Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 10 (23): 3042-3050, 2010
ISSN 1812-5654

Robust Logistic Diagnostic for the Identification of High Leverage Points in Logistic Regression Model

[1]B.A. Syaiba and [1,2]M. Habshah
[1]Department of Mathematics, Faculty of Science,
Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia
[2]Institute for Mathematical Research, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia

Abstract: High leverage points are observations that have outlying values in covariate space. In logistic regression model, the identification of high leverage points becomes essential due to their gross effects on the parameter estimates. Currently, the distance from the mean diagnostic method is used to identify the high leverage points. The main limitation of the distance from the mean diagnostic method is that it tends to swamp some low leverage points even though it can identify the high leverage points correctly. In this study, we propose a new diagnostic method for the identification of high leverage points. Robust approach is firstly used to identify suspected high leverage points by computing the robust mahalanobis distance based on minimum volume ellipsoid or minimum covariance determinant estimators. For confirmation, the diagnostic procedure is used by computing the group deleted potential. We called this proposed diagnostic method the robust logistic diagnostic. The performance of the proposed diagnostic method is then investigated through real examples and monte carlo simulation study. The result of this study indicates that the proposed diagnostic method ensures only correct high leverage points are identified and free from swamping and masking effects.

Key words: Logistic regression model, high leverage points, masking, swamping, robust mahalanobis distance, group deleted potential

INTRODUCTION

Diagnostic is one of important part that needs to be considered when analyzing data since leverage outliers can easily biased the parameter estimates and obscure other observations in logistic regression model. Leverage outliers are outlying points with the respect to the explanatory variables. It is referred as bad leverages or high leverage points if their presences have high influence on the model fit and depart away from the fitted pattern set by the rest of data (Croux and Haesbroeck, 2003; Imon, 2006). Rousseeuw (1991) declared that the high leverage points do not fit the model at all and they are the most dangerous kind of outliers because they have the largest effect on the classical Maximum Likelihood Estimation (MLE). The presence of high leverage points causes more difficulties to logistic regression model. The effect from the high leverage points is more severe than other kind of bad points. Imon (2006) pointed out that high leverage points are not only responsible for producing wrong parameter estimates but also capable of masking the high leverage points. Masking effect (false negative) occurs when outlying points go undetected because of the presence of another high leverage points. Meanwhile, swamping effect (false positive) occurs when good points are incorrectly identified as bad points (Hadi and Simonoff, 1993; Imon, 2005, 2006; Imon and Apu, 2007; Imon and Hadi, 2008; Nurunnabi *et al*., 2009; Habshah *et al*., 2009).

According to Imon (2006), the assessment of high leverage points are equally important as the detection of residual outliers in regression analysis. It is now evident that the high leverage points have huge tendency to break the covariate pattern which result in biased parameter estimates especially for the MLE with zero breakdown point or robust estimators with small breakdown point (Brown *et al*., 1980; Pregibon, 1981; Jennings, 1986; Williams, 1987; Bedrick and Hill, 1990; Munier, 1999; Hosmer and Lemeshow, 2000; Imon, 2006; Imon and Hadi, 2008). Even a single high leverage point is enough to suffer the estimates thus result in completely erroneous estimation. Therefore, it is important to identify the high leverage points but it is more critical to detect them correctly at very early stage especially when the dimension (number of covariates) is high due to masking and swamping effects.

There are a number of detection methods for high leverage points in logistic regression model (Hoaglin and

Corresponding Author: Syaiba Balqish Ariffin, Department of Mathematics, Faculty of Science, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia

Welsch, 1978; Vellman and Welsch, 1981; Jennings, 1986; Hosmer Lemeshow, 2000). Hoaglin and Welsch (1978) and Vellman and Wesch (1981) based their studies on the leverage values while Jenning (1986) work based on the estimated logistic probability. The most recent technique in the identification of high leverage points in logistic regression model is based on the Distance from the Mean (DM) method which is proposed by Imon (2006). The results of his study signified that the DM method is very effective in the identification of high leverage points while the methods based on the leverage values and estimated logistic probability failed to identify high leverage points correctly. However, the weakness of the DM method is that even though it can identify high leverage points correctly, it may suffer from masking and swamping effects. This situation is not desirable since that purpose of diagnostic method is to pinpoint the high leverage points correctly, after which these bad points are decided to be removed or corrected. Low leverage points have little effect on model fit. Therefore, deleting the low leverage points after diagnostic procedure may reduce the precision of parameter estimates. This work has motivated us to develop a novel detection method of high leverage points in logistic regression model, since the detection of high leverage points is one of the most important issues in regression analysis.

MATERIALS AND METHODS

Diagnostic from the mean: In this section, we will briefly review the detection of high leverage point using current method. We begin with some introduction on logistic regression model. Consider a multiple logistic regression model:

$$Y = \pi(x)+\varepsilon \tag{1}$$

Where:

$$\pi(x) = \frac{\exp(\eta)}{1+\exp(\eta)} \tag{2}$$

with $\eta = \beta_0+\beta_1X_1+\beta_2X_2+ \cdots +\beta_pX_p = X\beta$. Here, Y is an n×1 vector of response. Let, $y_i = 0$ if the ith unit does not have the characteristic and $y_i = 1$ if the ith unit does possess that characteristics. X is an n×k matrix of explanatory variables with k = p+1. $\beta^T = (\beta_0, \beta_1, \beta_2,..., \beta_p)$ is the vector of regression parameters and ε is an n×1 vector of unobserved random errors. The quantity π_i is known as probability or fitted value for the ith covariate. The model given in Eq. 2 satisfies $0\leq\pi_i\leq1$. The fitted values in logistic regression model are calculated for each covariate pattern which depend on the estimated probability for the covariate pattern, denote as $y_i = \hat{\pi}_i$. Thus, the ith residuals are defined as:

$$\hat{\varepsilon}_i = y_i - \hat{\pi}_i, i = 1, 2,..., n \tag{3}$$

Suppose that there are J distinct values of the observed x. We denote the number of cases $x = x_j$ by m_j where, j = 1, 2,..., J. We also call m_j as a number of covariate patterns. We define the number of the covariate pattern to be equal to the number of observations. The number of covariate pattern, m_j may be some number less than the number of observations, n_i if there are identical observations in the data. In linear regression model, the hat matrix plays an extremely important role in the analysis. This matrix provides the fitted values as the projection of the outcome variable into covariate space. Using weighted least squares linear regression as a model, Pregibon (1981) derived a linear approximation to the fitted values, which yields a hat matrix for logistic regression, which is:

$$H = V^{1/2}X\left(X^TVX\right)^{-1}X^TV^{1/2} \tag{4}$$

where, V is j×j diagonal matrix with element $v_j = m_j\hat{\pi}_j\left(1-\hat{\pi}_j\right)$. Thus, the diagonal elements of the hat matrix are called the leverage values shown as:

$$h_j = m_j\hat{\pi}_j\left(1-\hat{\pi}_j\right)x_j^T\left(X^TVX\right)^{-1}x_j = v_jb_j \tag{5}$$

Where:

$$b_j = x_j^T\left(X^TVX\right)^{-1}x_j \tag{6}$$

As mentioned in previous section, there is evident in the logistic regression model that the most extreme points in the covariate pattern may have the smallest leverage values. Therefore, identical method to detect the high leverage points in logistic regression model based on the leverage values in linear regression model is unsuccessful. First evident was pointed out by Hosmer and Lemeshow (2000) and then was highlighted by Imon (2006). Therefore, we are not considering leverage values in our next section. Further explanation on the disadvantages of leverage values in identifying the high leverage points in logistic regression model were well explained by Imon (2006). Imon (2006) pointed out that, in logistic regression model, a quantity that increases with the distance from the mean is denoted as b_j. He proposed to use this quantity that he called Distance from the Mean (DM) for the identification of high leverage points. He also suggested a suitable cut-off point for b_j written as:

$$b_j \geq \text{Median}(b_j) + cMAD(b_j) \quad (7)$$

where, MAD (b_j) = Median $\{|b_j\text{-Median}(b_j)|\}/0.6745$ and c is an appropriately chosen constant such as 2 or 3.This confidence bound for location parameter, which was first introduced by Hadi (1992) in regression diagnostics has been used by many authors (Imon, 2005; Imon and Apu, 2007; Habshah *et al*., 2009). This form is analogous to a confidence interval for a location and dispersion parameters where mean and standard deviation which is not robust to extreme points are replaced by median and Median Absolute Deviation (MAD), respectively as robust measures.

Robust logistic diagnostic: As already mentioned, Imon (2006) proposed a distance from the mean values for the identification of high leverage points. He has shown through some real examples that the DM values correctly identify all the high leverage points. However, there is a possibility for the DM values to swamp some low leverage points as high leverage points or to mask some high leverage points as low leverage points. The low leverage points are less harmful compared to the high leverage points depending in their outlying magnitude but elimination of the low leverage points may contribute to a loss of efficiency and precision of the parameter estimates. Therefore, we need detection techniques that can correctly identify the high leverage points and free from swamping and masking problems. The work of Imon (2006) has motivated us to propose a new improved detection method. Our new proposed method is called the Robust Logistic Diagnostic (RLGD). The RLGD method incorporates the Distance from the Mean (DM) technique proposed by Imon (2006) and the Diagnostic Robust Generalized Potentials (DRGP) method proposed by Habshah *et al.* (2009). Following the idea of Habshah *et al.* (2009), on the first stage, the suspected high leverage points are identified by robust estimator either using Minimum Covariance Determinants (MCD) or Minimum Volume Ellipsoid (MVE) (Rousseeuw, 1984). Then, diagnostic approach is employed to confirm our suspicion.

On the second stage of the RLGD method, we compute the potential based on the distance from the mean for logistic regression model. We assume that d observations among a set of n observations are deleted. Let us denote R to be a set of cases remaining in the analysis and D to be a set of cases deleted. Hence, R contains (n-d) cases after d cases are deleted. We assume that these observations are the last d rows of X, Y and V so that:

$$X = \begin{bmatrix} X_R \\ X_D \end{bmatrix}, \quad Y = \begin{bmatrix} Y_R \\ Y_D \end{bmatrix} \text{ and } V = \begin{bmatrix} V_R & 0 \\ 0 & V_D \end{bmatrix}$$

Let, $\hat{\beta}^{(-D)}$ be the corresponding vector of estimated coefficients after d cases are deleted.

The fitted values for the entire data are defined as:

$$\hat{\pi}_i^{(-D)} = \frac{\exp(x_i^T\hat{\beta}^{(-D)})}{1+\exp(x_i^T\hat{\beta}^{(-D)})}, \quad i = 1,2,...,n \quad (8)$$

Then we define the residual as:

$$\hat{\varepsilon}_i^{(-D)} = y_i - \hat{\pi}_i^{(-D)} \quad (9)$$

with corresponding variance and covariate pattern:

$$v_j^{(-D)} = m_j\hat{\pi}_j^{(-D)}\left(1-\hat{\pi}_j^{(-D)}\right), \; j = 1,2,...,J \quad (10)$$

Again, we consider Eq. 6:

$$b_j = x_j^T\left(X^TXV\right)^{-1}x_j$$

Let:

$$\tilde{X} = V^{\frac{1}{2}}X \quad (11)$$

Then, it can be shown that:

$$b_j = x_j^T\left(\tilde{X}^T\tilde{X}\right)^{-1}x_j \quad (12)$$

Thus, group deleted distance from mean based on group deleted cases D is:

$$b_j^{(-D)} = x_j^T\left(\tilde{X}_R^T\tilde{X}_R\right)^{-1}x_j \quad (13)$$

Giving a simple relationship between potential values proposed by Hadi (1992) and Eq. 13 gives:

$$b_j^{(-D+j)} = x_j^T\left(\tilde{X}_R^T\tilde{X}_R + x_jx_j^T\right)^{-1}x_j = \frac{b_j^{(-D)}}{1+b_j^{(-D)}} \quad (14)$$

Based on group deleted cases indexed by D, by adopting distance from mean, let us define the group deleted potential denoted by:

$$p_{jj}^{*(-D)} = \begin{cases} b_j^{(-D)}/1+b_j^{(-D)} & ; j \in R \\ b_j^{(-D)} & ; j \in D \end{cases} \quad (15)$$

Since, the distribution of $p^{*}_{jj}{}^{(-D)}$ is unknown, we apply cut-off point based on median and MAD for $p^{*}_{jj}{}^{(-D)}$ as suggested by Hadi (1992). Hence, any observation corresponding to excessively large potential values with cut-off point:

$$p_{jj}^{*(-D)} > \text{Median}\,(p_{jj}^{*(-D)}) + \text{cMAD}(p_{jj}^{*(-D)}) \quad (16)$$

where:

$$\text{MAD}\left(p_{jj}^{*(-D)}\right) = \text{Median}\left\{\left|p_{jj}^{*(-D)} - \text{Median}\left(p_{jj}^{*(-D)}\right)\right|\right\} / 0.6745$$

shall be declare as high leverage point. The RLGD method is summarized as follows:

- **Step 1:** For each ith point, compute RMD_i using either MCD or MVE estimators
- **Step 2:** An ith point with RMD_i>Median (RMD_i) + cMAD (RMD_i), are suspected as high leverage points and included in the deleted set D. The remaining points are put into the set R
- **Step 3:** Based on the above set D and R, compute the $p^{*}_{jj}{}^{(-D)}$
- **Step 4:** Any deleted points with $p^{*}_{jj}{}^{(-D)}$>Median ($p^{*}_{jj}{}^{(-D)}$) + cMAD ($p^{*}_{jj}{}^{(-D)}$), are finalized and declared as high leverage points

RESULTS

We investigate the usefulness of the proposed RLGD method on several well-known real data and compared the results with the DM method.

The prostate cancer data: We first consider the Prostate Cancer (PC) data by Brown *et al*. (1980). Here the main objective was to see whether two continuous variables which are an elevated level of acid phosphates (AP) in the blood serum and age of patients (AGE) together with three categorical variables (X-RAY, STAGE and GRADE) would be of value for predicting whether or not PC patients also had lymph node involvement (LNI). It has been reported by Imon (2006) that the original data on the 53 patients may contain three high leverage points (case 24, 25 and 53). Here the response is nodal involvement with Y = 1 denoting the presence of nodal involvement and Y = 0 indicating the absence of such involvement.

The character plot of the PC data is presented in Fig. 1 where, AP is plotted against AGE and the character corresponding to occurrence Y = 1 and non occurrence Y = 0 are denoted by symbols triangle and circle, respectively.

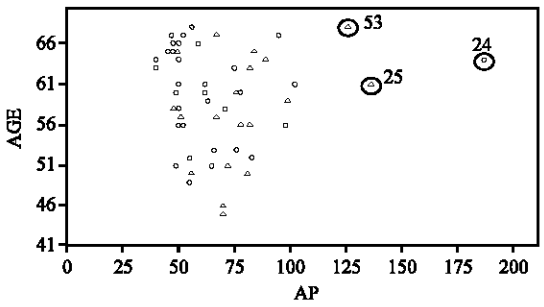

Fig. 1: Scatter plot of AP vs. AGE for PC data

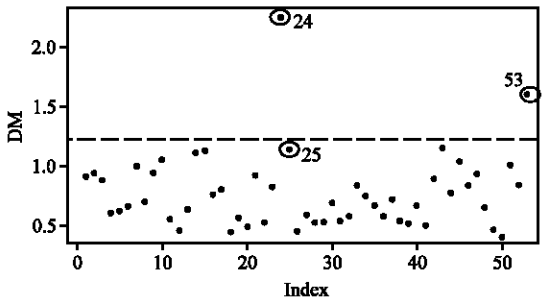

Fig. 2: Index plot of DM for PC data

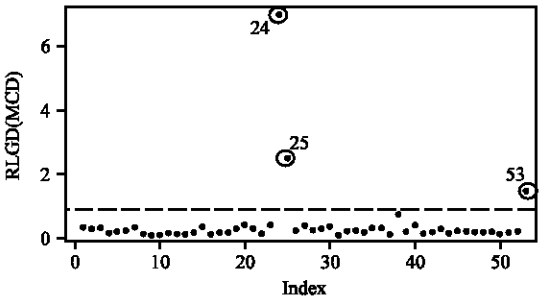

Fig. 3: Index plot of RLGD(MCD) for PC data

Figure 2 and 3 show the index plot of DM and RLGD(MCD) for PC data. To confirm these suspected high leverage points, we shall apply the DM method and the RLGD method. The DM and RLGD values are presented in Table 1.

The vaso-constriction skin digits data: We now consider another data as real example given by Finney (1947). The original data was obtained to study the effect of the two continuous variable (RATE) and (VOL) of air inspired on a transient Vaso-constriction in the Skin of the Digits (VSD) as a binary response. The nature of the measurement process was such that only the occurrences and non-occurrence of VSD could be reliably measured. There are many versions of VSD data available in the literature. These data consist of 39 observations. We

Table 1: High leverage points diagnostics for PC data

	Cut-off points		
	1.2215	1.3977	0.796
		RLGD	RLGD
ID	DM	MCD	MVE
1	0.9133	0.3752	0.3215
2	0.9408	0.3071	0.2702
3	0.8859	0.3492	0.3010
4	0.6092	0.1706	0.1538
5	0.6283	0.2180	0.1924
6	0.6647	0.2616	0.2276
7	0.9998	0.3849	0.3281
8	0.7040	0.1567	0.1409
9	0.9495	0.1156	0.0961
10	1.0543	0.1119	0.0927
11	0.5523	0.1642	0.1486
12	0.4606	0.1857	0.1379
13	0.6420	0.1740	0.1219
14	1.1124	0.1960	0.1787
15	1.1332	0.4047	0.3459
16	0.7650	0.1342	0.1214
17	0.8012	0.1885	0.1680
18	0.4507	0.2429	0.1707
19	0.5643	0.4069	0.2788
20	0.4927	1.2239	0.4057
21	0.9218	0.3402	0.2949
22	0.5296	0.1950	0.1512
23	0.8400	1.1208	0.3854
24	2.2609	10.5537	6.0047
25	1.1496	3.8698	2.1711
26	0.4501	0.3402	0.2301
27	0.5971	0.4472	0.3776
28	0.5295	0.2658	0.2317
29	0.5265	0.3157	0.2730
30	0.6928	0.4313	0.3635
31	0.5318	0.1205	0.1064
32	0.5802	0.2432	0.2185
33	0.8430	0.2410	0.2100
34	0.7490	0.1929	0.1722
35	0.6739	0.3454	0.2971
36	0.5806	0.3585	0.3073
37	0.7230	0.1497	0.1317
38	0.5412	1.1788	0.3975
39	0.5155	0.2930	0.1955
40	0.6689	0.6246	0.3688
41	0.5003	0.1699	0.1226
42	0.8928	0.2678	0.1913
43	1.1560	0.4053	0.2769
44	0.7664	0.2207	0.1586
45	1.0401	0.3116	0.2565
46	0.8382	0.2845	0.1921
47	0.9378	0.2975	0.1946
48	0.6504	0.1960	0.1787
49	0.4660	0.2624	0.1857
50	0.4031	0.1598	0.1251
51	1.0089	0.2613	0.1732
52	0.8452	0.5150	0.2357
53	1.6041	2.3174	1.2915

replace the variable (RATE) for the cases 32 as 0.300 instead of 0.030 (Pregibon, 1981; Kunsch *et al*., 1989).

The character plot of the VSD data is presented as shown in Fig. 4 where VOL is plotted against RATE and the character corresponding to occurrence Y = 1 and non occurrence Y = 1 are denoted by symbols triangle and circle, respectively.

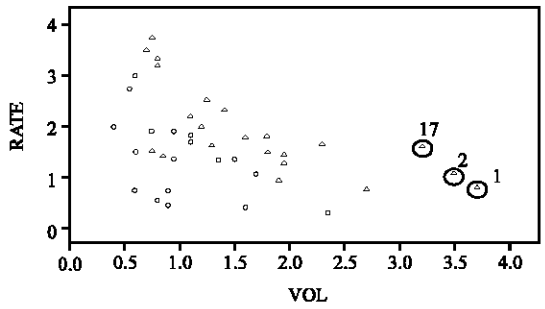

Fig. 4: Scatter plot of VOL vs. RATE for VSD data

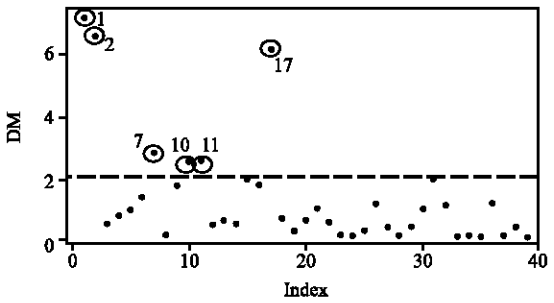

Fig. 5: Index plot of DM for VSD data

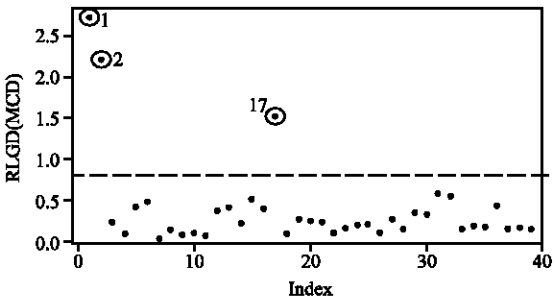

Fig. 6: Index plot of RLGD(MCD) for VSD data

Figure 5 and 6 show the index plot of DM and RLGD(MCD) for VSD data. To confirm these suspected high leverage points, we shall apply the DM method and the RLGD method. The DM and RLGD values are shown in Table 2.

The erythrocyte sedimentation rate data: Our final real data is Erythrocyte Sedimentation Rate (ESR) data. Here the main objective is to see whether the levels of two plasma protein (i.e., fibrinogen and γ-globulin) in blood plasma would be the factor to increase the ESR for healthy individual. The study was carried out by Institute of Medical Research, Kuala Lumpur, Malaysia involving 32 patients and the original data were collected by Collett and Jemain (1985). Here, the continuous variables are

Table 2: High leverage points diagnostics for VSD data

	Cut-off points		
	2.141	0.8074	0.8074
ID	DM	RLGD MCD	RLGD MVE
1	7.1683	2.7303	2.7303
2	6.5763	2.2240	2.2240
3	0.6162	0.2400	0.2400
4	0.8963	0.1021	0.1021
5	1.0582	0.4167	0.4167
6	1.4585	0.4867	0.4867
7	2.8867	0.0374	0.0374
8	0.2441	0.1349	0.1349
9	1.8504	0.0848	0.0848
10	2.6006	0.1082	0.1082
11	2.6257	0.0736	0.0736
12	0.5932	0.3692	0.3692
13	0.7287	0.4106	0.4106
14	0.6260	0.2192	0.2192
15	2.0529	0.5212	0.5212
16	1.8620	0.3963	0.3963
17	6.1514	1.5332	1.5332
18	0.8001	0.0938	0.0938
19	0.3943	0.2784	0.2784
20	0.7416	0.2572	0.2572
21	1.1046	0.2364	0.2364
22	0.6838	0.0965	0.0965
23	0.2756	0.1630	0.1630
24	0.2563	0.1988	0.1988
25	0.4079	0.2127	0.2127
26	1.2576	0.1091	0.1091
27	0.5140	0.2705	0.2705
28	0.2798	0.1542	0.1542
29	0.5389	0.3509	0.3509
30	1.0896	0.3316	0.3316
31	2.0737	0.5783	0.5783
32	1.2135	0.5494	0.5494
33	0.2090	0.1476	0.1476
34	0.2588	0.1965	0.1965
35	0.2121	0.1713	0.1713
36	1.2708	0.4410	0.4410
37	0.2798	0.1542	0.1542
38	0.5035	0.1651	0.1651
39	0.1962	0.1544	0.1544

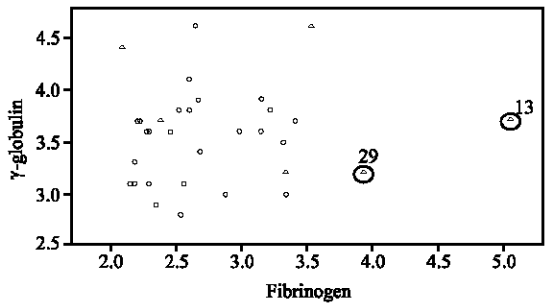

Fig. 7: Scatter plot of fibrinogen vs. γ-globulin for ESR data

(fibrinogen and γ-globulin) versus the binary response of ESR.

The character plot of the ESR data is presented as shown in Fig. 7 where, fibrinogen is plotted against γ-globulin and the character corresponding to occurrence Y = 1 and non occurrence Y = 0 are denoted by symbols triangle and circle, respectively.

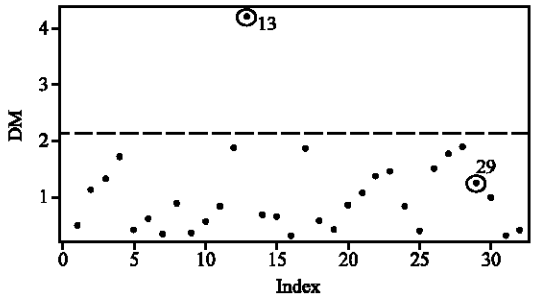

Fig. 8: Index plot of DM for ESR data

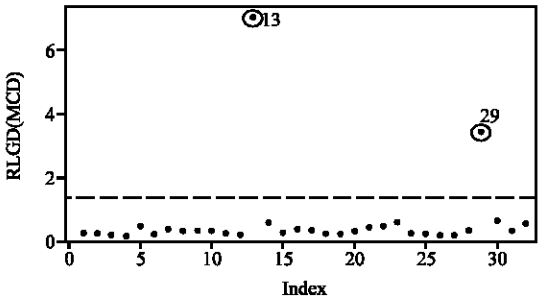

Fig. 9: Index plot of RLGD(MCD) for ESR data

Figure 8 and 9 show the index plot of DM and RLGD(MCD) for ESR data. To confirm these suspected high leverage points, we shall apply both the DM method and the RLGD method. Table 3 shows the high leverage points diagnostics for the ESR data.

Monte carlo simulation study: A simulation study is conducted to further assess the performance of the RLGD method and the DM method. Following, Croux and Haesbroeck (2003) work, three different types of data are considered namely the uncontaminated (Type 1), 5% moderate contaminated of high leverage points (Type 2) and 5% extreme contaminated of high leverage points (Type 3). Explanatory variables for uncontaminated data are generated according to a standard normal distribution $x_1 \sim N(0, 1)$ and $x_2 \sim N(0, 1)$ with number of observations, n = 100. Setting the true parameters as $\beta = (\beta_0, \beta_1, \beta_2)^T = (0.5, 1, -1)^T$ and the response is defined as the following model equations:

$$y_i = \begin{cases} 0 & \text{if } \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_i < 0 \\ 1 & \text{if } \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_i \geq 0 \end{cases} \quad (17)$$

where, the error terms is generated according to a logistic distribution, $\varepsilon_i \sim \Lambda$ (0, 1). The explanatory variables are

Table 3: High leverage points diagnostics for ESR data

	Cut-off points		
	2.1398	1.7399 RLGD	1.7399 RLGD
ID	DM	MCD	MVE
1	0.5029	0.2908	0.2908
2	1.1461	0.3404	0.3404
3	1.3224	0.2360	0.2360
4	1.7103	0.2100	0.2100
5	0.4322	0.5869	0.5869
6	0.6192	0.2624	0.2624
7	0.3432	0.4752	0.4752
8	0.8955	0.3510	0.3510
9	0.3616	0.4177	0.4177
10	0.5817	0.3482	0.3482
11	0.8319	0.2897	0.2897
12	1.8937	0.2879	0.2879
13	4.2287	9.3067	9.3067
14	0.6935	0.6777	0.6777
15	0.6648	0.2943	0.2943
16	0.3153	0.4967	0.4967
17	1.8564	0.7611	0.7611
18	0.5789	0.3225	0.3225
19	0.4379	0.2838	0.2838
20	0.8656	0.3430	0.3430
21	1.0788	0.5413	0.5413
22	1.3765	0.4917	0.4917
23	1.4669	0.6425	0.6425
24	0.8461	0.2925	0.2925
25	0.4134	0.2944	0.2944
26	1.5190	0.2284	0.2284
27	1.7664	0.2085	0.2085
28	1.8906	0.4257	0.4257
29	1.2513	4.4955	4.4955
30	1.0171	0.7175	0.7175
31	0.3058	0.4187	0.4187
32	0.4123	0.6002	0.6002

generated according to a standard normal distribution z_1~N (0, 1) and z_2~N (0, 1). We considered different percentage of contamination denoted as s, such that s = (50, 10, 15 and 20%) with magnitude of outlying shift distance in X-space for Type 2 and 3 are taken as $\delta = 5$ and $\delta = 10$, respectively. The new x values are defined as $x^*_1 = z_1 + \delta$, $x^*_2 = z_2 - \delta$ and the response is defined as the following model equations:

$$y_i^* = \begin{cases} 0 & \text{if } \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \varepsilon_i \geq 0 \\ 1 & \text{if } \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \varepsilon_i < 0 \end{cases} \quad (18)$$

The performance of the DM method and the RLGD method is evaluated based on the probability of the Detection Capability (DC) and the False Alarm Rate (FAR) (Kudus *et al*., 2008). These measures are computed over M = 1000 replications. The FAR is the probability of swamping occur and the DC is probability of masking occur with c = 3 for cut-off point median and MAD. After we apply the diagnostic procedures (the DM and RLGD methods) the high leverage points are assigned with weights w_i of 1 and 0 for otherwise. Let say, $\bar{w}_1$ is average for weights of the uncontaminated, $\bar{w}_1 = \sum_{i=1}^{n} w_i / n$ and $\bar{w}_2$ is average for weights of the contaminated, $\bar{w}_2 = \Sigma_{j=1}^{s} w_j / s$ yielding $FAR = \sum_{i=1}^{M} \bar{w}_1 / M$ and $DC = \sum_{i=1}^{M} \bar{w}_2 / M$. A good diagnostic method is the one which has probability of the FAR closest to 0 and the DC closest to 1. The higher probability of FAR shows that many low leverage points are swamped after the diagnostic method. Meanwhile for the diagnostic method that masked some high leverages, the probability of DC will be less than 1. Smaller probability of the DC suggests that the detection method fail to identify the high leverage points correctly because of the high leverage points are masked. Simulation result on the identification of the high leverage points based on the DM method and the RLGD method are shown in Table 4 and 5.

Table 4: The measures of performance on the diagnostic methods on moderate contamination

		Moderate contamination				
Detection methods	Measures of performance	0%	5%	10%	15%	20%
DM	FAR	0.1300	0.1292	0.0932	0.0687	0.0507
	DC	-	0.9996	0.9807	0.8911	0.7359
RLGD	FAR	0	0.0329	0.0224	0.0133	0.0077
MVE	DC	-	0.9998	0.9997	0.9990	0.9974
RLGD	FAR	0	0.0336	0.0228	0.0141	0.0083
MCD	DC	-	0.9996	0.9996	0.9993	0.9976

Table 5: The measures of performance on the diagnostic methods on extreme contamination

		Moderate contamination				
Detection methods	Measures of performance	0%	5%	10%	15%	20%
DM	FAR	0.1300	0.1652	0.1397	0.1171	0.0977
	DC	-	1	1	1	1
RLGD	FAR	0	0.0331	0.0227	0.0137	0.0082
MVE	DC	-	1	1	1	1
RLGD	FAR	0	0.0338	0.0231	0.0145	0.0090
MCD	DC	-	1	1	1	1

DISCUSSION

The scatter plot of AP versus AGE (Fig. 1) clearly shows that observations 24, 25 and 53 may severely distort the covariate pattern. They may be considered as high leverage points. We first apply the DM method proposed by Imon (2006). The DM method gives the upper cut-off point 1.2215 when, the constant c is set as 2. It is very important to point out that the case 25 of the DM was found to be 1.1496 instead of 1.2496 as reported by Imon (2006). We have double checked this result and confirmed that the DM value of case 25 should be 1.1496. Therefore, we can say that the DM method can identify cases 24 and 53 correctly but masks case 25 (Table 1). From Fig. 2, the index plot of DM clearly shows that cases 25 is masked by another high leverage points although the value for case 25 where, AP = 136 are larger than the value for case 53, AP = 126 (Brown *et al*., 1980). Now, we

apply our newly proposed methods namely, the RLGD (MCD) and RLGD (MVE) to identify the high leverage points for this data. Let us first focus to the result of the first stage of our proposed methods. When, we employ the RMD based on MVE and MCD with the constant c is set as 2, we identify more than three cases as suspected high leverage points. Based on RMD(MCD), we identify cases 20, 23, 24, 25, 38, 40, 52 and 53 as suspected high leverage points. From RMD(MVE), we identify cases 24, 25, 40, 45 and 53 as suspected high leverage points. These results show that the MVE estimator detect less suspected high leverage points compared to the MCD estimator. Perhaps, if we set the constant c as 3, we may identify these three high leverage points correctly in the first stage for the MVE and MCD estimators. Then we perform the deletion set D with the suspected high leverage points. We compute the group deleted potential for the whole set based on D and the results are presented in Table 1. We observed from this table, that the RLGD(MCD) and RLGD(MVE) values are much larger than the cut-off point for cases 24, 25 and 53 which reveal that these three points are high leverage points. Similar conclusion may be drawn from the index plot of RLGD(MCD) as shown in Fig. 3. All these three suspected cases are clearly separated from the rest of the data.

The scatter plot of VOL versus RATE (Fig. 4) clearly shows that observations 1, 2 and 17 may severely distort the covariate pattern. They may be considered as high leverage points for the variable VOL. The index plot of DM shows that, this diagnostic method can identify cases 1, 2 and 17 as the high leverage point but also swamps low leverage of cases 7, 10 and 11. Here, the deletion set D contains cases 1, 2 and 17 based on RMD(MCD) and RMD(MVE). Contrary from the DM method, after re-estimated the model with these cases of D and computed the RLGD values for the whole set, our proposed method (Table 2) showed better result by detecting three suspected high leverage points for the cases 1, 2 and 17 correctly as we suspected previously as shown in Fig. 4. Similar conclusions on index plot of the DM values and RLGD(MCD) values may be shown from Fig. 5 and 6.

The scatter plot of fibrinogen versus γ-globulin (Fig. 7) clearly suggests that observations 13 and 29 may be outlying in the covariate space of fibrinogen. We observe from this table that the DM method can only detect single suspected high leverage point for case 13 and masks the case 29. The robust diagnostic based on RMD(MCD) and RMD(MVE) at first stage identify cases 13, 17 and 29 as suspected high leverage points. Then, we apply the RLGD method for the entire data based on a set that excluded these suspected high leverage points and the result are presented in Table 3. As to be expected, the RLGD values corresponding to cases 13 and 29 exceed the cut-off point and hence can be successfully identified as high leverage points. We observed similar picture when we look at the index plot of DM values and RLGD (MCD) values as shown in Fig. 8 and 9.

A good method of identifying the high leverage points is the method which performs the Detection Capability (DC) closer or exactly 1. If the weight of contaminated contains value 0 (suppose to be all 1) during simulation, therefore, $\bar{w}_2$ is not equal to 1. A good method of identifying the high leverage points also indicates that False Alarm Rate (FAR) should be closer or exactly 0. If the weight of uncontaminated contains value 1 (suppose to be all 0) during simulation, therefore, $\bar{w}_1$ is not equal to 0. Refer to Table 4 and 5, in general, all detection method give good results. In the moderate and extreme contamination, the DM method performs less efficient compared to the RLGD method. Even when there is no high leverage point, the DM method show probability of FAR equals to 0.13. When the percentage of high leverage points increases, the DM method masked some of the high leverages points and swamped less low leverage points. The RLGD method based on the MVE estimator performs slightly better compared to the RLGD method based on the MCD estimator where the most of the high leverage points are identified correctly and swamped very less of the low leverage points.

CONCLUSIONS

Here, we establish the fact in the logistic regression the DM method which was proposed by Imon (2006) suffered from the masking and swamping problems. Therefore, the DM method may not be able to identify the high leverage points correctly. Then we proposed a new method for the identification of high leverage points, namely the Robust Logistic Diagnostic (RLGD) which incorporates both robust methods in their first and second stages. The advantages of the RLGD method give us an idea by showing the suspected high leverage points at the first stage that may consist of low leverage points as well as high leverage points. Finally, we confirm the high leverage points at the second stage. The setting of constant for the cut-off point is also important in order to swamp less low leverages depending on how far the cases are outlying in the covariate space and the number of the regressor variables that exists in the data. The numerical examples signify that the RLGD method is proven to be very effective in the identification of high leverage points when the DM method is less effective. The RLGD method have better detection probability and have false alarm rate

up to 20% of contamination in the data. In general, the RLGD method is more efficient in identifying the high leverage points compared to the DM method.

REFERENCES

Bedrick, E.J. and J.R. Hill, 1990. Outlier tests for logistic regression: A conditional approach. Biometrika, 77: 815-827.

Brown, B.W., R.G. Miller, B. Efron and L.E. Moses, 1980. Prediction Analyses for Binary Data. In: Biostatistics Casebook, Miller, R.G., B. Efron, B.W. Brown and L.E. Moses (Eds.). John Wiley and Sons, Inc., New York, pp: 3-18.

Collett, D. and A.A. Jemain, 1985. Residuals, outliers and influential observations in regression analysis. Sains Malaysiana, 14: 493-511.

Croux, C. and G. Haesbroeck, 2003. Implementing the bianco and yohai estimator for logistic regression. Comput. Statist. Data Anal., 44: 273-295.

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

Habshah, M., M.R. Norazan and A.H.M.R. Imon, 2009. The performance of diagnostic-robust generalized potentials for the identification of multiple high leverage points in linear regression. J. Applied Statist., 36: 507-520.

Hadi, A.S., 1992. A new measure of overall potential influence in linear regression. Comput. Statist. Data Anal., 14: 1-27.

Hadi, A.S. and J.S. Simonoff, 1993. Procedure for the identification of multiple outliers in linear models. J. Am. Statistical Assoc., 88: 1264-1272.

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

Hosmer, W.D. and S. Lemeshow, 2000. Applied Logistic Regression. 2nd Edn., John Wiley and Sons, Inc., New York, pp: 392.

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

Imon, A.H.M.R., 2006. Identification of high leverage points in logistic regression. Pak. J. Statist., 22: 147-156.

Imon, A.H.M.R. and M.R. Apu, 2007. Identification of multiple high leverage points using robust mahalanobis distance. J. Statist. Stud., 32: 929-946.

Imon, A.H.M.R. and A.S. Hadi, 2008. Identification of multiple outliers in logistic regression. Commun. Statist. Theory Meth., 37: 1697-1709.

Jennings, D.E., 1986. Outliers and residual distributions in logistic regression. J. Am. Statist. Assoc., 81: 987-990.

Kudus, A., I.N. Akma and D. Isa, 2008. Simulation on group deleted generalized potentials for diagnostics of censored survival regression. Proceedings of the 16th National Mathematical Science Symposium, June 2-5, Kota Bharu, Kelantan, Malaysia, pp: 280-287.

Kunsch, H.R., L.A. Stefanski and R.J. Carroll, 1989. Conditionally unbiased bounded influence estimation in general regression models with applications to generalized linear models. J. Am. Statist. Assoc., 84: 460-466.

Munier, S., 1999. Multiple outlier detection in logistic regression. J. Statist. Stud., 3: 117-126.

Nurunnabi, A.A.M., A.H.M.R. Imon and M. Nasser, 2009. Identification of Multiple Influential Observations in Logistic Regression. Taylor and Francis, Inc., Philadelphia, PA.

Pregibon, D., 1981. Logistic regression diagnostics. Ann. Statist., 9: 705-724.

Rousseeuw, P.J., 1984. Least median of squares regression. J. Am. Statist. Assoc., 79: 871-880.

Rousseeuw, P.J., 1991. A diagnostic plot for regression outliers and leverage points. J, Comput. Statist. Data Anal., 11: 127-129.

Vellman, P.F. and R.E. Welsch, 1981. Efficient computing of regression diagnostics. J. Am. Statist., 27: 234-242.

Williams, D.A., 1987. Generalized linear model diagnostics using the deviance and single case deletions. J. Applied Statist., 36: 181-191.