Journal of Applied Statistics

ISSN: 0266-4763 (Print) 1360-0532 (Online) Journal homepage: https://www.tandfonline.com/loi/cjas20

Identification of multiple high leverage points in logistic regression

A.H.M. Rahmatullah Imon & Ali S. Hadi

To cite this article: A.H.M. Rahmatullah Imon & Ali S. Hadi (2013) Identification of multiple high leverage points in logistic regression, Journal of Applied Statistics, 40:12, 2601-2616, DOI: 10.1080/02664763.2013.822057

To link to this article: https://doi.org/10.1080/02664763.2013.822057

Journal of Applied Statistics, 2013
Vol. 40, No. 12, 2601–2616, http://dx.doi.org/10.1080/02664763.2013.822057

Taylor & Francis
Taylor & Francis Group

Identification of multiple high leverage points in logistic regression

A.H.M. Rahmatullah Imon[a]* and Ali S. Hadi[b]

[a]*Department of Mathematical Sciences, Ball State University, Muncie, IN, USA;*
[b]*Department of Mathematics, The American University in Cairo, Cairo, Egypt*

(*Received 26 September 2012; accepted 1 July 2013*)

Leverage values are being used in regression diagnostics as measures of unusual observations in the X-space. Detection of high leverage observations or points is crucial due to their responsibility for masking outliers. In linear regression, high leverage points (HLP) are those that stand far apart from the center (mean) of the data and hence the most extreme points in the covariate space get the highest leverage. But Hosemer and Lemeshow [*Applied logistic regression*, Wiley, New York, 1980] pointed out that in logistic regression, the leverage measure contains a component which can make the leverage values of genuine HLP misleadingly very small and that creates problem in the correct identification of the cases. Attempts have been made to identify the HLP based on the median distances from the mean, but since they are designed for the identification of a single high leverage point they may not be very effective in the presence of multiple HLP due to their masking (false–negative) and swamping (false–positive) effects. In this paper we propose a new method for the identification of multiple HLP in logistic regression where the suspect cases are identified by a robust group deletion technique and they are confirmed using diagnostic techniques. The usefulness of the proposed method is then investigated through several well-known examples and a Monte Carlo simulation.

Keywords: logistic regression; covariates; high leverage points; masking; swamping; group deletion; robust regression; deletion median distance from the median, Monte Carlo simulation

1. Introduction

Regression diagnostics methods have become essential parts of logistic regression [2,19] in the recent years. In regression diagnostics the issue of outliers, influential observation and high leverage points (HLP) are discussed together. In linear regression an outlier is a point that deviates from the linear relationship determined from the other points, or at least from the majority of those points. Observations corresponding to exceptionally large residuals are termed outliers. Observations whose presence or absence can make a huge impact on the fitting of the model and hence the resulting analyses are called influential. HLP are unusual observations in the X-space. Chatterjee and Hadi [6] discussed the interrelationships among these three types of cases. Here, we only

*Corresponding author. Email: rimon@bsu.edu

note that influential observations need not be outliers in the sense of having large residuals. It is generally believed that outliers would be highly influential. But that is not always true. Andrews and Pregibon [1] have presented some examples where outlying observations have little influence on the results. Their examples illustrate the existence of outliers that do not matter. However, HLP are likely to be influential, but it is also not always the case. To quote Chatterjee and Hadi [6], 'As with outliers, high leverage points need not be influential, and influential observations are not necessarily high leverage points.'

A large body of literature is now available for the identification of multiple outliers [15] and multiple influential observations [18] in logistic regression, but not too much work has been done in the identification of multiple HLP in logistic regression. We often observe that HLP greatly affect the fitted values and consequently its presence might cause all kind of interpretative problems such as erroneous goodness-of-fit statistics, wrong odds ratios, wrong Wald statistics, etc. So we need to detect such observations and study their impact on the model. In Section 2, we introduce some commonly used diagnostics for the identification of HLP. In linear regression, the leverage values measure the distance of each observation from the mean and observations possessing excessive high leverage values are termed as HLP. But it is now evident that in logistic regression, the most extreme points in the covariate space may have the smallest leverage [13]. Thus, the commonly used leverage measures may fail to identify the HLP in logistic regression. As a remedy to this problem, Imon [15] proposed a method based on the median distances from the mean (MDM). These distances were defined in such a way that the factors that are responsible for yielding small leverages for genuine HLP are corrected. Although it is observed that the identification of a single HLP is often achieved satisfactorily by the use of the distances from the mean, these distances may be ineffective when a group of HLP's are present in the data. We anticipate that leverages or distances from the mean based on group deletions may produce better results in this situation. We introduce median deletion distance from the mean (MDDM) in Section 3 and propose a method for the identification of multiple HLP's in logistic regression using the MDDM. The performance of this newly proposed method is then investigated in Section 4 by few well-known examples. In Section 5, we report a Monte Carlo simulation study which is designed to investigate the performance of the proposed method under a variety of situations.

2. Identification of HLP in logistic regression

Leverage values play an extremely important role in logistic regression. These are the quantities that provide the fitted values as the projection of the outcome variable into the covariate space. So it is really important to know which observations in the X-space are responsible for yielding unusual fitted responses.

Let us consider a binary logistic model

$$E(Y \mid X)\pi = (X), \tag{1}$$

where

$$\pi(X) = \frac{\exp(Z)}{1 + \exp(Z)} \tag{2}$$

with

$$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p = X\beta. \tag{3}$$

Here, Y is an $n \times 1$ vector of binary responses; where $y_i = 0$ if the ith unit does not have the characteristic and $y_i = 1$ if the ith unit does possess that characteristic, $X$is an $n \times k$ matrix containing the data for each case with $k = p + 1$, $\beta^T = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)$ is the vector of regression

parameters. The quantity π_i is known as probability for the ith factor/covariate. The model

$$Y \sim \text{Bernoulli}(\pi(X)) \tag{4}$$

satisfies the important requirement that $0 \leq \pi_i \leq 1$ and will be a satisfactory model in many applications.

The ordinary least-squares method is the most commonly used method for estimating parameters in linear regression. This method can be extended to logistic regression as well, but it would violate most of the traditionally used assumptions under which the least-squares estimators possess good properties. The maximum likelihood method based on iterative-reweighed least-squares algorithm [21] has been in use instead in logistic regression. Let $\hat{\beta}$ denote the vector of estimated coefficients. Thus the fitted values for the logistic regression model are

$$\hat{\pi}(x_i) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k)}, \quad i = 1, 2, \ldots, n.$$

The fitted values in logistic regression are the estimated probabilities denoted by

$$\hat{y}_i = \hat{\pi}_i, \quad i = 1, 2, \ldots, n. \tag{5}$$

In logistic regression, we have Bernoulli random variable and as a result the variance is a function of the conditional mean [13], i.e.

$$\text{Var}(y_i/x_i) = v_i = \widehat{\pi}_i(1 - \widehat{\pi}_i), \quad i = 1, 2, \ldots, n. \tag{6}$$

We have already mentioned that the unusual set observations in the X-space are known as HLP. These are the quantities that provide the fitted values as the projection of the outcome variable into the covariate space. So it is really important to know which observations in the X-space are affecting the fitted responses. According to Hocking and Pendleton [12], 'High-leverage points ... are those for which the input vector x_i is, in some sense, far from the rest of the data.'

For linear regression, the well-known measure of leverage is given by the diagonal elements of the weight (or hat or leverage) matrix

$$W = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}. \tag{7}$$

For a perfect balanced design, the ith diagonal element of W can be written as

$$w_i = \frac{1}{n} + \frac{(x_{i1} - \bar{x}_{.1})^2}{\sum (x_{i1} - \bar{x}_{.1})^2} + \frac{(x_{i2} - \bar{x}_{.2})^2}{\sum (x_{i2} - \bar{x}_{.2})^2} + \cdots + \frac{(x_{ip} - \bar{x}_{.p})^2}{\sum (x_{ip} - \bar{x}_{.p})^2}. \tag{8}$$

Here, w_i is the Euclidean distance between the ith vector and the center of gravity $(1/n, \bar{x}_{.1}, \bar{x}_{.2}, \ldots, \bar{x}_{.p})$ for all vectors. Thus, w_i's indicate how much the corresponding vector deviates from the bulk of the values of the explanatory variable. The larger the value of w_i the more extreme is the corresponding vector of observations of the explanatory variable. If we look at the expression (8) we clearly observe that w_i contains both mean and sum of squared deviations from the mean both of which are largely affected by any unusual value of X. Much work has been done on the identification of HLP in linear regression [3,9,11,14,22]. A linear regression-like approximation for logistic regression was derived by Pregibon [19]. Based on his work, we can define a hat

matrix for logistic regression as

$$H = V^{1/2}X(X^{\mathrm{T}}VX)^{-1}X^{\mathrm{T}}V^{1/2}, \tag{9}$$

where V is the diagonal matrix with general elements v_i defined in Equation (6). The ith diagonal element of the matrix H can be expressed as

$$h_i = \hat{\pi}_i(1-\hat{\pi}_i)x_i^{\mathrm{T}}(X^{\mathrm{T}}VX)^{-1}x_i, \tag{10}$$

where $x_i^{\mathrm{T}} = [1, x_{1i}, x_{2i}, \ldots, x_{pi}]$ is the $1\times k$ vector of observations corresponding to the ith case. Now whatever be the choice of the hat matrix, W or H,

$$\sum_{i=1}^{n} w_i = \mathrm{Trace}(W) = k = \mathrm{Trace}(H) = \sum_{i=1}^{n} h_i, \tag{11}$$

where $k = p + 1$. Since the average value of h_i is k/n, Hoaglin and Welsch [11] consider observations unusual when h_i exceeds $2k/n$ which is also known as *twice-the-mean* (2M) rule. Vellman and Welsch [22] suggest considering the *thrice-the-mean* (3M) rule where h_i is considered as large when it exceeds $3k/n$. Other popular methods for the detection of HLP are the method suggested by Huber [14], the method based on Mahalanobis distance [20] and the method based on potentials as suggested by Hadi [9].

Although the h_i values as defined in Equation (10) are very popular measures of leverages and observations possessing large h_i values are known as HLP, but they have potential disadvantages as well. In linear regression, the leverage value is a monotonic increasing function of the distance of a covariance pattern from the mean. But Hosmer and Lemeshow [13] pointed out that in logistic regression, the most extreme points in the covariate space may not necessarily have high leverage if its weight is very small. For a logistic regression model, the ith leverage value is

$$h_i = \hat{\pi}_i(1-\hat{\pi}_i)x_i^{\mathrm{T}}(X^{\mathrm{T}}VX)^{-1}x_i = v_i b_i, \tag{12}$$

where

$$b_i = x_i^{\mathrm{T}}(X^{\mathrm{T}}VX)^{-1}x_i. \tag{13}$$

If we look at the leverage value as defined in Equation (12), we will observe that a quantity that does increase with the distance from the mean (DM) is b_i as defined in Equation (13). But to compute the leverage value, this quantity is multiplied by another quantity $\hat{\pi}_i(1-\hat{\pi}_i)$. For an extreme data point in the X-space, it is expected that the quantity $\hat{\pi}_i$ should be very close to 0 or 1 which automatically implies that the product $\hat{\pi}_i(1-\hat{\pi}_i)$ should be very close to 0. Hence even if the quantity b_i is large, its corresponding h_i could be very small making the procedure of identifying HLP on the magnitude of h_i very cumbersome. Hosmer and Lemeshow [13] suggest focusing on b_i if we are only interested in measuring the distance, however, they did not suggest any identification method based on these quantities. Imon [16] suggested that it is not easy to derive a theoretical distribution of b_i, but it does not make any problem to obtain a suitable confidence bound type cut-off point for them. He considered b_i to be large if

$$b_i > \mathrm{Median}(b_i) + 3\,\mathrm{MAD}(b_i). \tag{14}$$

This type of form, analogous to a confidence bound for a location parameter, which was first introduced by Hadi [8] in regression diagnostics and then used by many others [8,15,18]. The rule given in Equation (14) is based on the median of distances from the mean and for this reason we call it MDM.

3. Identification of multiple HLP

All the detection methods discussed in the previous section are designed for the identification of a single high leverage point. But we often observe that after the deletion of one high leverage point another observation may emerge as a point of extremely high leverage that was not visible at first. This effect is generally known as masking, [20] for which high leverage cases appear as points previously of low leverage. The opposite effect of masking is known as swamping [2] for which for which low leverage cases are classified as HLP. Mainly because of these two effects, single case deleted diagnostics have been proved to be ineffective and therefore a group deletion version of leverage measure is required for logistic regression.

Let us now partition the entire data set into two groups. We assume that d observations among a set of n observations are omitted before the fitting of the model. Here, we denote a set of $(n-d)$ cases 'remaining' in the analysis by R and a set of d cases 'deleted' by D. Without loss of generality, assume that the deleted cases are the last of $d$rows of X, Y and V so that

$$X = \begin{bmatrix} X_R \\ X_D \end{bmatrix} \quad Y = \begin{bmatrix} Y_R \\ Y_D \end{bmatrix} \quad V = \begin{bmatrix} V_R & 0 \\ 0 & V_D \end{bmatrix}. \tag{15}$$

When a group of observations indexed by D is omitted, we estimate the parameter with the remaining observations. We denote the corresponding vector of estimated coefficients by $\hat{\beta}^{(-D)}$. Thus the corresponding fitted values for the logistic regression model are

$$\hat{\pi}_i^{(-D)} = \frac{\exp(x_i^{\mathrm{T}}\hat{\beta}^{(-D)})}{1+\exp(x_i^{\mathrm{T}}\hat{\beta}^{(-D)})}, \quad i = 1, 2, \ldots, n. \tag{16}$$

Here, we define the ith deletion variances and deletion leverages for the entire data set as

$$v_i^{(-D)} = \hat{\pi}_i^{(-D)}(1-\hat{\pi}_i^{(-D)}), \tag{17}$$

$$h_i^{(-D)} = \hat{\pi}_i^{(-D)}(1-\hat{\pi}_i^{(-D)})x_i^{\mathrm{T}}(X_R^{\mathrm{T}}V_RX_R)^{-1}x_i. \tag{18}$$

Using the above results and also using linear regression-like approximation, we define the deletion distance from the mean (DDM) for the entire data set after the deletion of a group of suspect HLP indexed by D, as

$$b_i^{(-D)} = x_i^{\mathrm{T}}(X_R^{\mathrm{T}}V_RX_R)^{-1}x_i, \quad i = 1, 2, \ldots, n. \tag{19}$$

It should be noted that $b_i^{(-D)}$ is the ith diagonal element of $X(X_R^{\mathrm{T}}V_RX_R)^{-1}X^{\mathrm{T}}$ matrix. Like b_i there exists no finite upper bound for DDMs. Hence, following Hadi [9] we consider DDM to be large if

$$b_i^{(-D)} > \mathrm{Median}(b_i^{(-D)}) + 3\,\mathrm{MAD}(b_i^{(-D)}). \tag{20}$$

The rule given in Equation (20) is based on the median of deletion distances from the mean and for this reason we call it MDDM.

Although the expression for DDM in Equation (19) is available for any arbitrary set of deleted cases, D, the choice of such a set is very important since the omission of this group determines leverages for the whole set. Hence, we suggest a two-step procedure for identifying multiple HLP in logistic regression.

Step 1: At first we suggest employing any suitable robust multivariate technique such as the robust Mahalanobis distance based on minimum volume ellipsoid (MVE) or minimum covariance determinants (MCD) proposed by Rousseeuw and Leroy [20] or the block adaptive computationally efficient outlier nominator proposed by Billor and Hadi [4], or the diagnostic-robust generalized potentials proposed by Habshah *et al.* [8] to identify suspect HLP.

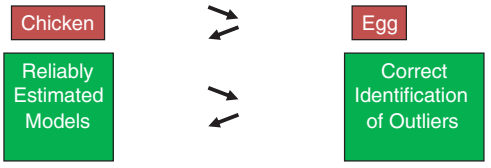

Figure 1. Identification of outliers is a circular problem.

Step 2: After the omission of suspect cases, the model is refitted with the rest of the data and the DDM values are computed for the entire data set. Observations corresponding to DDM values satisfying rule (20) are declared as HLP.

Unfortunately, because of the insidious nature of the masking effect, the identification of the outliers in multivariate data is not an easy task. Here, we present a figure (Figure 1) taken from [10] which gives a pictorial illustration of this problem.

The assumed models usually depend on unknown parameters. The parameters have to be estimated from the data. But the estimates will be affected by the outliers when the data contain outliers. Consequently, the estimates will be wrong, and some or all of the outliers may be left undetected due to masking and/or swamping. To quote Hadi *et al.* [10]:

> These incorrect classifications continue the cycle by leading to wrong parameter estimates. All classical statistical methods are affected by the masking and swamping problems. Thus, in order for us to obtain reliably estimated models, we need to know the outliers in the data. But to know the outliers in the data, our estimates should not be affected by the outliers. Outlier identification is indeed very much like the chicken-and-egg problem.

The conventional robust detection methods only consider the first step. When we apply robust techniques such as MCD or MVE we do not need to know how many suspect HLP are there. These techniques compute diagnostics on all possible subsets of observations where the estimation subset is almost half the original set. The good thing about the robust techniques is that since it focuses on the most compact data set, the group which contains all high leverage cases can easily be identified. But on many occasions we lose information in this process. If the number of contaminations is small then omitting almost 50% of the cases every time may make the detection technique too much prone in declaring cases as HLP when they are not. Cook and Hawkins [7] applied MCD and MVE in the detection of outliers in a sample that was generated as a genuine normal distribution. There was nothing wrong with these data, but these robust methods identified 6 out of 20 observations as outliers. Cook and Hawkins [7] described this situation as 'outliers everywhere'. On the other hand for the diagnostic approach, we have to have a group of suspect or candidate cases (all of them need not to be genuine HLP) before applying diagnostics there. If we consider the full sample results, then the presence of HLP contaminate the entire leverage structure in such a way that we may not identify the genuine cases at the first place or by successive applications of diagnostics. Our proposed method is a combination of both diagnostic and robust approaches. The advantage of our method is that we are applying robust techniques such as MCD or MVE initially, so our group of suspect cases D definitely contains all potential HLP if any. But this is not our final choice. We are ready to put back all non-high leverage cases into the estimation subset so that our estimation subset no longer contains only 50% of the cases, it may be the entire data set if there is no high leverage cases, or a group which contains a much bigger subset than what MCD or MVE yield. Thus, we expect that our conclusion should be more accurate than the robust methods because we are allowing much more information than robust methods.

4. Examples

In this section we consider few real-world and artificial examples to investigate the performance of the proposed method in the identification of multiple HLP in logistic regression.

Table 1. Brown cancer data.

Index	LNI	AP	Index	LNI	AP	Index	LNI	AP	Index	LNI	AP
1	0	48	15	0	47	29	0	50	43	1	81
2	0	56	16	0	49	30	0	40	44	1	76
3	0	50	17	0	50	31	0	55	45	1	70
4	0	52	18	0	78	32	0	59	46	1	78
5	0	50	19	0	83	33	1	48	47	1	70
6	0	49	20	0	98	34	1	51	48	1	67
7	0	46	21	0	52	35	1	49	49	1	82
8	0	62	22	0	75	36	0	48	50	1	67
9	1	56	23	1	99	37	0	63	51	1	72
10	0	55	24	0	**187**	38	0	102	52	1	89
11	0	62	**25**	1	**136**	39	0	76	**53**	1	**126**
12	0	71	26	1	82	40	0	95			
13	0	65	27	0	40	41	0	66			
14	1	67	28	0	50	42	1	84			

Note: The bold values are used to highlight the unusual points.

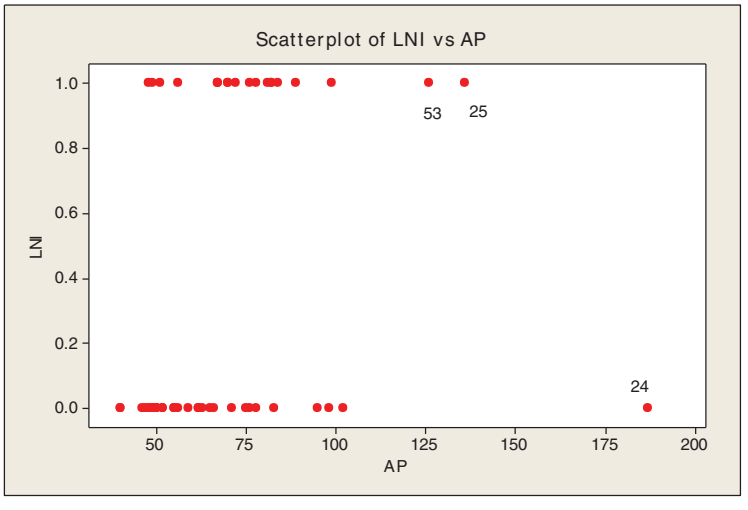

Figure 2. Scatter plot of LNI versus AP for Brown data.

4.1 *Brown cancer data*

We first consider a cancer data set given by Brown [5]. The original objective of the author was to see whether an elevated level of acid phosphates (AP) in the blood serum would be of value for predicting whether or not prostate cancer patients also had lymph node involvement (LNI). The data set additionally contains data on the four more commonly used regressors, but we use here only AP in illustrating simple logistic regression with 53 cases. The observations from 53 patients are given in Table 1.

This data set has been extensively analyzed by many authors [16–18,21]. The index plot of each of the variables shows that only the variable AP may contain few HLP. We present the scatter plot of LNI against AP in Figure 2, which indicates that observations 24, 25 and 53 may be HLP.

We first apply the commonly used diagnostics for the identification of HLP and the results are presented in Table 2. Here, the mean value of the leverages is 0.0377. Both the twice-the mean rule (cut-off = 0.0755) and the thrice-the-mean rule (cut-off = 0.1132) can successfully identify three observations (cases 24, 25 and 53) as HLP without swamping a good one. The

Table 2. High leverage diagnostics for Brown [5] data.

Index	$\hat{\pi}$	Lev	DM	DDM	Index	$\hat{\pi}$	Lev	DM	DDM
1	0.2793	0.0319	0.1584	0.2345	28	0.2876	0.0298	0.1455	0.2067
2	0.3133	0.0159	0.0741	0.1421	29	0.2876	0.0298	0.1455	0.2067
3	0.2876	0.0298	0.1455	0.2067	30	0.2477	0.0415	0.2229	0.3771
4	0.2960	0.0279	0.1338	0.1820	31	0.3090	0.0253	0.1186	0.1509
5	0.2876	0.0298	0.1455	0.2067	32	0.3266	0.0226	0.1028	0.1204
6	0.2835	0.0308	0.1518	0.2202	33	0.2793	0.03190	0.15842	0.2345
7	0.2712	0.0341	0.1727	0.2654	34	0.2918	0.0288	0.13951	0.1940
8	0.3402	0.0212	0.0943	0.1057	35	0.2835	0.0308	0.15185	0.2202
9	0.3133	0.0159	0.0741	0.1421	36	0.2793	0.0319	0.1584	0.2345
10	0.3090	0.0253	0.1186	0.1509	37	0.3448	0.02080	0.0920	0.1024
11	0.3402	0.0212	0.0943	0.1057	38	0.5384	0.06262	0.2520	0.5446
12	0.3826	0.0202	0.0857	0.1041	39	0.4070	0.0222	0.0920	0.1306
13	0.3541	0.0203	0.0886	0.0981	40	0.5028	0.04670	0.1879	0.4102
14	0.3635	0.0200	0.0864	0.0970	41	0.3588	0.0201	0.0873	0.0971
15	0.2752	0.0330	0.1654	0.2496	42	0.4469	0.0293	0.1185	0.2139
16	0.2835	0.0308	0.1518	0.2202	43	0.4318	0.0260	0.1062	0.1768
17	0.2876	0.0298	0.1455	0.2067	44	0.4070	0.0222	0.0920	0.1306
18	0.4168	0.0235	0.0967	0.1467	45	0.3778	0.0201	0.0854	0.1011
19	0.4418	0.0281	0.1141	0.2007	46	0.4168	0.0235	0.0967	0.1467
20	0.5181	0.0533	0.2135	0.4803	47	0.3778	0.0201	0.0854	0.1011
21	0.2960	0.0279	0.1338	0.1820	48	0.3635	0.0200	0.0863	0.0970
22	0.4020	0.0217	0.0901	0.1237	49	0.4368	0.0271	0.1100	0.1883
23	0.5231	0.0555	0.2226	0.5052	50	0.36351	0.0200	0.0863	0.0970
24	0.8685	**0.2586**	**2.2646**	**5.7680**	51	0.3874	0.0205	0.0863	0.1078
25	0.7000	**0.1645**	**0.7833**	**1.9781**	52	0.4722	0.0362	0.1453	0.2914
26	0.4368	0.0271	0.1100	0.1883	53	0.6555	**0.1330**	**0.5891**	**1.4742**
27	0.2477	0.0415	0.223	0.3771					

Note: The bold values are used to highlight the unusual points.

MDM method proposed by Imon [15] gives the cut-off point 0.2580 and hence can correctly identify cases 24, 25 and 53 as HLP. We then apply our newly proposed diagnostic to the Brown data. We employ the MCD method to identify the suspect cases and observations 24, 25 and 53 are identified as suspects. The diagnostics based on the DDM gives a cut-off value 0.7043. We observe from Table 2 that the DDM values for the observations 24, 25 and 53 are bigger than the cut-off value. Thus, our proposed method can correctly identify all three HLP without swamping any low leverage cases.

We observe exactly the same kind of scenario when we look at the index plots of different leverage measures as given in Figure 3. All the four methods (2M, 3M, MDM and MDDM) considered here can correctly identify all three HLP and do not swamp in any low leverage cases.

4.2 *Modified Brown data*

Now we modify the original Brown [2] data in a way that was done by Imon and Hadi [17] in an experiment for the identification of multiple outliers. Here we insert seven more HLP (cases 54 through 60) with $X = 200$, 220, 240, 260, 280, 300 and 320, respectively. The scatter plot of LNI against AP as shown in Figure 4 indicates that 10 observations (case 24, 25, 53–60) may be potentially HLP for this data set.

Table 3 gives leverage diagnostics for the modified Brown data. Here, the mean value of the leverages is 0.0333. The 2M rule where the cut-off point is 0.6667 can successfully identify all the seven new HLP (cases 54–60) but fails to identify the previous ones (cases 24, 25 and 53). The performance of the 3M rule is even worse in this situation. Although it correctly identifies five HLP's (cases 56–60), it masks the other five (cases 24, 25, 53, 54 and 55). The MDM method

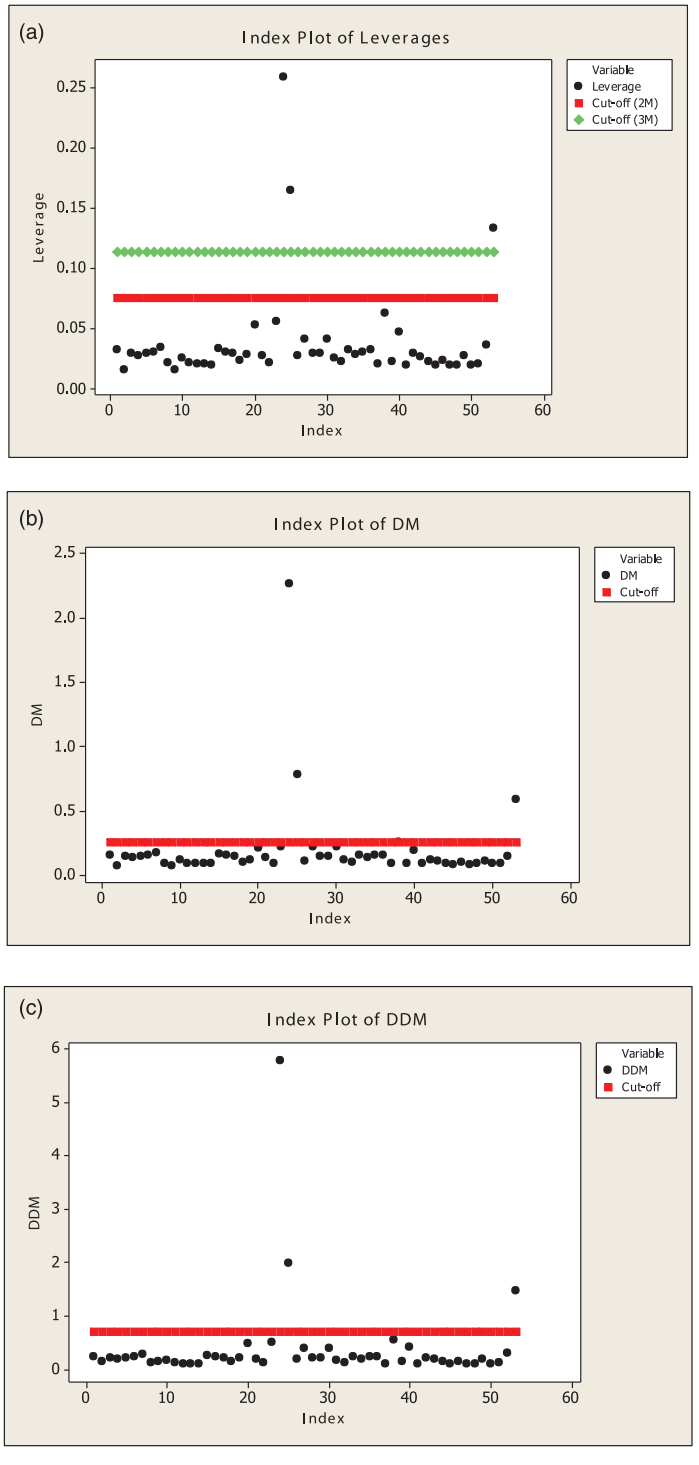

Figure 3. Index plot of (a) leverages (b) DM's and (c) DDM's for the Brown data.

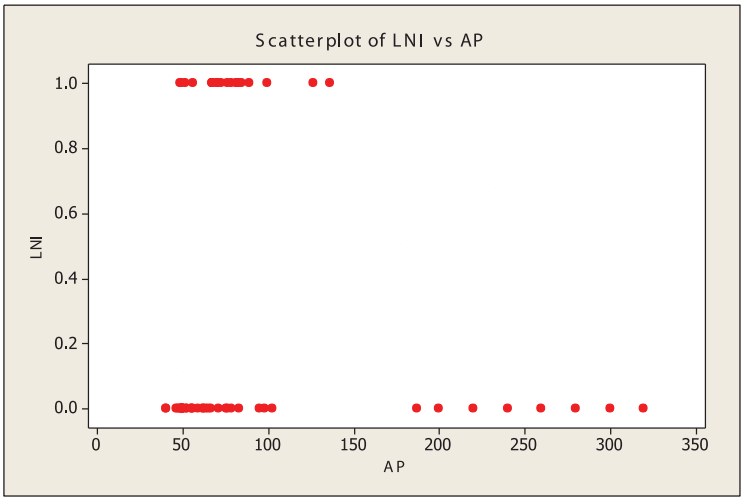

Figure 4. Scatter plot of LNI versus AP for the modified Brown data.

Table 3. High leverage diagnostics for the modified Brown data.

Index	$\hat{\pi}$	Lev	DM	DDM	Index	$\hat{\pi}$	Lev	DM	DDM
1	0.3836	0.0250	0.1057	0.2345	31	0.3743	0.02220	0.0949	0.1509
2	0.3730	0.0219	0.0935	0.1421	32	0.3690	0.0209	0.0898	0.1204
3	0.3809	0.0241	0.1024	0.2067	33	0.3836	0.0250	0.1058	0.2345
4	0.37830	0.0233	0.0992	0.1820	34	0.3796	0.0237	0.1008	0.1940
5	0.3809	0.0241	0.1024	0.2067	35	0.3823	0.0246	0.1040	0.2202
6	0.3823	0.0246	0.1040	0.2202	36	0.3836	0.0250	0.1058	0.2345
7	0.3863	0.0259	0.1093	0.2654	37	0.3638	0.0198	0.0856	0.1024
8	0.3651	0.0201	0.0865	0.1057	38	0.3144	0.0189	0.0876	0.5446
9	0.3730	0.0219	0.0935	0.1421	39	0.3469	0.0176	0.0775	0.1306
10	0.3743	0.0222	0.0949	0.1509	40	0.3230	0.0178	0.0814	0.4102
11	0.3651	0.0201	0.0865	0.1057	41	0.3598	0.0191	0.0829	0.0971
12	0.3534	0.0182	0.0796	0.1041	42	0.3367	0.0172	0.0769	0.2139
13	0.3611	0.0193	0.0837	0.0981	43	0.3405	0.0172	0.0767	0.1768
14	0.3585	0.0189	0.0821	0.0970	44	0.3469	0.0176	0.0775	0.1306
15	0.3850	0.0254	0.1076	0.2496	45	0.3546	0.0183	0.0801	0.1011
16	0.3823	0.0246	0.1040	0.2202	46	0.3444	0.0174	0.0770	0.1467
17	0.3809	0.0241	0.1024	0.2067	47	0.3546	0.0183	0.0801	0.1011
18	0.3444	0.0174	0.0770	0.1467	48	0.3585	0.0189	0.0821	0.0970
19	0.3380	0.0172	0.0768	0.2007	49	0.3393	0.0172	0.0767	0.1883
20	0.3193	0.0182	0.0838	0.4803	50	0.3585	0.0189	0.0821	0.0970
21	0.3783	0.0233	0.0992	0.1820	51	0.3521	0.0180	0.0790	0.1078
22	0.34826	0.0177	0.07780	0.1237	52	0.33043	0.0173	0.0782	0.2914
23	0.3180	0.0184	0.0847	0.5052	53	0.2859	0.0261	0.1280	**1.4742**
24	0.2209	0.0628	**0.3651**	**5.7680**	54	0.20851	0.07274	**0.4405**	**7.0602**
25	0.2745	0.0306	**0.1537**	**1.9781**	55	0.1904	0.0884	**0.5737**	**9.3070**
26	0.3393	0.0172	0.0767	0.1883	56	0.1736	**0.1044**	**0.7276**	**11.8675**
27	0.3944	0.0290	0.1213	0.3771	57	0.1580	**0.1200**	**0.9022**	**14.7416**
28	0.3809	0.0241	0.1024	0.2067	58	0.1435	**0.1349**	**1.0975**	**17.9293**
29	0.38095	0.0241	0.1024	0.2067	59	0.1301	**0.1487**	**1.3135**	**21.4307**
30	0.3944	0.0290	0.1213	0.3771	60	0.1179	**0.1612**	**1.5503**	**25.2458**

Note: The bold values are used to highlight the unusual points.

proposed by Imon [15] gives the cut-off point 0.1494 and hence correctly identifies nine HLP but fails to detect the case 53. We then apply our newly proposed diagnostic to the Brown data. We first employ the MCD for finding out the suspect HLP and it identifies 10 observations (cases 24,

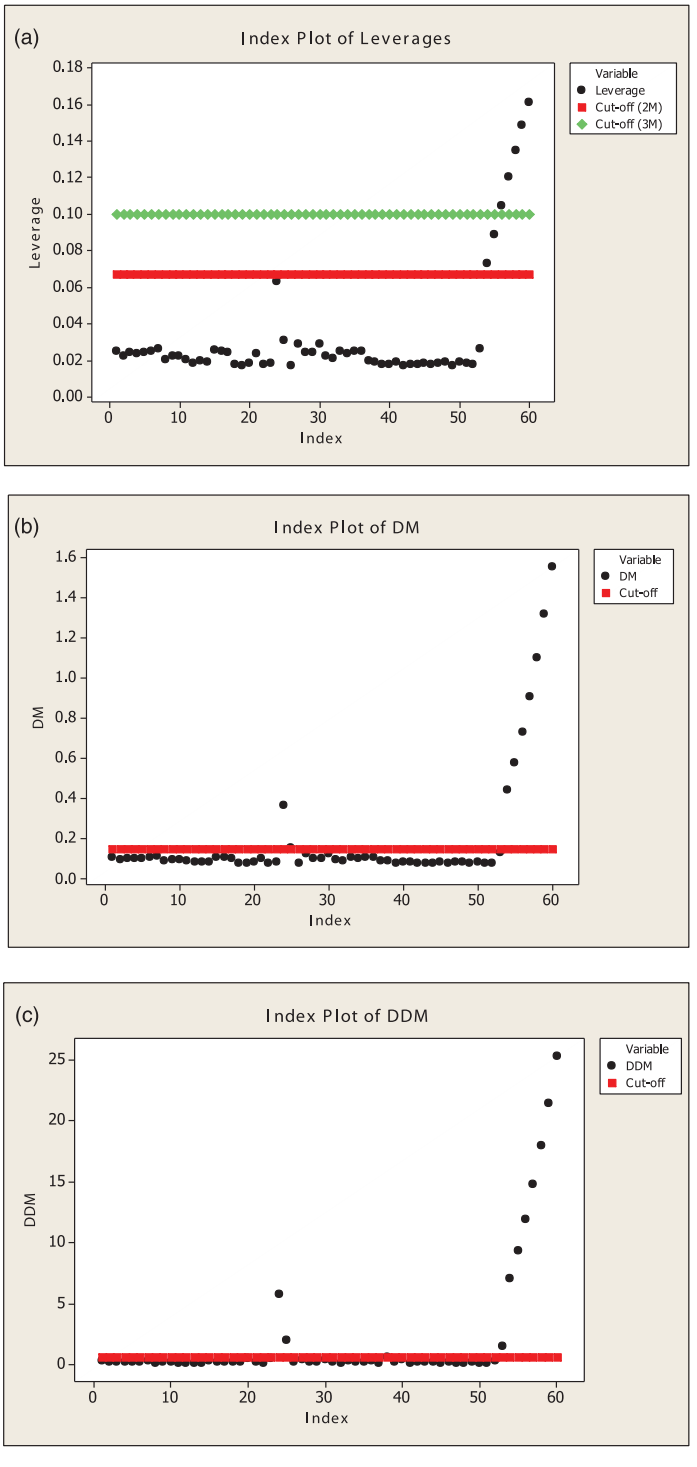

Figure 5. Index plot of (a) leverages (b) DM's and (c) DDM's for the modified Brown data.

25, 53–60) as suspects. The diagnostics based on the DDM gives a cut-off value 0.5797. We observe from Table 3 that the DDM values for the observations 24, 25 and 53–60 are bigger than the cut-off value. Thus, our proposed method can correctly identify all 10 HLP without swamping any low leverage cases.

We observe from Figure 5(a) that the 2M rule and the 3M rule fail to identify five and three HLP's, respectively. Figure 5(b) shows that the MDM rule clearly identifies eight cases, one is on the border line and one is undetected. Figure 5(c) shows that the newly proposed MDDM method can correctly identify all 10 HLP's and they are clearly separated from the rest of the data.

4.3 *Artificial high leverage data*

Finally, we consider an artificial data. The first two columns of Table 4 give this single predictor data set that contains twenty observations with six HLP. The first 14 observations of X are generated from Uniform (10, 20) and the last six observations are set at 30, 32, 34, 36, 38 and 40, respectively. Since the mean of the distribution of the first 14 points is 20 with a standard deviation 2.88, the first possible high leverage point is located at least in three sigma distance away from the upper end point of this distribution. The values of Y are assigned in a way that the first five values are set at 0, the next five at 1 and we repeat the whole sequence once again to generate all twenty observations.

The scatter plot of Y versus X as shown in Figure 6 clearly shows the existence of six HLP for this artificial data. In the presence of the last five HLP, the first one may not look unusual but when we omit the last five, the first one emerges as extremely unusual. This is a perfect example of masking which has been extensively discussed in [20]. Our artificial data set contains 30% HLP. To achieve the highest possible 50% breakdown properties, many studies in the robustness literature consider artificial data seta that possess 50% unusual cases, but we feel that such a high percentage of outliers do not make much practical sense as the routine data typically contains roughly 10% unusual cases. Leverage diagnostics for this artificial data are shown in Table 4.

Table 4. High leverage diagnostics for the artificial data.

Index	X	Y	$\hat{\pi}$	Leverage	DM	DDM
1	10.2379	0	0.2338	0.0952	0.1380	1.3726
2	11.2833	0	0.2582	0.0891	0.0722	0.9281
3	11.6081	0	0.2661	0.0873	0.0536	0.8146
4	11.8463	0	0.2720	0.0859	0.0406	0.7387
5	13.6030	0	0.3178	0.0764	0.0406	0.3729
6	14.2217	1	0.3349	0.0736	0.0629	0.3253
7	14.2455	1	0.3356	0.0735	0.0637	0.3243
8	14.6833	1	0.3479	0.0716	0.0775	0.3173
9	15.4942	1	0.3714	0.0688	0.0986	0.3603
10	15.6448	1	0.3758	0.0683	0.1019	0.3763
11	16.1364	0	0.3904	0.0671	0.1113	0.4459
12	16.5975	0	0.4043	0.0662	0.1183	0.5355
13	16.9506	0	0.4150	0.0657	0.1224	0.6200
14	19.2271	0	0.4857	0.0674	0.1235	1.4954
15	30.0000	0	0.7853	0.1457	0.4707	**13.4049**
16	32.0000	1	0.8246	0.1566	0.6990	**17.0273**
17	34.0000	1	0.8581	0.1628	**0.9433**	**21.0917**
18	36.0000	1	0.8860	0.1640	**1.2308**	**25.5981**
19	38.0000	1	0.9091	0.1608	**1.5524**	**30.5466**
20	40.0000	1	0.9278	0.1541	**1.9080**	**35.9370**

Note: The bold values are used to highlight the unusual points.

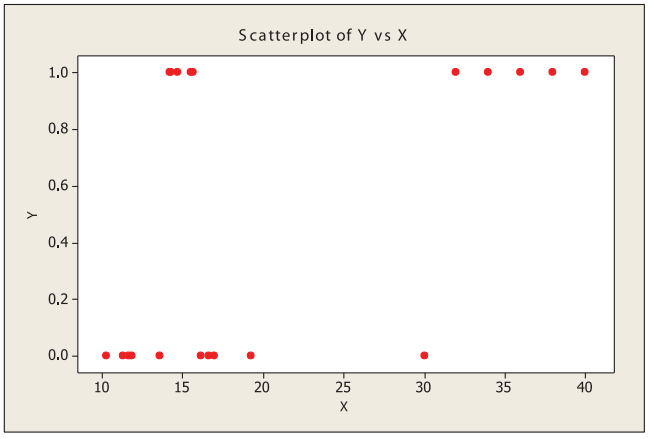

Figure 6. Scatter plot of Y versus X for the artificial data.

Here, the average leverage is 0.1. We observe that the 2M rule with the cut-off point 0.2 fails to identify even a single high leverage point. Similar remarks apply with the 3M rule. The cut-off value for the MDM method 0.9043 and thus it can correctly identify four HLP (cases 17–20), but fails to identify the other two (cases 15 and 16). When the MCD is employed for finding out the suspect HLP, the last six observations are flagged as suspects. Diagnostics based on the DDM yield cut-off value which is 2.7877. Results presented in Table 4 show that the DDM values for observations 15–20 exceed the cut-off points and hence all HLP are correctly identified.

We observe from Figure 7(a) that the 2M rule and the 3M rule fail to identify even a single observation as HLP. Figure 7(b) shows that the MDM rule identifies four out of six cases. Figure 7(c) shows that the newly proposed MDDM method can correctly identify all six HLP's and they are clearly separated from the rest of the data.

5. Monte Carlo simulations

In this section, we report a Monte Carlo simulation which is designed to investigate the performances of different measures of leverages in logistic regression Here we consider three different sample sizes, $n = 20$, 40 and 100, we generated the X values from Uniform (10, 20). Here, we consider four different percentages, i.e. 0%, 10%, 20% and 30% of HLP. The X value corresponding to the first high leverage value is set at 30 and the next values have an increment of 2 each. The first 25% observations of Y are set at 0 and the next 25% values of Y are 1 each and we replicate this design once again to generate the entire Y. For each different sample, we apply all four different leverage identification rules; 2M, 3M, MDM and MDDM, and compute the correct identification rate (IR) and the swamping rate (SR) in terms of percentages. It is worth mentioning that the IR is the percent of all points correctly identified. For simulation, we use a program written in S-Plus where MCD was used for finding the suspect cases for the proposed MDDM method. We run 10,000 simulations for each combination and these results are presented in Table 5.

The above simulation results show that the SR of all four methods considered in this study are low as they have been always less than 1%, but the rates of 2M and MDM are relatively higher than 3M and MDDM. In terms of the identification of HLP, the thrice-the-mean rule performs worst overall. It has a very low correct IR which also deteriorates with the increase in level of contamination and sample sizes. The performance of the twice-the-mean rule is not satisfactory either. For a sample size of 100 and in the presence of 30% HLP, it can correctly identify even

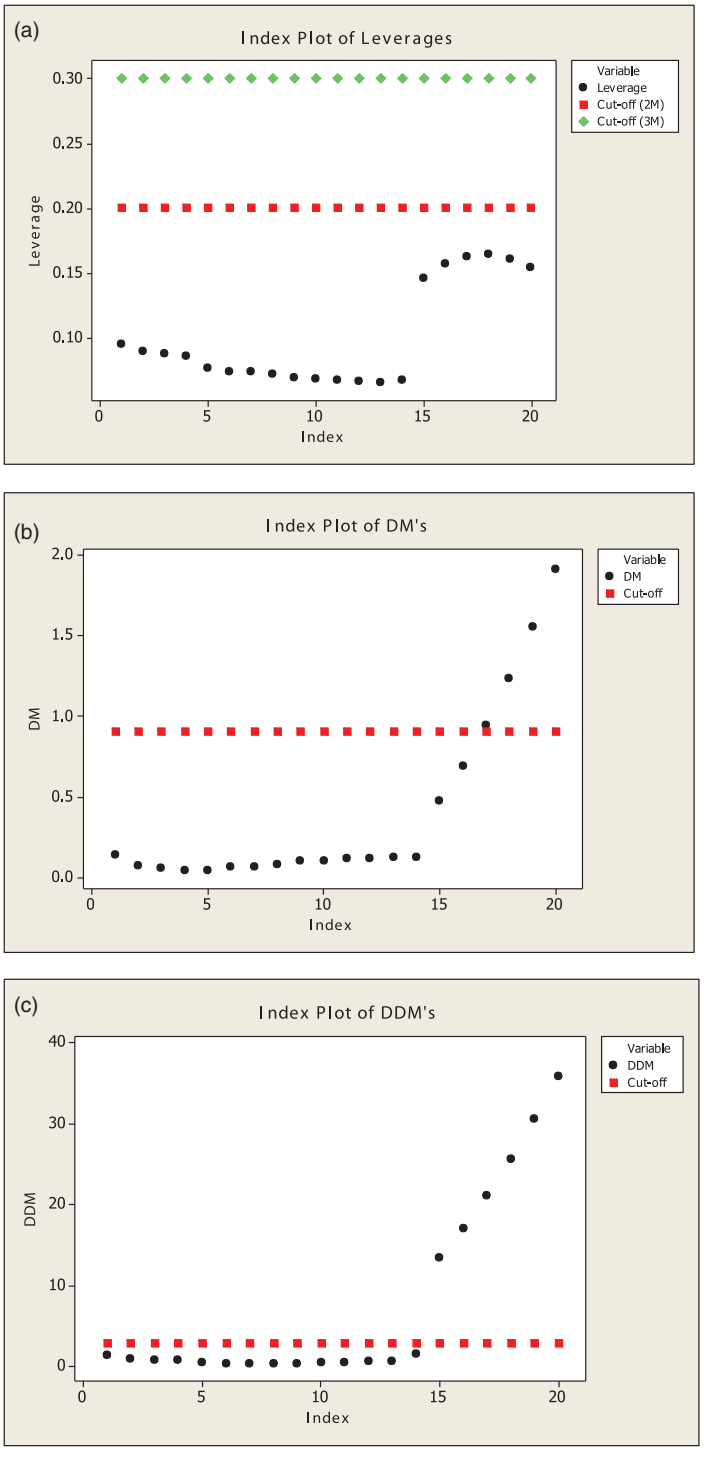

Figure 7. Index plot of (a) leverages (b) DM's and (c) DDM's for the artificial data.

Table 5. Identification and SRs of different leverage detection rules.

		Percentages of HLP							
		IR (%)				SR (%)			
Sample size	Rules	0%	10%	20%	30%	0%	10%	20%	30%
$n = 20$	2M	–	35.23	14.01	2.67	0.4488	0.2350	0.1278	0.0569
	3M	–	18.87	3.53	0.13	0.0231	0.0119	0.0087	0.0006
	MDM	–	100.00	92.73	89.14	0.2450	0.1156	0.0806	0.0138
	MDDM	–	100.00	100.00	100.00	0.0131	0.0069	0.0031	0.0006
$n = 40$	2M	–	22.58	6.95	1.34	0.4123	0.1974	0.0998	0.0367
	3M	–	7.88	1.17	0.02	0.0147	0.0067	0.0017	0.0000
	MDM	–	99.93	89.67	76.87	0.1244	0.0754	0.0378	0.0032
	MDDM	–	100.00	100.00	100.00	0.0023	0.001	0.0000	0.0000
$n = 100$	2M	–	12.35	3.73	0.37	0.3511	0.1114	0.0547	0.0135
	3M	–	1.89	0.35	0.01	0.0029	0.0000	0.0000	0.0000
	MDM	–	96.75	80.95	63.41	0.0756	0.0327	0.0136	0.0004
	MDDM	–	100.00	100.00	100.00	0.0000	0.0000	0.0000	0.0000

less than 1% HLP's. The performance of this method also deteriorates with the increase in level of contamination and sample sizes. The performance of the MDM is quite satisfactory. But its performance also tends to deteriorate when the sample size get increases and the percentage of HLP's is higher. The newly proposed MDDM outperforms all other methods. It maintains 100% correct IR throughout the experiment.

6. Conclusions

In this paper we first observe that the traditionally used measures of leverages may not able to focus on all the potential HLP when they occur in a group. As a remedy to this problem, we propose a new method for the identification of HLP based on the median of group deleted distances from the mean. Both the numerical examples and Monte Carlo simulation results show that the proposed method performs superbly in the identification of multiple HLP when the traditional methods fail to do so.

Acknowledgements

The authors gratefully acknowledge valuable comments and suggestions of the reviewers of this paper.

References

[1] D.F. Andrews and D. Pregibon, *Finding the outliers that matter*, J. Roy. Stat. Soc. Ser. B 40 (1978), pp. 85–93.
[2] V. Barnett and T.B. Lewis, *Outliers in Statistical Data*, Wiley, New York, 1995.
[3] D.A. Belsley, E. Kuh, and R.E. Welsch, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, Wiley, New York, 1980.
[4] N. Billor, A.S. Hadi, and F. Velleman, *BACON: Blocked adaptive computationally-efficient outlier nominator*, Comput. Stat. Data Anal. 34 (2000), pp. 279–298.
[5] B.W. Brown, Jr, *Prediction analysis for binary data*, in *Biostatistics Casebook*, R.G. Miller, Jr, B. Efron, B. W. Brown, Jr, and L.E. Moses, eds., Wiley, New York, 1980, pp. 3–18.
[6] S. Chatterjee and A.S. Hadi, *Influential observations, high leverage points, and outliers in linear regression*, Stat. Sci. 1 (1986), pp. 379–393.
[7] R.D. Cook and H. Hawkins, *Comments on unmasking multivariate outliers and leverage points by P.J. Rousseeuw and B.C. van Zomeren*, J. Am. Stat. Assoc. 85 (1990), pp. 648–651.
[8] M. Habshah, R. Norazan, and A.H.M.R. Imon, *The performance of diagnostic-robust generalized potentials for the identification of multiple high leverage points in linear regression*, J. Appl. Stat. 36 (2009), pp. 507–520.

[9] A.S. Hadi, *A new measure of overall potential influence in linear regression*, Comput. Stat. Data Anal. 14 (1992), pp. 1–27.
[10] A.S. Hadi, A.H.M.R. Imon, and M. Werner, *Detection of outliers*, Wiley Int. Rev. Comput. Stat. 1 (2009), pp. 57–70.
[11] D.C. Hoaglin and R.E. Welsch, *The hat matrix in regression and ANOVA*, Am. Stat. 32 (1978), pp. 17–22.
[12] R.R. Hocking and O.J. Pendleton, *The regression dilemma*, Comm. Stat. Theory Methods 12 (1983), pp. 497–527.
[13] D.W. Hosmer and S. Lemeshow, *Applied Logistic Regression*, Wiley, New York, 1980.
[14] P.J. Huber, *Robust Statistics*, Wiley, New York, 1981.
[15] A.H.M.R. Imon, *Identifying multiple influential observations in linear regression*, J. Appl. Stat. 32 (2005), pp. 929–946.
[16] A.H.M.R. Imon, *Identification of high leverage points in logistic regression*, Pak. J. Stat. 22 (2006), pp. 147–156.
[17] A.H.M.R. Imon and A.S. Hadi, *Identification of multiple outliers in logistic regression*, Comm. Stat. Theory Methods. 37 (2008), pp. 1967–1709.
[18] A.A.M. Nurunnabi, A.H.M.R. Imon, and M. Nasser, *Identification of multiple influential observations in logistic regression*, J. Appl. Stat. 37 (2010), pp. 1605–1624.
[19] D. Pregibon, *Logistic regression diagnostics*, Ann. Stat. 9 (1981), pp. 705–724.
[20] P.J. Rousseeuw and A.M. Leroy, *Robust Regression and Outlier Detection*, Wiley, New York, 1987.
[21] T.P. Ryan, *Modern Regression Methods*, Wiley, New York, 1997.
[22] P.F. Velleman and R.E. Welsch, *Efficient computing in regression diagnostics*, Am. Stat. 35 (1981), pp. 234–242.