Malaysian Journal of Mathematical Sciences 10(3): 283–295 (2016)

MALAYSIAN JOURNAL OF MATHEMATICAL SCIENCES

Journal homepage: http://einspem.upm.edu.my/journal

Detection of High Leverage Points Using a Nonparametric Cut-off Point for the Robust Mahalanobis Distance

A. H. M. Rahmatullah Imon *[1] and M. Rasheduzzaman Apu[2]

[1] *Department of Mathematical Sciences, Ball State University, USA*

[2] *Mercantile Bank Ltd., Bangladesh*

E-mail: imon_ru@yahoo.com

**Corresponding author*

ABSTRACT

Mahalanobis distance (MD) is a widely used multivariate technique for measuring dispersion. Rousseuw and Leroy (1987) advocated using MD as a measure of high leverage points in linear regression. Since MD's are non-robust in the presence of high leverage points they suggested using robust version of Mahalanobis distance. They also proposed a cut-off point for MD's which follows a square root of Chi-square distribution with the degrees of freedom equals to the dimension of the explanatory variables. But we see a major problem in it. In regression we do not assume normality assumption for the explanatory variables, sometimes the explanatory variables may be indicator or categorical variables. Moreover, the explanatory variables are treated as fixed variables hence a chi-square cut-off point is not appropriate. In this paper we propose a nonparametric cut-off point for the robust Mahalanobis distance. This cut-off point does not require any distributional assumption of the explanatory variables. We employ this method to several well-known data sets and observed that the proposed method performs much better than the existing methods.

Keywords: Leverage, Mahalanobis distance, MVE, Swamping.

1. Introduction

Leverage values are being used in regression diagnostics as measures of influential observations in the **X**-space. Detection of high leverage observations or points is crucial due to their responsibility for masking outliers. Let us consider a k variable regression model

$$Y = X\beta + \epsilon \tag{1}$$

The least squares residual vector can be expressed in terms of the true disturbance vector as

$$\hat{\epsilon} = Y - \hat{Y} = (I - W)Y = (I - W)\epsilon \tag{2}$$

where the matrix

$$W = X(X^T X)^{-1} X^T \tag{3}$$

is generally known as weight matrix or leverage matrix. Observations corresponding to excessively large ϵ values are termed as outliers and the values of X which stand far away from the centre of the data are called high leverage points. According to Hocking and Pendleton (1983), *"high leverage points are those for which the input vector x_i, in some sense, far from the rest of the data."* The i-th diagonal element, w_{ii} of the weight matrix W is traditionally used as a measure of leverage of the response value y_i on the corresponding value $\hat{y}_i$ since the weight matrix W reflects joint effect of k regressors on the fitted responses. The average value of w_{ii} is k/n, where k is the number of the regressors (including the intercept term) and n is the total number of observations. Data points having large w_{ii} values are generally considered as high leverage points. Hoaglin and Welsch (1978) considered observations unusual when w_{ii} exceeded $2k/n$ which is also known as *twice-the-mean-rule*. Vellman and Welsch (1981) consider w_{ii} as large when it exceeds $3k/n$ which is known as *thrice-the-mean-rule*. For a definition of when w_{ii} is large, Huber (1981) suggested breaking the range of possible values, $(0 \leq w_{ii} \leq 1)$ into three intervals. Values $w_{ii} \leq 0.2$ appear to be safe, values between 0.2 and 0.5 are risky, and values above 0.5 should be avoided. All these cut-off points are rules of thumb and they have been proposed in the literature in this way mainly because of the fact that the explanatory variables are held fixed in the least squares approach.

Well known Mahalanobis distances are also suggested to use as measures of leverages in the literature. For the mean vector $\bar{X}$ and the variance covariance matrix S, Mahalanobis (1936) defined a multivariate dispersion matrix

$$M = \sqrt{(X - \bar{X}) S^{-1} (X - \bar{X})^T} \tag{4}$$

The diagonal elements of the M matrix as defined in (4) are known as Mahalanobis distances. Here we define the ith Mahalanobis distance

$$d_i = \sqrt{a_{ii}} \tag{5}$$

where $\sqrt{a_{ii}}$ is the ith diagonal element of M. Mahalanobis (1936) also showed under normality

$$M^2 = (X - \bar{X})S^{-1}(X - \bar{X})^T \sim \chi_p^2 \tag{6}$$

In a linear regression problem if X is the set of explanatory variables a very simple and common approach to consider observations to be points of high leverage, if they possess large Mahalanobis distance. Rousseeuw and Leroy (1987) show that

$$d_i = \sqrt{(n-1)(w_{ii} - 1/n)} \tag{7}$$

and thus Mahalanobis distance for each of the points has a one-one relationship with w_{ii}. They also pointed out that Mahalanobis distance consists of mean vector and variance covariance matrix both of them can be largely affected by high leverage points and may break down easily. For this reason a robust version of Mahalanobis distance is required. In section 2 we introduce the robust Mahalanobis distance. In section 3 we introduce leverage measures based on robust Mahalanobis distance (RMD). We believe that the existing cut-off point for RMD based on chi-square distribution is logically incorrect and for this reason we propose a new nonparametric cut-off point for the detection of high leverage points. In section 4 we presented a couple of well-known examples to show the advantage of using the proposed cut-off point.

2. Robust Mahalanobis Distance For the Detection of High Leverage Points

Here we introduce a robust version of Mahalanobis distances based on robust estimators for mean vector and variance-covariance matrix. Let us define a general form of Mahalanobis Squared Distance (MSD)

$$MD_i^2 = [x_i - T(X)][C(X)]^{-1}[x_i - T(X)]^T \tag{8}$$

Rousseeuw (1984) suggested a method where $T(X)$ is the center of the minimal volume ellipsoid covering at least h points of X, where h can be taken equal to $(n/2) + 1$. This is called the minimum volume ellipsoid (MVE) estimator. The corresponding covariance estimator is given by the ellipsoid itself, multiplied by a suitable factor to obtain consistency.

In most applications it is not feasible to consider all 'halves' of data. We start by drawing a sub sample of $(p + 1)$ different observations, indexed by

$J = (i_1, i_2, ..., i_{p+1})$. For this sub sample we determine the arithmetic mean and the corresponding covariance matrix, given by

$$\bar{X}_J = \frac{1}{p+1} \sum_{i \in J} x_i, c_J = \frac{1}{p} \sum_{i \in J} (x_i - \bar{x}_J)^T (x_i - \bar{x}_J) \tag{9}$$

where c_J is non-singular. The corresponding ellipsoid should then be inflated or deflated to contain exactly h points, which corresponds to computing

$$m_J^2 = med(x_i - \bar{x}_J) c_J^{-1} (x_i - \bar{x}_J)^T \tag{10}$$

The volume of the resulting ellipsoid, corresponding to $m_J^2 C_J$ is proportional to

$$[det(m_J^2 C_J)]^{1/2} = [det(C_J)]^{1/2} (m_J)^p \tag{11}$$

We repeat it for many J so that the above determinant becomes the minimum and its corresponding values yield $T(X) = \bar{x_J}$ and $C(X) = (\chi^2_{p,0.5})^{-1} m_J^2 C_J$ where $\chi^2_{p,0.5}$ is the median of the chi-squared distribution with p degrees of freedom. This correction factor is used because we assume that the data follow a multivariate normal distribution.

Another popular approach for finding robust estimates for mean vector and the variance covariance matrix where $T(X)$ is the mean of the h points of X for which the determinant of the covariance matrix is minimal. We call this method the minimum covariance determinants (MCD) suggested by Rousseeuw (1984). The standard choice of $h = [(n + p + 1)/2]$ for the MCD is proposed by Van Alest and Rousseeuw (2009). Conceptually MVE is related with the least median squares (LMS) and MCD is related with least trimmed of squares (LTS) regression. They are computationally extensive, but fortunately both the MVE and MCD are readily available in S-PLUS and R.

After obtaining the robust multivariate location and scale estimates given by either MVE or MCD, we compute the robust Mahalanobis distance

$$Md_i = \sqrt{[x_i - T(X)][C(X)]^{-1}[x_i - T(X)]^T} \tag{12}$$

Rousseeuw and Leroy (1987) suggested a cut-off point for Md_i as $\sqrt{\chi^2_{p,0.5}}$ at the 5% level of significance.

3. A New Cut-Off Point for the Robust Mahalanobis Distance

When we use the ordinary least squares (OLS) method for fitting a regression line, the resulting residuals are functions of leverages and true errors. Thus high leverage points together with large errors (outliers) may pull the fitted least squares line in a way that the fitted residuals corresponding to that outliers might be too small and this may cause masking (false negative) of outliers. For the same reason the residuals corresponding to inliers may be too large and this may cause swamping (false positive). Robust methods are proved to be useful to identify the genuine high leverage points, but they have a general tendency [see Cook and Hawkins (1990)] to declare innocent observations to be unusual. This effect is known as swamping and Davies et al. (2004) showed that swamping can severely affect the estimation of parameters. So we definitely want a high leverage detection method which is not affected by masking or swamping. But the other issue is more crucial here. The cut-off value $\chi^2_{p,0.5}$ for the robust Mahalanobis distance as shown in (8) is logically grounded from the assumption that the p-dimensional variables follow a multivariate normal distribution. But in a regression problem it is a common assumption that the explanatory variables are held as fixed and for this reason no probability distribution is assumed for them. Mainly because of this fact the traditionally used measures of leverages such as twice-the-mean rule, thrice-the-mean rule and Huber's rule as mentioned in section 2 yield thumb rules for cut-off points of w_{ii}. No body even thought about the distributional properties of w_{ii} because they are fixed as the X matrix is fixed. Hence assumption regarding the X matrix to follow a multivariate normal distribution is logically incorrect and hence the resulting Mahalanobis distance should not be associated with a chi-square distribution. At this point the immediate question comes to our mind is what should be an appropriate cut-off point for Mahalanobis distance in a regression problem. Here we propose a new cut-off point for Mahalanobis distances and/or robust Mahalanobis distances. This cut-off point is purely empirical in nature and it does not require any distributional assumption regarding the distance values Md_i. For a set of Md_i values, we call an observation as a high leverage point if it satisfies the condition

$$Md_i > Median(Md_i) + 3MAD(Md_i) \tag{13}$$

where $MAD(Md_i)$ is defined as

$$MAD(Md_i) = Median|Md_i - median(Md_i)|/0.6745 \tag{14}$$

This nonparametric type cut-off point was first proposed by Hadi (1992) and later used by Imon (2002) and many others.

4. Examples

In this section we compare the performance of our proposed cut-off rule with the existing rule. We have considered two well-known data sets Brown (1980) data and Finney (1947) data.

4.1 Brown Data

We first consider a cancer data set given by Brown (1980). The original objective of the author was to see whether an elevated level of acid phosphates (A.P.) in the blood serum would be of value for predicting whether or not prostate cancer patients also had lymph node involvement (L.N.I). The data set additionally contains data on the four more commonly used regressors, but we use here only two variables A.P. and the 'Age' of the patients in illustrating simple logistic regression with 53 cases. The observations from 53 patients are given in Table 1.

Table 1: Brown cancer data

Index	A.P	Age	Index	A.P	Age	Index	A.P	Age	Index	A.P	Age
1	48	66	15	47	67	29	50	64	43	81	50
2	56	68	16	49	51	30	40	63	44	76	60
3	50	66	17	50	56	31	55	52	45	70	45
4	52	56	18	78	60	32	59	66	46	78	56
5	50	58	19	83	52	33	48	58	47	70	46
6	49	60	20	98	56	34	51	57	48	67	67
7	46	65	21	52	67	35	49	65	49	82	63
8	62	60	22	75	63	36	48	65	50	67	57
9	56	50	23	99	59	37	63	59	51	72	51
10	55	49	24	187	64	38	102	61	52	89	64
11	62	61	25	136	61	39	76	53	53	126	68
12	71	58	26	82	56	40	95	67			
13	65	51	27	40	64	41	66	53			
14	67	67	28	50	61	42	84	65			

This data set has been analyzed extensively by many authors [see Imon and Hadi (2008, 2013)] and it has been reported that 3 observations (cases 22, 23 and 53) are genuine high leverage points in this data. The scatterplot as shown in Figure 1 also supports their findings.

Now we compute robust Mahalanobis distance for this data set based on MVE, MCD and the results with the existing and proposed cut-off points are presented in Table 2. If we look at the MVE and MCD values we clearly see that they are very different. MVE values are on average much higher than the

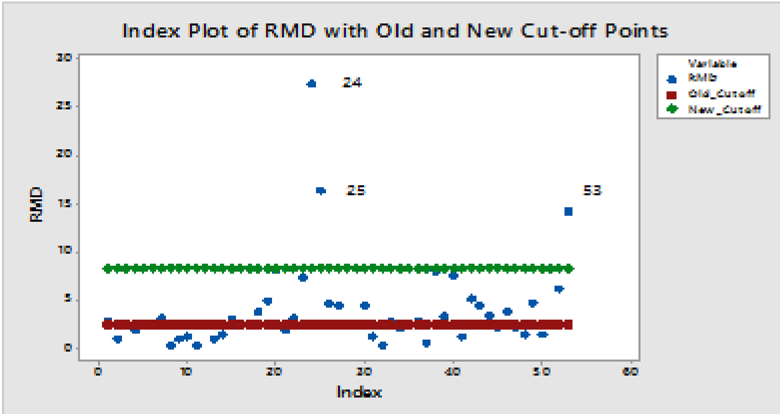

Figure 1: Scatter plot of Age vs A.P. for Brown data

corresponding MCD values and the former one has much higher dispersion as well since the average MVE and MCD are 3.868 and 2.061 respectively with their respective standard deviations 4.488 and 2.005. But we are surprised to see that the existing method uses the same cut-off points for both of these two methods. Results presented in Table 2 clearly shows that how ineffective is the existing method to identify the high leverage points. There are only 3 high leverage points in the data but the existing cut-off point based on MVE identifies 29 observations out of 53 as high leverage points. It is simply absurd. The performance of MCD is slightly better although it identifies 10 observations as high leverage points. When we employ our proposed method it correctly identifies the 3 observations and no more points are swamped in as high leverage points.

Similar remarks may apply with Figure 2 where we present an index plot of robust Mahalanobis distances with the existing and proposed cut-off points. This figure shows that the proposed method can locate the three high leverage points clearly but the existing methods identify more than half of the observations as high leverage points.

4.2 Finney Data

Next we consider another cancer data set given by Finney (1947). The original data set were obtained to study the effect of the rate and volume of air inspired on a transient vaso-constriction in the skin of the digits. The nature of the measurement process was such that only the occurrence and nonoccurrence

Table 2: Robust Mahalanobis distances for Brown data

Index	Mahalanobis Distance		Identified Cases Cut-off (2.447)		Identifications based on Proposed Cut-off Points	
	MVE	MCD	MVE	MCD	MVE(8.3487)	MCD(3.8347)
1	2.7145	1.2516	1	0	0	0
2	0.9804	1.4629	0	0	0	0
3	2.2810	1.1778	0	0	0	0
4	1.8474	0.9481	0	0	0	0
5	2.2810	0.9098	0	0	0	0
6	2.4977	0.8847	1	0	0	0
7	3.1480	1.2646	1	0	0	0
8	0.3202	0.4143	0	0	0	0
9	0.9804	1.6118	0	0	0	0
10	1.1971	1.8293	0	0	0	0
11	0.3202	0.5533	0	0	0	0
12	2.2711	1.1017	0	0	0	0
13	0.9705	1.2124	0	0	0	0
14	1.4041	1.8217	0	0	0	0
15	2.9313	1.4039	1	0	0	0
16	2.4977	1.8932	1	0	0	0
17	2.2810	1.1203	0	0	0	0
18	3.7885	1.9001	1	0	0	0
19	4.8723	2.1225	1	0	0	0
20	8.1238	3.5561	1	1	0	0
21	1.8474	1.2769	0	0	0	0
22	3.1382	1.9166	1	0	0	0
23	7.2346	3.8214	1	1	0	0
24	27.4161	12.5386	1	1	1	1
25	16.3610	7.4766	1	1	1	1
26	4.6556	2.0554	1	0	0	0
27	4.4486	1.6658	1	0	0	0
28	2.2810	0.7824	0	0	0	0
29	2.2810	0.9524	0	0	0	0
30	4.4486	1.6567	1	0	0	0
31	1.1971	1.3364	0	0	0	0
32	0.3301	1.2145	0	0	0	0
33	2.7145	1.0984	1	0	0	0
34	2.0642	0.9156	0	0	0	0
35	2.4977	1.1008	1	0	0	0
36	2.7145	1.1502	1	0	0	0
37	0.5370	0.4010	0	0	0	0
38	7.9909	3.7513	1	1	0	0
39	3.3550	1.5339	1	0	0	0
40	7.4735	3.7647	1	1	0	0
41	1.1873	0.9464	0	0	0	0
42	5.0891	2.9502	1	1	0	0
43	4.4388	2.0436	1	0	0	0
44	3.3550	1.7111	1	0	0	0
45	2.0544	2.1464	0	0	0	0
46	3.7885	1.6858	1	0	0	0
47	2.0544	1.9976	0	0	0	0
48	1.4041	1.8217	0	0	0	0
49	4.6556	2.5524	1	1	0	0
50	1.4041	0.7067	0	0	0	0
51	2.4879	1.4139	1	0	0	0
52	6.1729	3.2989	1	1	0	0
53	14.1933	7.1220	1	1	1	1

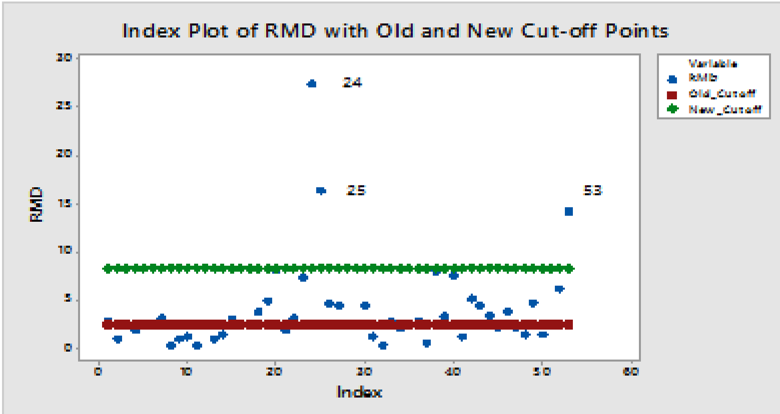

Figure 2: Index plot of robust Mahalanobis distances for Brown data

of vaso-constriction could be reliably measured and this data is presented in Table 3.

Table 3: Finney Cancer Data

Index	Volume	Rate	Index	Volume	Rate	Index	Volume	Rate
1	3.70	0.825	14	1.40	2.330	27	1.80	1.500
2	3.50	1.090	15	0.75	3.750	28	0.95	1.900
3	1.25	2.500	16	2.30	1.640	29	1.90	0.950
4	0.75	1.500	17	3.20	1.600	30	1.60	0.400
5	0.80	3.200	18	0.85	1.415	31	2.70	0.750
6	0.70	3.500	19	1.70	1.060	32	2.35	0.030
7	0.60	0.750	20	1.80	1.800	33	1.10	1.830
8	1.10	1.700	21	0.40	2.000	34	1.10	2.200
9	0.90	0.750	22	0.95	1.360	35	1.20	2.000
10	0.90	0.450	23	1.35	1.350	36	0.80	3.330
11	0.80	0.570	24	1.50	1.360	37	0.95	1.900
12	0.55	2.750	25	1.60	1.780	38	0.75	1.900
13	0.60	3.000	26	0.60	1.500	39	1.30	1.625

Finney data set has also been extensively analyzed by many authors [see Imon and Hadi (2009, 2013)] and it has been reported that 3 observations (cases 1, 2 and 17) are genuine high leverage points. This kind of findings is supported by Figure 3 as well.

When we compute robust Mahalanobis distances by the MVE method AND the MCD as shown in Table 4, we observe that the existing cut-off point can

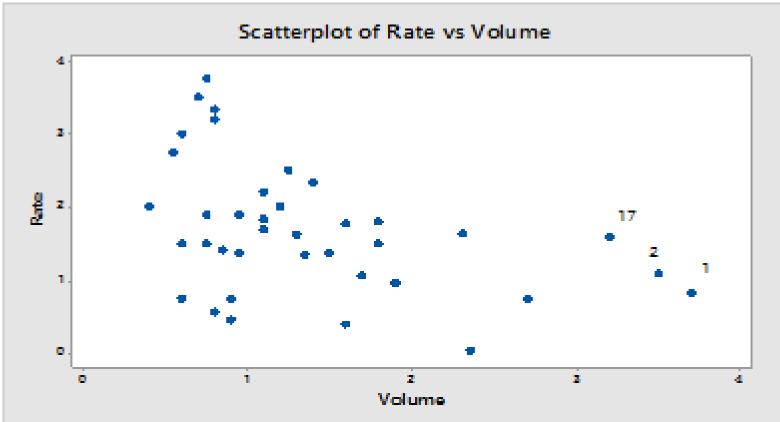

Figure 3: Scatter plot of Rate vs Volume for Finney data

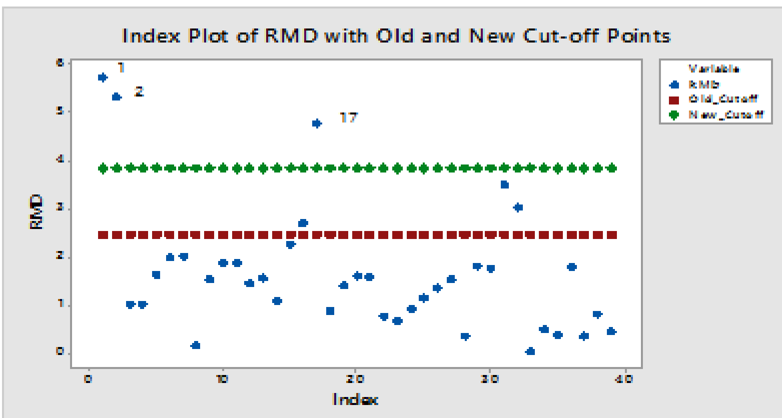

Figure 4: Index plot of robust Mahalanobis distances for Finney data

successfully identify the 3 genuine cases but at the same time both of them swamp in 3 good cases (16, 31 and 32) as well. But the proposed method correctly identifies the 3 observations without swamp in any good case.

We observe exactly same picture when we look at Figure 4 where we present an index plot of robust Mahalanobis distances with the existing and proposed cut-off points. This figure shows that the proposed method can locate the three high leverage points clearly but the existing methods identify 3 more observations as high leverage points.

Table 4: Robust Mahalanobis distances based on MVE for Finney data

Index	Mahalanobis Distance		Identified Cases Cut-off (2.447)		Identifications based on Proposed Cut-off Points	
	MVE	MCD	MVE	MCD	MVE(8.827)	MCD(4.178)
1	5.73219	4.79209	1	1	1	1
2	5.32267	4.44493	1	1	1	1
3	1.00416	1.22161	0	0	0	0
4	1.01001	0.95435	0	0	0	0
5	1.63098	2.04027	0	0	0	0
6	1.99275	2.44218	0	0	0	0
7	2.00697	1.98295	0	0	0	0
8	0.14051	0.14630	0	0	0	0
9	1.52233	1.55587	0	0	0	0
10	1.87784	1.95751	0	0	0	0
11	1.87232	1.91859	0	0	0	0
12	1.43250	1.63277	0	0	0	0
13	1.55923	1.85866	0	0	0	0
14	1.06975	1.15026	0	0	0	0
15	2.26533	2.77222	0	0	0	0
16	2.70194	2.22844	1	1	0	0
17	4.76606	4.00749	1	1	1	1
18	0.87594	0.83690	0	0	0	0
19	1.39284	1.11870	0	0	0	0
20	1.60071	1.31336	0	0	0	0
21	1.57406	1.47117	0	0	0	0
22	0.74977	0.71942	0	0	0	0
23	0.66168	0.46962	0	0	0	0
24	0.90508	0.65193	0	0	0	0
25	1.13112	0.90583	0	0	0	0
26	1.33950	1.24734	0	0	0	0
27	1.52808	1.18909	0	0	0	0
28	0.34206	0.46677	0	0	0	0
29	1.80745	1.47418	0	0	0	0
30	1.75091	1.71090	0	0	0	0
31	3.50089	2.89976	1	1	0	0
32	3.01687	2.71063	1	1	0	0
33	0.02376	0.23051	0	0	0	0
34	0.47515	0.71276	0	0	0	0
35	0.36568	0.48214	0	0	0	0
36	1.77791	2.21215	0	0	0	0
37	0.34206	0.46677	0	0	0	0
38	0.79492	0.81042	0	0	0	0
39	0.43467	0.23288	0	0	0	0

5. Conclusion

In this paper our main objective was to propose a new cut-off point for robust Mahalanobis distance to identify multiple high leverage points because the existing cut-off point logically looks incorrect. Another drawback of the existing method is that it often severely gets affected by swamping and identifies too many cases unnecessarily. Our proposed method is based on nonparametric approach and empirical in nature. So it does not require any table and is very easy to compute. A couple of well-known data sets clearly show the advantage of using the proposed cut-off point instead of the existing one. No matter whether we compute robust Mahalanobis distances by the MVE or the MCD when we employ the existing cut-off points for leverage measures it swamps a huge number of observations. But when we employ the proposed cut-off point it successfully identifies all high leverage points and does not swamp even a single observation.

References

Brown, B. W. Jr. 1980. Prediction analysis for binary data in Biostatistics Casebook, Miller, R. G. Jr., Efron, B., Brown, B. W. Jr. and Moses, L. E. (Eds.), New York: Wiley.

Cook, R.D., and Hawkins, D.M. 1990. Comment on Unmasking multivariate outliers and leverage points by Rousseeuw, P. J. and van Zomeren, B. C. Journal of the American Statistical Association, 85: 640-44.

Davies, P., Imon, A. H. M. R. and Ali, M. M. 2004. A conditional expectation method for improved residual estimation and outlier identification in linear regression, International Journal of Statistical Sciences, 3: 191-208.

Finney, D.J. 1947. The estimation from individual records of the relationship between dose and quantal response. Biometrika, 34: 320-334.

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

Hoaglin, D. C. and Welsch, R. E. 1978. The hat matrix in regression and ANOVA, The American Statistician, 32: 17-22.

Hocking, R. R. and Pendleton, O. J. 1983. The regression dilemma, Communications in Statistics- Theory Methods, 12: 497-527.

Huber, P. J. 1981. Robust Statistics, New York: Wiley.

Imon, A. H. M. R. 2002. Identifying multiple high leverage points in linear regression, Journal of Statistical Studies, Special volume in honor of Professor Mir Masoom Ali, 207-218.

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

Imon, A. H. M. R. 2006. Identification of high leverage points in logistic regression, Pakistan Journal of Statistics, 22 : 147-156.

Imon, A. H. M. R. and Hadi, A. S. 2008. Identification of multiple outliers in logistic regression, Communication in Statistics-Theory Methods, 37: 1967 -1709.

Imon, A.H.M.R. and Hadi, A.S. 2013. Identification of multiple high leverage points in logistic regression, Journal of Applied Statistics, 40: 2601-2616.

Mahalanobis, P. C. 1936. On the generalised distance in statistics, Proceedings of the National Institute of Sciences of India 2: 4955..

Nurunnabi, A. A. M., Imon, A. H. M. R. and Nasser, M. 2010. Identification of multiple influential observations in logistic regression, Journal of Applied Statistics, 37: 1605-1624.

Rousseeuw, P. J. 1984. Least median of squares regression, Journal of the American Statistical Association,79: 871-80.

Rousseeuw, P.J. and Leroy, A. M. 1987. Robust Regression and Outlier Detection, New York: Wiley.

Van Aelst, S., and Rousseeuw, P.J. 2009. Minimum volume ellipsoid, Wiley Interdisciplinary Reviews: Computational Statistics, 1, 71-82.

Velleman, P. F. and Welsch, R. E. 1981. Efficient computing in regression diagnostics, The American Statistician, 35: 234-242.