World Applied Sciences Journal 14 (8): 1125-1130, 2011
ISSN 1818-4952
© IDOSI Publications, 2011

Diagnostics for Residual Outliers Using Deviance Component in Binary Logistic Regression

[1]*Sanizah Ahmad*, [2,3]*Habshah Midi and* [1]*Norazan Mohamed Ramli*

[1]Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Shah Alam, Malaysia
[2]Department of Mathematics, Universiti Putra Malaysia, Serdang, Malaysia
[3]Institute for Mathematical Research, Universiti Putra Malaysia, Serdang, Malaysia

Abstract: Detection of outliers based on residuals has received great interest in logistic regression. These methods like Pearson residuals and deviance residuals are only reliable for identifying a single outlier but fails for multiple outliers due to the masking and swamping problems. Therefore it is necessary to detect these outliers and take appropriate measures to obtain a good fit. In this study, we developed a new diagnostic method on the identification of residual outliers in logistic regression based on deviance component. The performance of the proposed diagnostic method is investigated through numerical examples and Monte Carlo simulation study. The result indicates that the proposed method manages to correctly identify all the outliers.

Key words: Outliers · Pearson residuals · Deviance residuals · Masking · Simulation

INTRODUCTION

Diagnostic methods on detecting outliers are commonly used in all branches of regression analysis including logistic regression. In recent years, diagnostics has become an essential part of logistic regression [1, 2]. It is well known the estimation of maximum likelihood estimator (MLE) can be severely affected in the presence of outliers. It is often observed outliers greatly affect the covariate pattern and consequently their presence will give misleading interpretations. Detection of outliers based on residuals received great attention in linear regression [3] and logistic regression [4-6]. Some of the diagnostic methods available for the identification of outliers in logistic regression are Pearson residuals and deviance residuals. These methods, however, are only able to identify a single outlier. If the data contains multiple outliers, these methods fail to detect them due to the masking and swamping problems [7, 8]. Therefore it is necessary to detect these outliers and take appropriate measures to obtain a good fit. In this paper, we focus mainly on residual outliers. First we define on residual outliers in logistic regression and discuss on commonly used diagnostics. Next we propose a new diagnostic method on the identification of residual outliers based on deviance component. The performance of the proposed method is investigated through numerical example and simulation experiment. Throughout this paper, residual outliers will also be termed as outliers.

MATERIALS AND METHODS

Outliers in Logistic Regression: Given a binary response variable Y and a $p \times 1$ vector X of independent variables, the logistic regression model is of the form.

$$\pi(x_i) = \exp\left(x_i^T\beta\right) \Big/ \left[1 + \exp\left(x_i^T\beta\right)\right],\ i = 1,\cdots,n \tag{1}$$

With $\beta^T = (\beta_0,\beta_1,...,\beta_p)$ being the vector of parameters. The response variable is denoted by y_i = 1 or 0 with probabilities π_i and $1-\pi_i$, respectively. Classically one uses the maximum likelihood method based on iterative reweighted least squares to get the estimated coefficients $\hat{\beta}$. Then we obtain the fitted values, $\hat{\pi}(x_i)$ where $0 \leq \pi(x_i) \leq 1$, by substituting $\hat{\beta}$ in expression (1).

Measures of agreement between an observation on a response variable and the corresponding fitted value are known as *residuals* [2]. Unlike ordinary regression, residual outliers in logistic regression have to be rethought in the context of binary data in which all the *y*s are 0 or 1. An error in the y direction can only occur as a transposition $0 \rightarrow 1$ or $1 \rightarrow 0$ [6]. In binary data, it is possible to encounter outliers when the values of the explanatory variables are not extreme. This type of outlier is also

Corresponding Author: Sanizah Ahmad, Faculty of Computer and Mathematical Sciences,
Universiti Teknologi MARA 40450 Shah Alam, Selangor Darul Ehsan, Malaysia.
Tel: +603-55435458, Fax: +603-55435501.

known as y-outlier or residual outlier or misclassified observation. An outlier occurs when $y = 1$ and the corresponding fitted probability is zero, or when $y = 0$ and the fitted probability is unity.

Residual Analysis for Outlier Measures: Many detection methods have been proposed for identifying outliers in logistic regression base on residual measures [1, 2, 4-6]. If we use the linear regression-like approximation [4] for the ith covariance pattern, it is observed that the ith residual is defined as:

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

In logistic regression, the residual is important in detecting ill-fitting points [7] but the residuals defined in (1) are unscaled, hence not applicable in detecting outliers. The scaled version is known as the Pearson residual (PR) with the general form of.

$$r_{Pi} = \left(y_i - \hat{\pi}_i\right)\Big/\sqrt{\hat{\pi}_i\left(1-\hat{\pi}_i\right)},\; i = 1,2,...,n \tag{2}$$

An observation is declared a residual outlier if its corresponding Pearson residual exceeds the value 3 in absolute term [9, 7], that matches with the 3σ distance rule used in the normal theory. A better procedure is to have approximate unit variance by dividing (2) with the standard error given by $se(y_i - \hat{\pi}_i) = \sqrt{v_i(1-h_i)}$ where $v_i = \hat{\pi}_i(1-\hat{\pi}_i)$ and h_i is the ith diagonal element of the $n \times n$ matrix $H = V^{1/2}X(X^TVX)^{-1}X^TV^{1/2}$. V is a diagonal matrix with diagonal elements v_i. Therefore, the resulting standardized Pearson residuals (SPR) is defined as:

$$r_{SPi} = \left(y_i - \hat{\pi}_i\right)\Big/\sqrt{v_i\left(1-h_i\right)},\; i = 1,2,...,n \tag{3}$$

Another type of residual constructed from the deviance which is useful for identifying potential outliers is the deviance residuals (DR) given by

$$D = 2\sum_{i=1}^{n}\left\{y_i\log\left(y_i/\hat{\pi}_i\right) + \left(1-y_i\right)\log\left[\left(1-y_i\right)/\left(1-\hat{\pi}_i\right)\right]\right\}^2 \tag{4}$$

The ith individual component is

$$r_{D_i} = \pm\sqrt{-2\left[y_i\log\hat{\pi}_i + \left(1-y_i\right)\log\left(1-\hat{\pi}_i\right)\right]}. \tag{5}$$

In particular, $r_{D_i} = -\sqrt{-2\log(1-\hat{\pi}_i)}$ if $y_i = 0$ and $r_{D_i} = \sqrt{-2\log\hat{\pi}_i}$ if $y_i = 1$. For all of the above methods, the ith observation may be declared as an outlier if the absolute residual measure is greater than 3.

A New Method for the Identification of Outliers: For binary data, an outlier occurs when $y_i = 1$ and the corresponding fitted probability is near zero, or when $y_i = 0$ and the fitted probability is near unity [2]. It is observed from (4) there is a component in the ith deviance residual which can be expressed as:

$$d\left(y_i, \hat{\pi}_i\right) = \begin{cases} 2\log\left[1/\left(1-\hat{\pi}_i\right)\right] \text{ if } y_i = 0 \\ 2\log\left(1/\hat{\pi}_i\right) \text{if } y_i = 1 \end{cases} \tag{6}$$

Where $0 \le \hat{\pi}_i < 1$ for $y_i = 0$ and $0 < \hat{\pi}_i \le 1$ for $y_i = 1$. In this paper, we call $d(y_i, \hat{\pi}_i)$ as the deviance component (DEVC). The value of $d(y_i, \hat{\pi}_i)$ is always nonnegative. It is observed that this deviance component is useful for identifying outliers. If $y_i = 1$ and $\hat{\pi}_i \to 0$, then $d(y_i, \hat{\pi}_i) = 2\log(1/\hat{\pi}_i)$ is large. If $y_i = 0$ and $\hat{\pi}_i \to 1$, then $d(y_i, \hat{\pi}_i) = 2\log\left[1/(1-\hat{\pi}_i)\right]$ is large. Here we propose a cut-off point for the identification of outliers using DEVC. One could consider DEVC to be large if:

$$\text{DEVC}_i > \text{Median}(\text{DEVC}_i) + 3.\ \text{MAD}\ (\text{DEVC}_i) \tag{7}$$

This type of cut-off value in (7) is analogous to a confidence bound for a location parameter, first introduced by [10] in regression diagnostics and later used by many others in linear regression [11, 8] and logistic regression [12, 7].

RESULTS AND DISCUSSIONS

We consider two numerical examples to investigate on the performance of the proposed diagnostic DEVC method with existing outlier measures for the identification of residual outliers in logistic regression.

Vaso-Skin Data: This is a well-known data set which was first introduced by [13] and later studied by [14, 4] and many others. This data, consisting 39 observations, was obtained in a carefully controlled study in human physiology where a reflex "vaso- constriction" may occur in the skin of the digits after taking a single deep breath.

The response variable y_i is the presence ($y_i = 1$) or absence ($y_i = 0$) of Vaso-constriction of the skin of the digits after air inspiration. The two explanatory variables are: x_1 the volume of air inspired and x_2 the inspiration rate (both in logarithms). Figure 1 provides a scatter plot

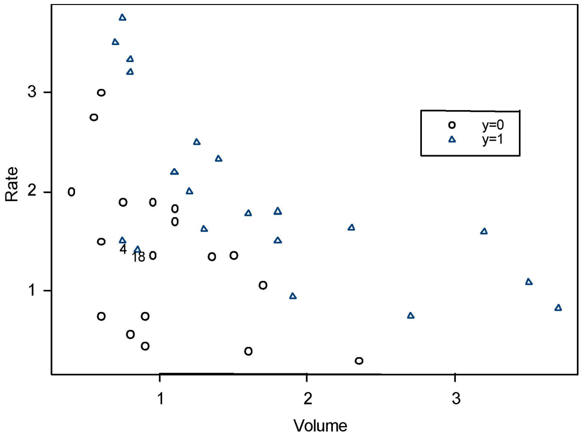

Fig. 1: Scatter plot of Vaso-skin Data

of the Vaso-skin data. [4] pointed out that this data set might contain two outliers (cases 4 and 18). However, it is quite difficult to identify cases 4 and 18 as outliers by looking at the scatter plot. When calculating the fitted probabilities, both cases (with y = 1) gives the values 0.07257 and 0.10297, respectively, which are considered closed to zero. This confirms both cases as residual outliers. Applying the newly proposed DEVC method, Table 1 presents the values of the Pearson and standardized Pearson residuals, deviance and the standardized deviance residuals and DEVC method for the Vaso-skin data. It seems that DR and SDR fail to identify any residual outliers in the data and PR only manage to identify case 4 as outlier. Both SPR and DEVC correctly identify both cases 4 and 18 as the outliers. The index plot in Figure 2 shows that DEVC can correctly and clearly identify the outliers compared to SPR where case 18 is slightly above the cut-off value.

ESR Data: The erythrocyte sedimentation rate (ESR) is the rate at which red blood cells (erythrocytes) settle out of suspension in blood plasma, when measured under standard conditions. The ESR data, collected by [15], is a study carried out by the Institute of Medical Research, Kuala Lumpur, Malaysia to examine the extent to which the ESR is related to two plasma proteins (explanatory variables): x_1 the level of fibrinogen (in gm/liter) and x_2 the level of gamma-globulin (in gm/liter), for a sample of 32 individuals. A healthy individual should have an erythrocyte sedimentation rate (ESR) less than 20mm/hour. The value of ESR is not that important, so the response variable is just

$$y_i = \begin{cases} 1 \text{ if ESR} < 20 \text{ or healthy} \\ 0 \text{ if ESR} \geq 20 \text{ or unhealthy} \end{cases}.$$

The scatter plot of the ESR data in Figure 3 does not clearly display the outliers. Table 2 suggests that DR and SDR are unable to identify any outliers, PR and SPR able to identify case 15 and DEVC identify cases 14, 15, 23 and 29. This shows that the existing diagnostic methods tends to mask the outliers when there are multiple of them. The index plots in Figure 4 clearly show DEVC performs best in detecting residual outliers for ESR data.

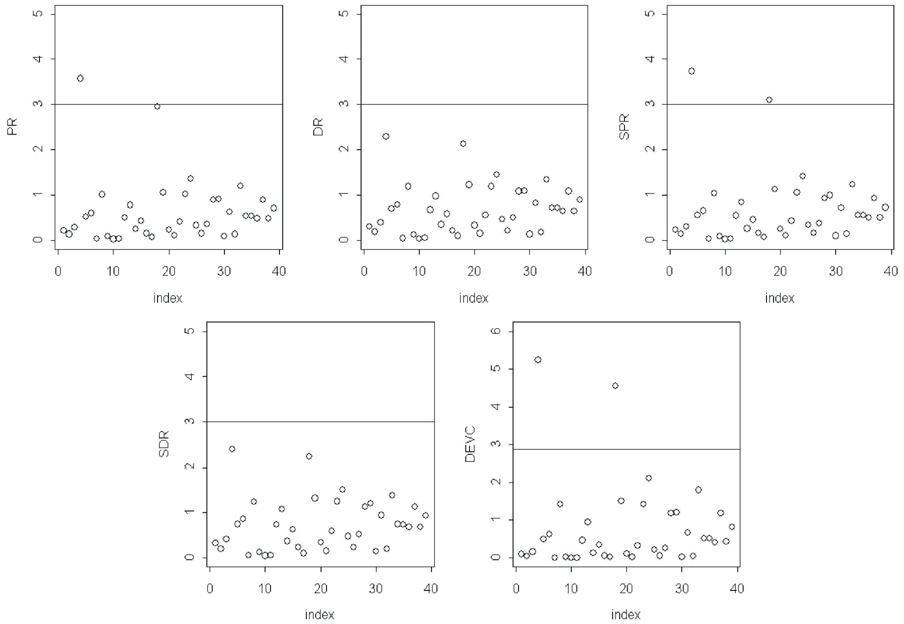

Fig. 2: Index plots of PR, DR, SPR, SDR and DEVC for Vaso-skin data

Table 1: Outlier Diagnostics for the Vaso-skin Data

	Cut-off						Cut-off				
	3.00	3.00	3.00	3.00	1.58		3.00	3.00	3.00	3.00	1.58
Obs	PR	DR	SPR	SDR	DEVC	Obs	PR	DR	SPR	SDR	DEVC
1	0.2214	0.3094	0.2326	0.3250	0.0957	21	0.1055	0.1489	0.1075	0.1516	0.0222
2	0.1343	0.1891	0.1373	0.1933	0.0358	22	0.4132	0.5615	0.4345	0.5905	0.3153
3	0.2887	0.4002	0.2976	0.4125	0.1601	23	1.0165	1.1913	1.0560	1.2376	1.4193
4	**3.5749**	2.2905	**3.7321**	2.3912	**5.2465**	24	1.3613	1.4481	1.4121	1.5021	2.0969
5	0.5226	0.6949	0.5542	0.7370	0.4829	25	0.3328	0.4584	0.3430	0.4724	0.2101
6	0.6017	0.7861	0.6513	0.8509	0.6180	26	0.1562	0.2196	0.1604	0.2255	0.0482
7	0.0314	0.0444	0.0315	0.0445	0.0020	27	0.3638	0.4985	0.3765	0.5159	0.2485
8	1.0157	1.1907	1.0452	1.2252	1.4177	28	0.8963	1.0859	0.9268	1.1229	1.1792
9	0.0904	0.1276	0.0918	0.1296	0.0163	29	0.9096	1.0981	0.9928	1.1985	1.2059
10	0.0277	0.0392	0.0278	0.0393	0.0015	30	0.0947	0.1336	0.0968	0.1366	0.0179
11	0.0352	0.0498	0.0354	0.0500	0.0025	31	0.6284	0.8159	0.7212	0.9363	0.6657
12	0.5066	0.6760	0.5493	0.7329	0.4570	32	0.1327	0.1869	0.1389	0.1956	0.0349
13	0.7778	0.9727	0.8504	1.0636	0.9461	33	1.2047	1.3391	1.2368	1.3748	1.7932
14	0.2528	0.3520	0.2598	0.3618	0.1239	34	0.5419	0.7176	0.5588	0.7400	0.5150
15	0.4283	0.5804	0.4581	0.6208	0.3369	35	0.5385	0.7136	0.5539	0.7340	0.5092
16	0.1560	0.2193	0.1592	0.2238	0.0481	36	0.4765	0.6397	0.5057	0.6789	0.4093
17	0.0698	0.0986	0.0704	0.0994	0.0097	37	0.8963	1.0859	0.9268	1.1229	1.1792
18	2.9515	2.1323	**3.0945**	2.2356	**4.5466**	38	0.4836	0.6483	0.5093	0.6828	0.4203
19	1.0598	1.2271	1.1332	1.3121	1.5058	39	0.7067	0.9001	0.7262	0.9250	0.8102
20	0.2385	0.3326	0.2449	0.3416	0.1106						

Table 2: Outlier Diagnostics for the ESR Data

	Cut-off						Cut-off				
	3.00	3.00	3.00	3.00	1.58		3.00	3.00	3.00	3.00	1.58
Obs	PR	DR	SPR	SDR	DEVC	Obs	PR	DR	SPR	SDR	DEVC
1	0.3573	0.4902	0.3667	0.5031	0.2403	17	0.5720	0.7524	0.7066	0.9295	0.5661
2	0.2152	0.3009	0.2206	0.3085	0.0905	18	0.3049	0.4216	0.3120	0.4314	0.1777
3	0.1766	0.2478	0.1801	0.2528	0.0614	19	0.3857	0.5266	0.3956	0.5401	0.2773
4	0.1497	0.2105	0.1525	0.2145	0.0443	20	0.2506	0.3490	0.2568	0.3577	0.1218
5	0.7734	0.9683	0.8158	1.0214	0.9377	21	0.2702	0.3755	0.2800	0.3890	0.1410
6	0.2887	0.4002	0.2953	0.4093	0.1601	22	0.7543	0.9492	0.9135	1.1495	0.9009
7	0.6973	0.8902	0.7252	0.9259	0.7925	23	2.6448	2.0390	2.8866	2.2254	**4.1576**
8	0.2458	0.3425	0.2519	0.3511	0.1173	24	0.2431	0.3389	0.2487	0.3467	0.1149
9	0.7050	0.8983	0.7351	0.9367	0.8070	25	0.4457	0.6020	0.4591	0.6200	0.3624
10	0.4872	0.6526	0.5108	0.6842	0.4259	26	0.1663	0.2335	0.1697	0.2383	0.0545
11	0.2455	0.3421	0.2511	0.3499	0.1170	27	0.1455	0.2046	0.1481	0.2084	0.0419
12	0.1507	0.2119	0.1539	0.2164	0.0449	28	0.1671	0.2347	0.1715	0.2408	0.0551
13	0.2674	0.3716	0.3115	0.4330	0.1381	29	1.1615	1.3069	1.3942	1.5687	**1.7081**
14	2.0407	1.8121	2.1610	1.9189	**3.2835**	30	0.4193	0.5692	0.4494	0.6100	0.3240
15	**3.4584**	2.2636	3.5433	2.3191	**5.1238**	31	0.4790	0.6428	0.4905	0.6582	0.4131
16	0.5581	0.7364	0.5748	0.7584	0.5423	32	0.6073	0.7924	0.6336	0.8267	0.6279

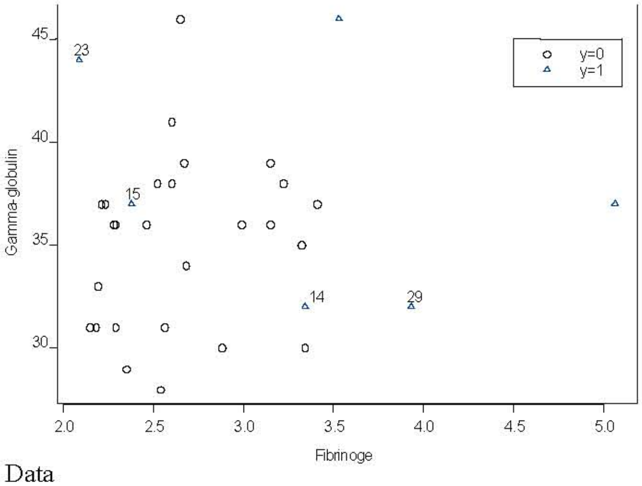

Fig. 3: Scatter plot of ESR Data

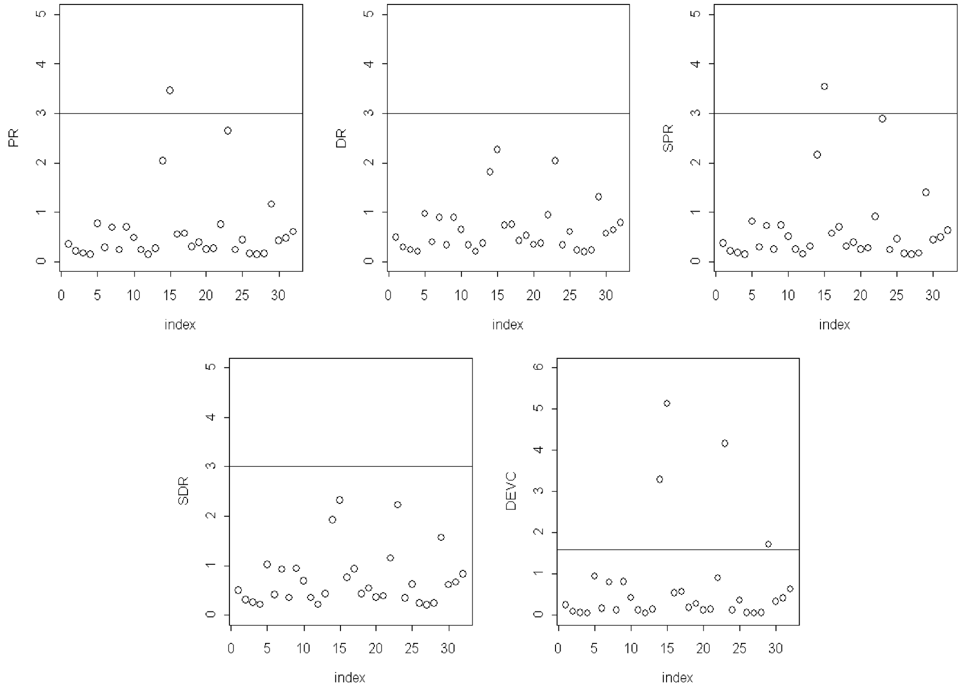

Fig. 4: Index plots of PR, DR, SPR, SDR and DEVC for ESR data

Table 3: Simulation Results of Outlier Diagnostics

Sample size	% of added outliers	No. Of outliers	PR	DR	SPR	SDR	DEVC
40	5	2	0	0	0	0	2
60		3	0	0	0	0	3
80		4	0	0	0	0	4
100		5	0	0	0	0	5
200		10	0	0	0	0	10
40	10	4	0	0	0	0	4
60		6	0	0	0	0	6
80		8	0	0	0	0	8
100		10	0	0	0	0	10
200		20	0	0	0	0	20

Monte Carlo Simulation Study: A Monte Carlo study was carried out to assess the performance of the proposed methods for different sample sizes of n = 40, 60, 80, 100 and 200 with 5% and 10% added outliers based on 1000 simulations. The outliers are added by considering y_i = 0 and the x_is are random numbers generated from Uniform (1.5, 2.0). A logistic regression model is generated with two independent normally distributed covariates and the response variable Y_i follow the model $\pi(x_i) = \exp\left(x_i^T \beta\right) \Big/ 1 + \exp\left(x_i^T \beta\right), i = 1,...,n$ with parameter p = 3, an intercept and two covariates and true parameter values β = (1,2,2). The simulation outputs are presented in Table 3.

Table 3 shows that the PR, DR, SPR and SDR methods are not able to identify any outliers when the data are contaminated with 5% and 10% residual outliers. In contrast, the newly proposed method has successfully identified all the outliers at both percentage of contamination. This simulation results suggest that the DEVC method is highly recommended to be used as a detection method for identifying multiple residual outliers in logistic regression.

It is important for statistics practitioners when analyzing data to be able to identify outliers if they exist so that appropriate measures may be taken. In this study, a new algorithm for the identification of residual outliers has been proposed based on the deviance component (DEVC). The performance of this new diagnostic method has been analyzed and compared with other existing diagnostic methods through numerical examples and simulation experiments. It is found that DEVC is able to identify the presence of multiple residual outliers successfully when other commonly used existing diagnostic methods fail to do so.

ACKNOWLEDGEMENTS

The authors would like to thank the Ministry of Higher Education (MOHE) Malaysia for providing the Fundamental Research Grant Scheme (FRGS) to carry out this study.

REFERENCES

1. Hosmer, D.W. and S. Lemeshow, 2000. Applied Logistic Regression. 2nd Edition., John Wiley and Sons, New York.
2. Collett, D., 2003. Modelling Binary Data. 2nd Edition, Chapman & Hall/CRC.
3. Cook, R.D. and S. Weisberg, 1982. Residuals and Influence in Regression. Chapman and Hall, New York.
4. Pregibon, D., 1981. Logistic Regression Diagnostics, Annals of Statistics, 9: 705-724.
5. Jennings, D.E., 1986. Outliers and Residual Distribution in Logistic Regression. Journal of American Statistical Association, 81: 987-990.
6. Copas, J.B., 1988. Binary Regression Model for Contaminated Data (with discussion). Journal of the Royal Statistical Society, Series B., 50: 225-265.
7. Imon, A.H.M.R. and A.S. Hadi, 2008. Identification of Multiple Outliers in Logistic Regression, Communications in Statistics Theory and Methods, 37: 1697-1709.
8. 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. Journal of Applied Statistics, 36: 507-520.
9. Ryan, T.P., 1997. Modern Regression Methods. Har/Dis Edn, Wiley, New York, USA.
10. Hadi, A.S., 1992. A New Measure of Overall Potential Influence in Linear Regression. Computational Statistics and Data Analysis, 14: 1-27.
11. Imon, A.H.M.R., 2005. Identifying Multiple Influential Observations in Linear Regression. Journal of Applied Statistics, 32: 929-946.
12. Imon, A.H.M.R., 2006. Identification of High Leverage Points in Logistic Regression. Pakistan Journal of Statistics, 22: 147-156.
13. Finney, D.J., 1947. The Estimation from Individual Records of the Relationship between Dose and Quantal Response. Biometrika, 34: 320-334.
14. Croux, C. and G. Haesbroeck, 2003. Inplementing the Bianco and Yohai Estimator for Logistic Regression. Computational Statistics and Data Analysis, 34: 320-334.
15. Collett, D. and A.A. Jemain, 1985. Residuals, Outliers and Influential Observations in Regression Analysis. Sains Malaysiana, 14: 493-511.