X))} \text{ for } i = 1, \ \dots, \ n \qquad (6)$$

where, $T_0(X)$ and $C_0(X)$ are the mean and covariance matrix of the clean data set. Two different cutoff points are considered, namely the $\sqrt{\chi^2_{k,0.975}}$ where k is the number of explanatory variables and a new proposed one, that is Median ($RDMD^{TS}$) +c Mad ($RDMD^{TS}$). The procedure of this method can be summarized in the following algorithm.

First step:

- Compute RMD_i(MCD) or RMD_i(MVE) for i = 1, …, n which is defined in equation (2) in multivariate cases (both x and y variables)
- Compare these values with $\sqrt{\chi^2_{p,0.975}}$ to detect outliers (if any) where p is the number of x and y variables together

Second step:

- Find the mean and the covariance matrix of the clean subset of the explanatory variables, after removing the suspected outliers in the first step
- Find the classical MD with the mean and covariance matrix of the clean data set in the first step for the entire data (for x variables only)
- Compare these values with $\sqrt{\chi^2_{k,0.975}}$ to detect high leverage points (if any) where k is the number of x variables. We refer to this method as ($RDMD^{TS}$)-chi-sq or:
- Compare these values with Median ($RDMD^{TS}$)+c Mad ($RDMD^{TS}$), to detect high leverage points (if any) where c is an appropriately chosen constant such as 2 or 3. We refer to this method as ($RDMD^{TS}$)-mad
- Those points with $(RDMD^{TS}) < \sqrt{\chi^2_{k,0.975}}$ or ($RDMD^{TS}$)<Median ($RDMD^{TS}$)+c Mad ($RDMD^{TS}$) are not considered high leverage points and are put back in the set of inliers

RESULTS

Numerical Examples: The two well-known data sets which are frequently referred to in the study of the identification of influential observations, high leverage points and outliers are considered in this study. It is important to note here that we changed the cutoff point of mad which is used by [6] to chi-square in the first step of the examples and also in the simulation study.

Hawkins-Bradu-Kass data: Hawkins *et al.*[9] constructed an artificial three-predictor data set containing 75 observations with 10 outliers (cases 1-10) and 14 high leverage points (cases 1-14). Most of the previous single case deletion identification methods fail to identify all of these influential observations. Some of them identify four high leverage points wrongly as outliers[23]. Table 1 shows the DRGP (MVE), DRGP (MCD), ($RDMD^{TS}$) (MVE), ($RDMD^{TS}$)(MCD), MD and their corresponding cutoff points.

Stack loss data: Here we consider the stack loss data[3] that have been extensively analyzed in the statistical literature. This three-predictor data set (Air flow, Cooling water inlet temperature and Acid concentration) contains 21 observations with five influential observations; three of them which (cases 1, 3 and 21) are high leverage outliers. One of the influential observations (case 4) is an outlier and another one (case 2) is a high leverage point. Table 2 illustrates the DRGP (MVE), DRGP (MCD), ($RDMD^{TS}$)(MVE),

Table 1: Diagnostic robust generalized potential based on MVE and MCD and two-step Robust diagnostics Mahalanobis distance based on MVE and MCD for hawkins-Bradu-Kass data

			($RDMD^{TS}$) (MVE)	($RDMD^{TS}$) (MCD)	
	DRGP (MVE)	DRGP (MCD)	$\sqrt{\chi^2_{3,0.975}}$ = (2.7955)	$\sqrt{\chi^2_{3,0.975}}$ = (2.7955)	MD
Index	MAD (0.2068)	MAD (0.2133)	MAD (3.5229)	MAD (3.6293)	$\sqrt{\chi^2_{3,0.975}}$ = 2.7955
1	14.5318	14.8768	29.2642	29.3577	1.9168
2	15.2960	15.6974	30.0246	30.1574	1.8558
3	17.0694	17.4605	31.7193	31.8077	2.3137
4	18.1231	18.5965	32.6845	32.8270	2.2297
5	17.4770	17.9118	32.0961	32.2165	2.1001
6	15.6763	16.0392	30.3961	30.4843	2.1462
7	15.7654	16.1511	30.4824	30.5906	2.0105
8	14.8954	15.2549	29.6285	29.7288	1.9193
9	17.1425	17.5823	31.7871	31.9185	2.2212
10	16.0898	16.4974	30.7947	30.9171	2.3335
11	22.5124	23.1043	36.4314	36.5932	2.4465
12	24.2013	24.7596	37.7742	37.8824	3.1083
13	22.7931	23.4013	36.6580	36.8278	2.6624
14	28.1638	29.1744	40.7515	41.1234	6.3816
15	0.0923	0.0934	2.0008	1.9934	1.8155
16	0.1107	0.1161	2.2127	2.2471	2.1514
17	0.0859	0.0859	1.9188	1.8983	1.3849
18	0.0277	0.0291	0.7782	0.8106	0.8482
19	0.0460	0.0469	1.2703	1.2713	1.1489
20	0.0978	0.0979	2.0675	2.0471	1.5914
21	0.0364	0.0374	1.0442	1.0537	1.0900
22	0.0723	0.0752	1.7302	1.7528	1.5488
23	0.0418	0.0426	1.1755	1.1786	1.0854
24	0.0483	0.0486	1.3178	1.3053	0.9712
25	0.0906	0.0959	1.9794	2.0225	0.7993
26	0.0699	0.0726	1.6942	1.7150	1.1684
27	0.0901	0.0910	1.9731	1.9628	1.4496
28	0.0377	0.0398	1.0779	1.1127	0.8678
29	0.0397	0.0418	1.1278	1.1586	0.5764
30	0.1136	0.1138	2.2441	2.2230	1.5689
31	0.0710	0.0777	1.7106	1.7878	1.8385
32	0.0736	0.0736	1.7499	1.7306	1.3072
33	0.0460	0.0462	1.2686	1.2559	0.9820
34	0.0974	0.1000	2.0620	2.0715	1.1750
35	0.0819	0.0825	1.8664	1.8542	1.2436
36	0.0424	0.0428	1.1904	1.1818	0.8508
37	0.0945	0.0992	2.0278	2.0623	1.8324
38	0.0566	0.0596	1.4748	1.5092	0.7521
39	0.0766	0.0780	1.7927	1.8820	1.2650
40	0.0377	0.0378	1.0771	1.0618	1.1120
41	0.0948	0.0980	2.0308	2.0479	1.6998
42	0.0868	0.0800	1.9312	1.9117	1.7650
43	0.1041	0.1067	2.1396	2.1464	1.8701
44	0.1025	0.1040	2.1218	2.1163	1.4204
45	0.0799	0.0881	1.8383	1.9271	1.0760
46	0.0893	0.0910	1.9631	1.9629	1.3442
47	0.1150	0.1300	2.2588	2.3849	1.9663
48	0.0828	0.0849	1.8783	1.8863	1.4242
49	0.0629	0.0644	1.5838	1.5888	1.5698
50	0.0560	0.0561	1.4640	1.4484	0.4240
51	0.0591	0.0634	1.5190	1.5728	1.3027
52	0.0983	0.0992	2.0731	2.1839	2.0761
53	0.1389	0.1389	2.6856	2.6598	2.2104
54	0.0859	0.0860	1.9197	1.8994	1.4143
55	0.0503	0.0505	1.3567	1.3430	1.2305
56	0.0682	0.0682	1.6679	1.6487	1.3311
57	0.0496	0.0543	1.3437	1.4152	0.8327
58	0.0743	0.0743	1.7599	1.7403	1.4044
59	0.0485	0.0486	1.3215	1.3053	0.5912

Table 1: Continue

60	0.1116	0.1127	2.2222	2.2114	1.8897
61	0.1140	0.1202	2.2481	2.2894	1.6749
62	0.0905	0.0952	1.9776	2.0140	0.7595
63	0.0781	0.0781	1.8138	1.7941	1.2923
64	0.0796	0.0796	1.8352	1.8151	0.9739
65	0.0646	0.0666	1.6120	1.6245	1.1482
66	0.0552	0.0552	1.4502	1.4326	1.2967
67	0.0219	0.0224	0.5320	0.5344	0.6298
68	0.1058	0.1072	2.1595	2.1526	1.5495
69	0.0721	0.0792	1.7280	1.8091	1.0705
70	0.0533	0.0539	1.4143	1.4084	0.9978
71	0.0344	0.0344	0.9880	0.9729	0.6429
72	0.0323	0.0324	0.9280	0.9140	1.0534
73	0.0521	0.0526	1.3930	1.3831	1.4722
74	0.0594	0.0605	1.5254	1.5253	1.6465
75	0.1097	0.1098	2.2023	2.1806	1.8992

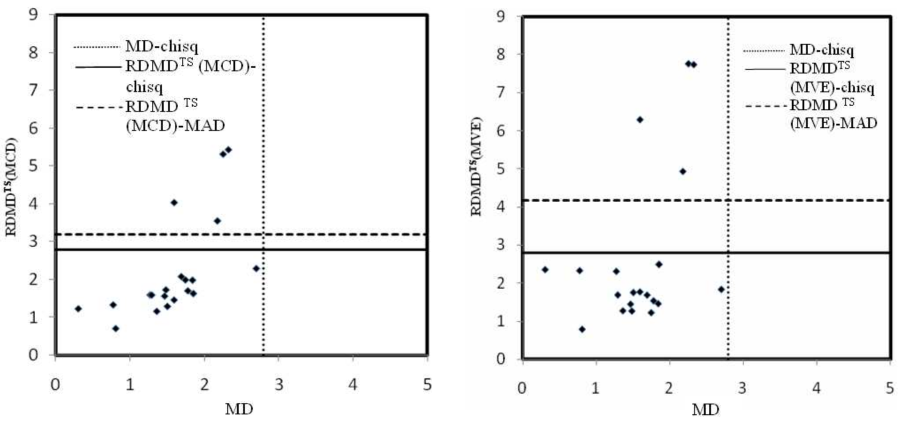
Table 2: Diagnostic robust generalized potential based on MVE and MCD and two-step robust diagnostics Mahalanobis distance based on MVE and MCD for stack loss data

			(RDMD^{TS}) (MVE)	(RDMD^{TS}) (MCD)	
	DRGP (MVE)	DRGP (MCD)	$\sqrt{\chi^2_{3,0.975}}$ = (2.796)	$\sqrt{\chi^2_{3,0.975}}$ = (2.796)	MD
Index	MAD (0.781)	MAD (1.063)	MAD (4.165)	MAD (3.199)	$\sqrt{\chi^2_{3,0.975}}$ = (2.796)
1	2.2214	2.4259	7.7595	5.3092	2.2536
2	2.3049	2.5304	7.7379	5.4260	2.3247
3	1.3005	1.4307	6.2906	4.0305	1.5937
4	0.2765	0.2871	2.3042	1.5883	1.2719
5	0.2133	0.2530	2.3501	1.2248	0.3034
6	0.2635	0.2877	2.3250	1.3259	0.7729
7	0.3944	0.4230	2.4873	1.6260	1.8527
8	0.3944	0.4230	2.4873	1.6260	1.8527
9	0.2229	0.2322	1.2689	1.1567	1.3606
10	0.4171	0.6825	1.2188	1.9861	1.7460
11	0.2489	0.3874	1.4417	1.5580	1.4657
12	0.4115	0.6765	1.4568	1.9797	1.8415
13	0.2314	0.3243	1.2625	1.7230	1.4826
14	0.2378	0.3175	1.5311	1.6991	1.7788
15	0.6137	0.7711	1.6796	2.0740	1.6902
16	0.3525	0.4016	1.6847	1.5860	1.2919
17	0.7604	1.0531	1.8307	2.2874	2.7000
18	0.2562	0.2738	1.7483	1.2868	1.5032
19	0.3213	0.3397	1.7646	1.4559	1.5932
20	0.0933	0.1179	0.7819	0.7014	0.8071
21	0.9128	1.1244	4.9305	3.5454	2.1768

(RDMD^{TS})(MCD), MD and their corresponding cutoff points. Another useful detection tool is proposed by Rousseeuw and Van Driessen[24] as DD plot. In this plot, the classical MD$_i$ is plotted vs. robust MD$_i$. The low leverage points should cluster below the cutoff point lines and the high leverage points will be separated from the bulk of the data and thus, will be located in the upper area of the cutoff points.

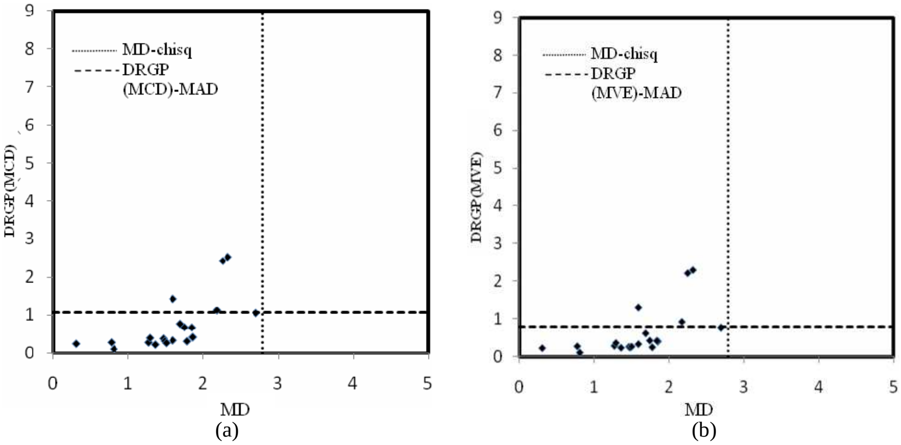
The DD plot of stack loss data set is shown in Fig. 1a (MD Vs RDMD^{TS} (MCD)), (b) (MD Vs RDMD^{TS} (MVE)) and Fig. 2a (MD Vs DRGP (MCD)) and 2b (MD Vs DRGP (MVE)). In both plot of Fig. 1, there are two cutoff point lines namely the Mad and the chi-square ($\sqrt{\chi^2_{3,0.975}}$), while there is only one cutoff point line (Mad) employed by DRGP in plot (a) and (b) of Fig. 2.

Simulation study: In order to investigate the merit of our newly proposed method, we designed a Monte Carlo simulation experiment. In this study, we compared the Robust Diagnostic Mahalanobis Distance (RDMD^{TS}) with other existed methods, with sample sizes equal to 20, 40, 60, 100 and 200. The first 100 (1-α) % observations of the three regressors from these sample sizes are produced from Uniform (0, 1) and the remaining 100α% observations are constructed as high leverage points. The high leverage points are generated with unequal weights,



(a) (b)

Fig. 1: (a): Mahalanobis distance against two-step robust diagnostic mahalanobis distance based on MCD, (b): Mahalanobis distance against two-step robust diagnostic mahalanobis distance based on MVE



(a) (b)

Fig. 2: (a): Mahalanobis distance against diagnostic robust generalized potential based on MCD, (b): Mahalanobis distance against diagnostic robust generalized potential based on MVE

where the last observations in each sample sizes are kept fixed at 10 value and the other high leverage points are the increments of five. We run 10000 simulations for these 5 different sample sizes. The results are illustrated in Table 3.

DISCUSSION

Let us focus our attention to the results of Hawkin's data presented in Table 1. The RMD (MCD) and the RMD (MVE) can detect 1-10 data as outliers. In addition to that RMD (MCD) identifies observations 11-14, 47 and 53 as outliers while RMD (MVE) swamp observations 11-14 and 53(not shown due to space limitations). Although these robust methods are more powerful than MD which can just detect 2 outliers, that is cases 12 and 14, they still can be improved so that their performance as high leverage detection tool is more powerful. As proposed in the second step of the (RDMD^{TS}), we should find the mean and covariance matrix of the clean data set for both RMD (MCD) and RMD (MVE) after deleting the suspected outlier group. Finally we can find the distance of the whole data set with this clean mean and clean covariance matrix for the x variables only. It is obvious from Table 1 that both of our proposed method and Habshah *et al.*[6] method can detect 14 high leverage points from both mad and chi-square cutoff points. However the values of (RDMD^{TS}) are further from their corresponding cutoff points compared to DRGP. Thus, this new method enhances the chance to detect these 14 observations as high leverage points.

Table 3: 10000 simulations for comparing RDMD^{TS} and DRGP based on (MCD) and (MVE)

			10000 simulations											
			RDMD^{TS} (MCD)				DRGP (MCD)		RDMD^{TS} (MVE)				DRGP (MVE)	
			Mad		Chi-sq		Mad		Mad		Chi-sq		Mad	
% HLP	n	2	1	2	1	2	1	2	1	2	1	2	1	2
5%	20	1	1.853	1	3.870	1	1.7520	1	0.351	1	0.465	1	0.805	1
	40	2	0.987	2	3.788	2	1.8000	2	0.138	2	0.221	2	0.625	2
	60	3	0.387	3	1.898	3	1.1000	3	0.076	3	0.135	3	0.533	3
	100	5	0.100	5	0.464	5	0.6270	5	0.034	5	0.076	5	0.461	5
	200	10	0.013	10	0.077	10	0.4790	10	0.008	10	0.039	10	0.453	10
10%	20	2	1.351	2	3.183	2	1.2370	2	0.213	2	0.329	2	0.595	2
	40	4	0.553	4	2.909	4	1.1120	4	0.082	4	0.178	4	0.405	4
	60	6	0.199	6	1.443	6	0.6600	6	0.036	6	0.120	6	0.316	6
	100	10	0.037	10	0.333	10	0.3170	10	0.012	10	0.069	10	0.239	10
	200	20	0.002	20	0.064	20	0.1950	20	0.002	20	0.039	20	0.177	20
15%	20	3	0.945	3	2.603	3	0.8450	3	0.119	3	0.222	3	0.416	3
	40	6	0.283	6	2.203	6	0.6590	6	0.046	6	0.151	6	0.254	6
	60	9	0.078	9	0.993	9	o.3305	9	0.016	9	0.092	9	0.164	9
	100	15	0.010	15	0.238	15	0.1330	15	0.003	15	0.063	15	0.103	15
	200	30	0.000	30	0.054	30	0.0630	30	0.000	30	0.039	30	0.058	30
20%	20	4	0.552	4	1.992	4	0.4610	4	0.585	4	0.132	4	0.270	4
	40	8	0.121	8	1.570	8	0.3280	8	0.016	8	0.111	8	0.127	8
	60	12	0.026	12	0.718	12	0.1400	12	0.007	12	0.086	12	0.076	12
	100	20	0.003	20	0.187	20	0.0500	20	0.001	20	0.063	20	0.041	20
	200	40	0.000	40	0.047	40	0.0120	40	0.000	40	0.035	40	0.011	40
25%	20	5	0.283	5	1.397	5	0.2340	5	0.028	5	0.070	5	0.188	5
	40	10	0.043	10	1.069	10	0.1410	10	0.007	10	0.087	10	0.065	10
	60	15	0.005	15	0.486	15	0.0570	15	0.002	15	0.068	15	0.032	15
	100	25	0.000	25	0.139	25	0.0130	25	0.000	25	0.057	25	0.011	25
	200	50	0.000	50	0.047	50	0.0020	50	0.000	50	0.039	50	0.002	50

1#: LLP = Low Leverage Points, 2#: HLP = High Leverage Points, where # denotes cardinality

Let us now focus to the Stack loss data where the RMD (MVE) can detect 4 outliers and another outlier which is case 2. Furthermore, RMD (MCD) can detect 4 outliers and cases 2, 13, 14, 20 as outliers as well. The RMD (MVE) and RMD (MCD) are not presented due to space constraint. After deleting the outliers from the data set and utilizing the mean and covariance matrix from the cleaned data set in the first step, the (RDMD^{TS}) can identify exactly 4 high leverage points. The DRGP (MCD) and DRGP (MVE) of Table 2 also can identify these 4 high leverage points. Like the results of Hawkin's Data, similar conclusion can be drawn from this example regarding higher chances of (RDMD^{TS}) for detection of high leverage points. The results of Table 2 show that (RDMD^{TS}) can detect these 4 high leverage points easily. Due to MD masking problem, it cannot detect any high leverage points.

By looking at Fig. 1 and 2, it is obvious that MD couldn't identify any high leverage points while the other 4 robust methods, can identify 4 high leverages easily.

Next, we will discuss the simulation results whether they confirm the conclusion of the numerical examples that our proposed method performs better than the DRGP and MD method. It can be observed from Table 3 that for small sample size, the (RDMD^{TS}) based on MCD or MVE with chi-square cutoff points swamp more low leverage points compared to (RDMD^{TS}) based on MCD or MVE with mad cutoff points. Nevertheless as soon as the number of sample sizes increases this cutoff point performs better and with this cutoff point we can find less low leverage but still it shows more low leverage than (RDMD^{TS})-mad. It is obvious from the results of Table 3 that (RDMD^{TS}) (MVE)-mad outperforms (RDMD^{TS}) (MCD)-mad in identifying less low leverages in small sample sizes.

In large sample sizes such as 200 (with 20 or 25% high leverage points) both of these two methods (RDMD^{TS}) (MVE)-mad and (RDMD^{TS}) (MCD)-mad are equally good and do credible job in detecting high leverage points. To compare (RDMD^{TS})-mad and DRGP based on MCD or MVE, we can say that the number of low leverage points which is identified are less when our newly proposed methods are used. When the sample size are 100 or 200 and 20 or 25% high leverage points are added, (RDMD^{TS})-mad can detect the exact high leverage points with no low leverage points while DRGP swamps some low leverages. When the number of sample size and high leverage points are very small, DRGP swamp less low leverage points compared to (RDMDTS)-mad (20 sample size and 5% high leverage points). When the number of high leverage points and the number of sample size increases, the ($RDMD^{TS}$)-mad overcome DRGP in detecting less low leverages.

CONCLUSION

The presence of high leverage points affects all least squares models, which are extensively used in data exploration and modeling. In multivariate cases the identification of high leverage points is much more difficult. Furthermore it is difficult to detect outliers in p-variate data when $p>2$, as one can no longer rely on visual inspection. Among all outlier detection tools, Mahalonobis Distance is more powerful to detect a single outlier. This approach is not applicable for multiple outliers because of the masking effect, by which multiple outliers do not necessarily have large Mahalonobis distance value. It is better to use distances based on robust estimators of multivariate location and scatter [23]. In regression analysis, the robust distances are computed from the explanatory variables which allow us to detect high leverage points. The main insight behind this study is to introduce a two step robust diagnostic methods based on Robust Mahalanobis distance. This relatively new method not only can detect exactly the high leverage points but also it can identify less number of low leverage points than the existing methods such as Diagnostic Robust Generalized Potential. To investigate the superiority of our new method, a Monte Carlo simulation is carried out. The results of this study indicate that for small sample sizes, the best detection method is ($RDMD^{TS}$) (MVE)-mad whereas there is not much difference between ($RDMD^{TS}$) (MVE)-mad and ($RDMD^{TS}$) (MCD)-mad for large sample sizes. Therefore, when the sample size is very small such as 20 and the number of high leverage is 5% of the data set, it is better to use DRGP (MVE) which can detect less low leverage points.

REFERENCES

1. Atkinson, A., 1994. Fast very robust methods for the detection of multiple outliers. J. Am. Stat. Assoc., 89: 1329-1339. http://www.questia.com/googleScholar.qst;jsessionid=JGJQQCDjZvh1FQwSntXskXM3FdSQwSJ1TzgzJLcg0vPhTsyynj2p!-1982163256!-953400118?docId=5002222122
2. Barnett, V. and T. Lewis, 1994. Outliers in Statistical Data. 3rd Edn., Wiley, New York, USA., ISBN: 10: 0471930946, pp: 604.
3. Brownlee, K.A., 1984. Statistical Theory and Methodology in science and Engineering. 2nd Edn., Krieger Pub Co., USA., ISBN: 10: 0898747481, pp: 590.
4. Chatterjee, S. and A.S. Hadi, 2006. Regression Analysis by Example. 4th Edn., Wiley, New York, USA., ISBN: 10: 0-471-74696-7, pp: 375.
5. Gnanadesikan, R. and J. Kettenring, 1972. Robust estimates, residuals and outlier detection with multiresponse data. Biometrics, 28: 81-124. http://links.jstor.org/sici?sici=0006-341X(197203)28%3A1%3C81%3ARERAOD%3E2.0.CO%3B2-W
6. Habshah, M., M.R. Norazan and A.H.M.R. Imon, 2008. The performance of diagnostic-robust generalized potentials for the identification of multiple high leverage points in linear regression. J. Applied Stat. (In press).
7. Hadi, A.S., 1992. A new measure of overall potential influence in linear regression. Comput. Stat. Data Anal., 14: 1-27. DOI: 10. 1016/0167-9473(92)90078-T
8. Hawkins, D.M., 1980. Identification of Outliers. 1st Edn., Springer, Londen, UK., ISBN: 10: 041221900X, pp: 188.
9. Hawkins, D.M., D. Bradu and G.V. Kass, 1984. Location of several outliers in multiple regression data using elemental sets. Technometrics, 26: 197-208. DOI: 10.2307/1267545
10. He, X., 1991. A local breakdown property of robust tests in linear regression. J. Multivariate Anal., 38: 294-305. DOI: 10.1016/0047-259X(91)90047-6
11. Hoaglin, D.C. and R.E. Welsch, 1978. The hat matrix in regression and ANOVA. Am. Stat. Assoc., 32: 17-22. http://ideas.repec.org/p/mit/sloanp/1920.html
12. Huber, P.J., 1981. Robust statistics. Wiley-Interscience, New York, USA., ISBN: 10: 0471418056, pp: 320.
13. Imon, A.H.M.R. and M.A.I. Khan, 2003. A solution to the problem of multcollinearity caused by the presence of multiple high leverage points. Int. J. Stat. Sci., 2: 37-50. http://ijss-ru0.tripod.com/abstract_vol2_2003.pdf.com
14. Imon, A.H.M.R., 2005. Identifying multiple influential observations in linear regression. J. Applied Stat., 32: 929-946. DOI: 10.1080/02664760500163599

15. Jureckova, J. and S. Portnoy, 1987. Asymptotic for one-step M-estimators in regression with application to combining efficiency and high breakdown point. Commun. Stat. Theor. Methods, 16: 2187-2199. DOI: 10.1080/03610928708829500
16. Mardia, K., J. Kent and J. Bibby, 1979. Multivariate Analysis. Academic Press, USA., ISBN: 10: 0124712525.
17. Maronna, R.A., 1976. Robust M-estimators of multivariate location and scatter. Ann. Stat., 4: 51-67. http://www.projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.aos/1176343347
18. Maronna, R.A. and V.J. Yohai, 1998. Robust Estimation of Multivariate Location and Scatter. In: Encyclopedia of Statistical Sciences, Kotz, S., C. Read and D. Banks (Eds.). Wiley-Interscience, New York, ISBN: 10: 0471118362, pp: 589-596.
19. Moller, S.F., J.V. Frese and R. Bro, 2005. Robust Methods for multivariate data analysis. J. Chemometr., 19: 549-563. DOI: 10.1002/cem.962
20. Pearson, E. and C. Chandra Sekar, 1936. The efficiency of statistical tools and a criterion for the rejection of outlying observations. Biometrika, 28: 308-320. http://biomet.oxfordjournals.org/cgi/content/citation/28/3-4/308
21. Rocke, D. and D. Woodruff, 2008. Computation of robust estimates of multivariate location and shape. Stat. Neerland., 47: 27-42. DOI: 10.1111/j.1467-9574.1993.tb01404.x
22. Rousseeuw, P.J., 1985. Multivariate estimation with high breakdown point. Math. Stat. Appli., B: 283-297. http://www.ams.org/mathscinet-getitem?mr=851060
23. Rousseeuw, P.J. and A. Leory, 1987. Robust Regression and Outlier Detection. 1st Edn., Wiley, New York, USA., ISBN: 10: 0471852333, pp: 352.
24. Rousseeuw, P. and K. Van Driessen, 1999. A fast algorithm for the minimum variance determinant estimator. Technometrics, 41: 212-223. DOI: 10.2307/1270566
25. Rousseeuw, P. and B. Van Zomeren, 1990. Unmasking multivariate outliers and leverage points. J. Am. Stat. Assoc., 85: 633-639. http://cat.inist.fr/?aModele=afficheN&cpsidt=19306453
26. Ryan, T., 1997. Modern Regression Methods. Har/Dis Edn., Wiley, New York, USA., ISBN: 10: 0471529125, pp: 515.
27. Vellman, P.F. and R.E. Welsch, 1981. Efficient computing of regression diagnostics. Am. Stat., 27: 234-242. http://www.jstor.org/pss/2683296
28. Werner, M., 2003. Identification of multivariate outliers in large data sets. PhD Thesis, University of Colorado at Denver, pp: 241. http://www-math.cudenver.edu/graduate/thesis/werner_thesis.pdf
29. Yohai, V.J., 1987. High breakdown point and high efficiency robust estimates for regression. Ann. Stat., 15: 642-656. DOI: 10.1214/aos/1176350366