International Journal of Statistical Sciences *ISSN 1683–5603*
Vol. 8, 2008, pp 33-43

Identification of Pattern-disrupting Outliers in Logistic Regression

A.T.M. Shabbir Ahmed Prodhan

Department of Statistics
University of Rajshahi
Rajshahi-6205, Bangladesh

A.H.M. Rahmatullah Imon

Department of Mathematical Sciences
Ball State University
Muncie, IN 47306, USA

[Received October 10, 2006; Revised October 22, 2008; Accepted November 1, 2008]

Abstract

The identification of outliers has been an area of a great deal of attention in statistical data analysis for many years, especially in regression analysis. In recent years, the use of logistic regression modeling has exploded and there has been a great amount of effort in research on all statistical aspects of the logistic regression model including the identification of outliers. An important source of outliers in logistic regression is the existence of observations that disrupt the covariate pattern. In this paper we propose a new method for the identification of outliers in logistic regression where outliers are caused by the disruption in covariate pattern. The advantage of using the proposed method in the identification of outliers is then investigated through several examples.

Keywords and Phrases: Logistic Regression, Outliers, Pattern-disruption, Pearson Residuals, Masking, Group Deletion, Generalized Standardized Pearson Residuals.

AMS Classification: Primary 62J02; Secondary 62J20.

1 Introduction

Diagnostic methods are commonly used in all branches of regression analysis. In recent years, diagnostics has become an essential part of logistic regression. We often observe that the presence of outliers can mislead our interpretation. Thus, we need to detect such observations and study their impact on the model. In this paper our main objective is to identify a group of observations that appear as outliers due to their failure to match with the usual pattern in a logistic regression. In section 2, we introduce a logistic regression model and a class of residuals that are commonly used in the identification of outliers. In this paper we consider observations as outliers if the values of the response variable do not match with usual pattern of their corresponding regressor values. This situation is known as pattern-disruption in the experimental design data [see Montgomery (2005)]. We propose a new method in section 3 to identify pattern-disrupting outliers in logistic regression. The usefulness of this newly proposed method is demonstrated in section 4 through a variety of examples.

2 Identification of Outliers

Consider a simple two variable regression model

$$y_i = \beta_0 + \beta_1 x_i + \in_i, \quad i = 1, 2, \cdots, n \tag{1}$$

We would logically let $y_i = 0$ if the i^{th} unit does not have the characteristic and $y_i = 1$ if the i^{th} unit does possess that characteristic. In linear regression, the ordinary least squares (OLS) method is commonly used for estimating parameters mainly because of tradition and ease of computation. We can use the OLS method for estimating parameters in logistic regression, but the maximum likelihood (ML) method based on the iterative-reweighed least squares algorithm [see Ryan (1997) has become more popular with the statisticians. The specific form of the logistic regression model we use in this paper is:

$$y_i = \pi_i (X) + \in_i, \quad i = 1, 2, \cdots, n \tag{2}$$

where π_i, known as probability for the i^{th} factor/covariate and is defined as

$$\pi_i(X) = \frac{\exp\left(x_i^T \beta\right)}{1 + \exp\left(x_i^T \beta\right)}$$

where $x_i^T = [1 \;\; x_i]$ and $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$. The model given in (2) satisfies the important requirement that $0 \leq \pi_i \leq 1$ and will be a satisfactory model in many applications. The model in matrix notation becomes

$$Y = \pi(X) + \in \tag{3}$$

where Y is an $n \times 1$ vector of the observed responses, X is an $n \times 2$ matrix that contains the explanatory variable including the intercept, $\pi(X)$ is an $n \times 1$ vector containing the values π_i and $\in$ is an $n \times 1$ vector of unobserved random disturbances.

After estimating the model by the ML method, we define the i^{th} residual as

$$\hat{\in}_i = y_i - \hat{\pi}_i, \quad i = 1, 2, \cdots, n \tag{4}$$

According to Barnett and Lewis (1994) outliers are those which stand apart from the rest of the data. Several versions of outliers for regression problems are discussed in the literature [see Ryan (1997)]. In this paper we consider the case where the residuals measure the extent of ill-fitted factor/covariate patterns. Hence the observations that fail to match with the usual pattern of the majority of data are expected to possess large residuals and hence are considered as suspect outliers. Here we introduce different types of residuals that are commonly used in diagnostics for the identification of outliers.

2.1 Pearson Residuals

The Pearson residuals are elements of the Pearson chi-square that can be used to detect ill-fitted factor/covariate patterns. The i^{th} Pearson residual is given by

$$r_i = \frac{y_i - \hat{\pi}_i}{\sqrt{v_i}} \quad i = 1, 2, \cdots, n \tag{5}$$

where $v_i = \hat{\pi}_i(1\text{-}\hat{\pi}_i)$. We call an observation outlier if its corresponding Pearson residual exceeds 3 in absolute term.

2.2 Standardized Pearson Residuals

Pregibon (1981) pointed out that the linear regression-like approximation for the i^{th} residual is given as

$$\hat{\in}_i = y_i - \hat{\pi}_i \approx (1 - h_i) y_i \tag{6}$$

where h_i is the i^{th} diagonal element of the matrix $H = V^{1/2} X (X^T V X)^{-1} X^T V^{1/2}$. Here V is an $n \times n$ diagonal matrix with general element $v_i = \hat{\pi}_i(1 - \hat{\pi}_i)$. Hence the variance of the i^{th} residual becomes $v_i(1-h_i)$ which suggests that the Pearson residuals do not have variance equal to 1. For this reason we could use the standardized Pearson residuals given by

$$r_{si} = \frac{y_i - \hat{\pi}_i}{\sqrt{v_i\,(1 - h_i)}}, \quad i = 1, 2, \ldots, n \tag{7}$$

The i^{th} observation is termed as outlier if $|r_{si}| > 3$.

2.3 Generalized Standardized Pearson Residuals

The main problem in using the above residuals in the identification of outliers is that the presence of outliers may distort the fitting of a logistic model in such a way that the resulting residuals often suffer from masking (for which outliers may possess relatively smaller residuals and consequently remain unidentified) and/or swamping (for which inliers may possess relatively bigger residuals and are identified as outliers). For this reason a group deletion version of standardized Pearson residuals is suggested in the literature for the identification of multiple outliers in logistic regression. Let us denote a set of cases 'remaining' in the analysis by R and a set of cases 'deleted' by D. Hence R contains $(n-d)$ cases after d cases in D are deleted. Without loss of generality, assume that these observations are the last d rows of X, Y and V so that

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

Let $\hat{\beta}^{(-D)}$ be the corresponding vector of estimated coefficients when a group of observations indexed by D is omitted. Thus the corresponding fitted values for the logistic regression model are

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

Hence the i^{th} deletion residual is defined as

$$\hat{\in}_i^{(-D)} = Y_i - x_i^T \hat{\beta}^{(-D)}, \quad i = 1, 2, \ldots, n \tag{9}$$

When a group of observations D is omitted, we define deletion weights (DW) for the entire data set as

$$h_i^{(-D)} = \hat{\pi}_i^{(-D)}(1 - \hat{\pi}_i^{(-D)}) x_i^T \left(X_R^T V_R X_R\right)^{-1} x_i, \quad i = 1, 2, \ldots, n \tag{10}$$

We also define

$$v_i^{(-D)} = \hat{\pi}_i^{(-D)}(1 - \hat{\pi}_i^{(-D)}) \tag{11}$$

Using the above results and also using linear-regression like approximation, Imon and Hadi (2008) define the i^{th} generalized standardized Pearson residual (GSPR) as

$$r_{si}^{(-D)} = \frac{y_i - \hat{\pi}_i^{(-D)}}{\sqrt{v_i^{(-D)}\left(1 - h_i^{(-D)}\right)}} \qquad for \quad i \in R$$

$$= \frac{y_i - \hat{\pi}_i^{(-D)}}{\sqrt{v_i^{(-D)}\left(1 + h_i^{(-D)}\right)}} \qquad for \quad i \in D \tag{12}$$

Residuals defined in (12) are analogous to residuals suggested by Hadi and Simonoff (1993), Atkinson (1994), Munier (1999) and Imon (2005).

3 Identification of Pattern-disrupting Outliers

In this section we outline a method for the identification of pattern-disrupting outliers. Here we consider observations as outliers whose response values do not match with the usual pattern of the values of the explanatory variables. For finding the pattern of the explanatory variables, we first split the variable X into two variables, X_0 and X_1, where X_0 denotes a variable that contains the values of X corresponding to the values $Y = 0$ and X_1 denotes a variable that contains the values of X corresponding to the values $Y = 1$. Next we find the center of these two variables. Traditionally the sample means of these two variables may be considered as their corresponding central values. Then we can construct a confidence bound-type interval for the variable X as

$$I : \bar{X} \pm 2\ St.\ dev.(X) \tag{13}$$

Since the sample means could be highly non-robust in the presence of outliers in X, we prefer their corresponding sample medians as the central values. Then we make a confidence bound-type interval for the variable X as

$$I : Median(X) \pm 2MAD(X) \tag{14}$$

where a robust alternative measure of dispersion is the median absolute deviation (MAD), which is defined for a variable X as

$$MAD(X) = Median\{|X_i - Median(X)|\}/0.6745 \tag{15}$$

Let X_U and X_L be the upper and the lower bounds of the interval I respectively as given in (14). We suspect that the observations are disrupting the pattern of the model if they satisfy any of the following cases:

Case 1: Median (X_0) < Median (X_1) but $x_{i0} > X_U$.
Case 2: Median (X_0) > Median (X_1) but $x_{i1} < X_L$.

Now we form the deletion set D with the observations that satisfy either of the two cases. Then we compute the generalized standardized Pearson residuals for the entire data set after the omission of the cases indexed by D. Observations possessing the GSPR values bigger than 3 in absolute terms are finally declared as outliers.

4 Examples

In this section we consider several data sets that have been frequently used in the study of the identification of outliers in logistic regression.

4.1 Modified Brown Data

We first consider the modified Brown data. The original data set was given by Brown (1980) where the original objective was to see whether an elevated level of acid phosphatase (X) in blood serum would be of value as an additional regressor for predicting whether or not prostate cancer patients also had lymph node involvement. The dependent variable is nodal involvement (Y), with 1 denoting the presence of nodal involvement and 0 indicating the absence of such involvement. This data set has been extensively analyzed by many authors [see Ryan (1997)]. It is now believed that the original data set with 53 observations contains one outlier (observation no. 24). Imon and Hadi (2008) modified this data set by putting 2 more outliers (cases 54 and 55) in it. This modified data is presented in Table 1.

Table 1: Modified Brown data

Index	LNI	A.P.	Index	LNI	A.P.	Index	LNI	A.P.	Index	LNI	A.P.
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	54	0	200
13	0	65	27	0	40	41	0	66	55	0	220
14	1	67	28	0	50	42	1	84			

From the scatter plot of this data as given in Figure 1, we observe that X_1 values are on the average bigger than the values of X_0. But three observations of X_0 are much bigger than any other values of X_1 and consequently they may be considered as suspect pattern-disrupting outliers. For this data set we obtain Median $(X_0) = 59.00$, Median $(X_1) = 78.00$, Median $(X) = 66.00$ and MAD $(X) = 23.72$ that gives $X_L = 18.56$ and $X_U = 113.44$.

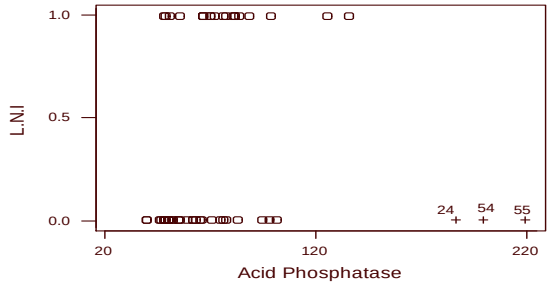

Figure 1: Scatter plot of Modified Brown Data

Table 2: Outlier diagnostics for modified Brown data

Index	PR	SPR	GSPR	Index	PR	SPR	GSPR	Index	PR	SPR	GSPR
1	-0.72	-0.73	-0.51	20	-0.79	-0.80	-1.68	39	-0.76	-0.76	-0.97
2	-0.73	-0.74	-0.61	21	-0.73	-0.74	-0.55	40	-0.78	-0.79	-1.56
3	-0.73	-0.73	-0.53	22	-0.76	-0.76	-0.95	41	-0.74	-0.75	-0.77
4	-0.73	-0.74	-0.55	23	1.27	1.29	0.64	42	1.30	1.32	0.89
5	-0.73	-0.73	-0.53	24	-0.91	-1.02	**-12.87**	43	1.31	1.32	0.95
6	-0.72	-0.73	-0.52	25	1.19	1.24	0.27	44	1.32	1.33	1.06
7	-0.72	-0.73	-0.48	26	1.31	1.32	0.93	45	1.33	1.35	1.22
8	-0.74	-0.75	-0.70	27	-0.71	-0.73	-0.42	46	1.32	1.33	1.01
9	1.37	1.38	1.70	28	-0.73	-0.73	-0.53	47	1.33	1.35	1.22
10	-0.73	-0.74	-0.59	29	-0.72	-0.73	-0.53	48	1.34	1.35	1.31
11	-0.74	-0.75	-0.70	30	-0.71	-0.73	-0.42	49	1.31	1.32	0.93
12	-0.75	-0.76	-0.86	31	-0.73	-0.74	-0.59	50	1.34	1.35	1.31
13	-0.74	-0.75	-0.75	32	-0.74	-0.74	-0.65	51	1.33	1.34	1.16
14	1.34	1.35	1.31	33	1.39	1.40	2.05	52	1.29	1.31	0.79
15	-0.72	-0.73	-0.49	34	1.38	1.40	1.91	53	1.21	1.25	0.34
16	-0.72	-0.73	-0.52	35	1.38	1.40	2.01	54	-0.93	-1.07	**-17.56**
17	-0.73	-0.73	-0.53	36	-0.72	-0.73	-0.51	55	-0.97	-1.16	**-28.23**
18	-0.76	-0.77	-1.02	37	-0.74	-0.75	-0.71				
19	-0.77	-0.77	-1.16	38	-0.79	-0.80	-1.85				

Now clearly Median (X_0) < Median (X_1) but we observe from Table 1 that there are 3 observations in this data set (case 24, 54 and 56) for which $x_{i0} > X_U$. Hence we form the deletion set D to compute the generalized standardized Pearson residuals. These values together with the Pearson residuals and the standardized Pearson residuals are given in Table 2.

We observe from the results given in Table 2 that both Pearson and standardized Pearson residuals fail to identify any one of the three outliers and these outliers get totally masked here. But the generalized standardized Pearson residuals can correctly identify all of them.

4.2 Artificial Data

Now we consider a set of artificial data for the detection of multiple outliers. We have generated data sets of size 40 each where the data for the explanatory variable X come from a uniform distribution in such a way that the X values corresponding to

$Y = 1$ are on the average bigger than those of $Y = 0$. We have named the data sets alphabetically as A, B, C, D and E and they are presented in Table 3.

Table 3: Artificial data sets

Index	Data Set A		Data Set B		Data Set C		Data Set D		Data Set E	
	Y	X	Y	X	Y	X	Y	X	Y	X
1	0	129.940	0	129.940	0	129.940	1	**70.000**	1	**80.000**
2	0	102.091	0	102.091	0	102.091	1	**75.000**	1	**85.000**
3	0	115.805	0	115.805	0	115.805	1	**80.000**	0	115.805
4	0	108.487	0	108.487	0	108.487	1	**85.000**	0	108.487
5	0	111.189	0	111.189	0	111.189	0	111.189	0	111.189
6	0	118.617	0	118.617	0	118.617	0	118.617	0	118.617
7	0	118.795	0	118.795	0	118.795	0	118.795	0	118.795
8	0	118.375	0	118.375	0	118.375	0	118.375	0	118.375
9	0	107.485	0	107.485	0	107.485	0	107.485	0	107.485
10	0	112.862	0	112.862	0	112.862	0	112.862	0	112.862
11	0	108.974	0	108.974	0	108.974	0	108.974	0	108.974
12	0	128.041	0	128.041	0	128.041	0	128.041	0	128.041
13	0	128.799	0	128.799	0	128.799	0	128.799	0	128.799
14	0	123.670	0	123.670	0	123.670	0	123.670	0	123.670
15	0	114.503	0	114.503	0	114.503	0	114.503	0	114.503
16	0	110.715	0	110.715	0	110.715	0	110.715	0	110.715
17	0	107.277	0	107.277	0	107.277	0	107.277	0	107.277
18	0	102.890	0	102.890	0	102.890	0	102.890	0	102.890
19	0	114.227	0	114.227	0	114.227	0	114.227	0	114.227
20	0	114.030	0	114.030	0	114.030	0	114.030	0	114.030
21	1	127.424	1	127.424	1	127.424	1	127.424	1	127.424
22	1	148.015	1	148.015	1	148.015	1	148.015	1	148.015
23	1	132.548	1	132.548	1	132.548	1	132.548	1	132.548
24	1	139.177	1	139.177	1	139.177	1	139.177	1	139.177
25	1	139.656	1	139.656	1	139.656	1	139.656	1	139.656
26	1	130.570	1	130.570	1	130.570	1	130.570	1	130.570
27	1	149.672	1	149.672	1	149.672	1	149.672	1	149.672
28	1	125.121	1	125.121	1	125.121	1	125.121	1	125.121
29	1	145.071	1	145.071	1	145.071	1	145.071	1	145.071
30	1	122.664	1	122.664	1	122.664	1	122.664	1	122.664
31	1	128.953	1	128.953	1	128.953	1	128.953	1	128.953
32	1	132.533	1	132.533	1	132.533	1	132.533	1	132.533
33	1	134.574	1	134.574	1	134.574	1	134.574	1	134.574
34	1	149.469	1	149.469	1	149.469	1	149.469	1	149.469
35	1	124.880	1	124.880	1	124.880	1	124.880	1	124.880
36	1	141.276	1	141.276	1	141.276	1	141.276	1	141.276
37	1	131.329	1	131.329	0	**165.000**	1	131.329	1	131.329
38	1	137.338	1	137.338	0	**170.000**	1	137.338	1	137.338
39	1	143.715	1	143.715	0	**175.000**	1	143.715	0	**165.000**
40	1	126.662	0	**165.000**	0	**180.000**	1	126.662	0	**170.000**

Data set A: Each value of Y perfectly matches with the pattern of X, i.e., each $x_{i0} < X_U$ and each$x_{i1} > X_L$. We observe from the scatter plot (see Figure 2) of this data that there exists no pattern-disrupting observation in this data.

Data set B: We have one observation (case 40) for which $x_{i0} > X_U$ and hence this data may have one pattern-disrupting outlier.

Data Set C: We observe from Table 3 and also from Figure 2 that this data set contains four observations (cases 37, 38, 39 and 40) for which $x_{i0} > X_U$.

Data Set D: So far we have considered examples where pattern-disrupting observations are occurring due to excessive large values of x_{i0}. In this data set we have four observations (cases 1, 2, 3 and 4) that has unusually low x_{i1} values such that $x_{i1} < X_L$.

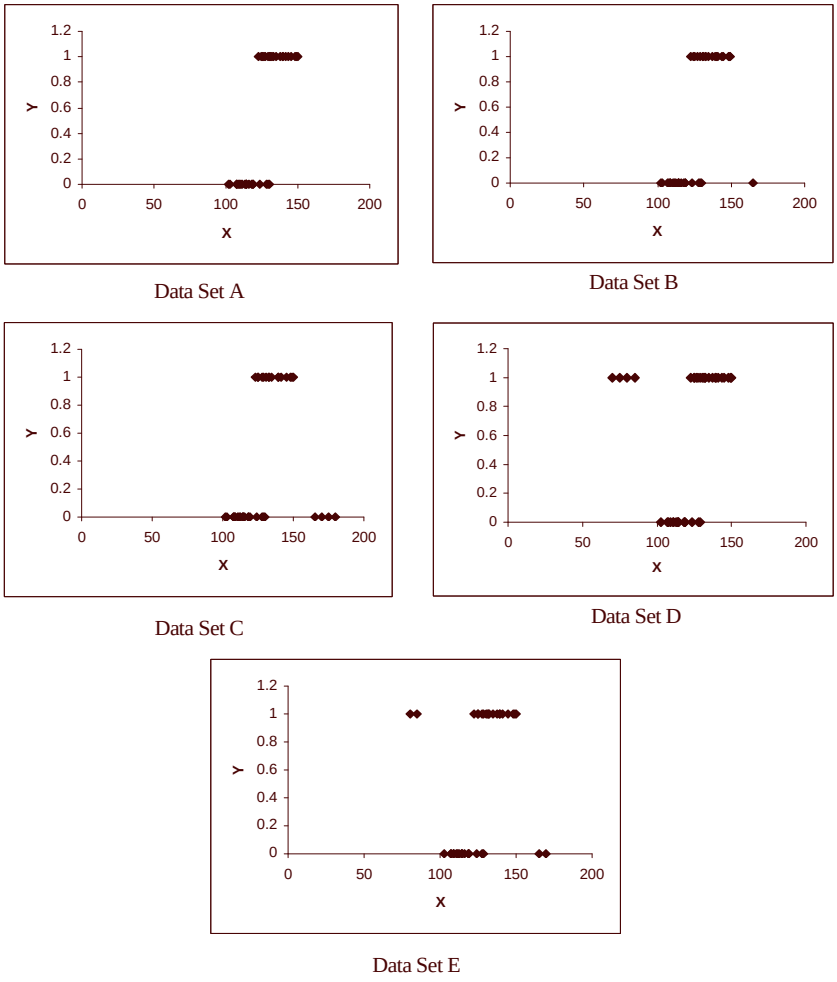

Figure 2: Scatter Plot for the Artificial Data Sets A - E

Data Set E: In the above examples we consider cases where the pattern-disrupting observations are occurring at the any one end of the data. The data set E gives us a situation where we have pattern-disrupting cases at the both ends of the data set. We observe from Figure 6 that the data set E contains two observations (cases 1 and 2) for which $x_{i1} < X_L$ and two observations (cases 39 and 40) for which $x_{i0} > X_U$.

For each artificial data set we have computed three sets of residuals, the Pearson residuals, the standardized Pearson residuals, and the generalized standardized Pearson residuals. Summary results are presented in Table 4.

Table 4: Diagnostics for the suspect outliers of the artificial data sets A-E

Data Set	Index of Suspect Case	PR	SPR	GSPR
A	—	—	—	—
B	40	**-15.86**	**-15.964**	**-584.53**
C	37	-1.426	-1.517	**-412.93**
	38	-1.542	-1.660	**-888.12**
	39	-1.669	-1.815	**-1911.18**
	40	-1.805	-1.984	**-4116.38**
D	1	1.790	1.977	**29088.64**
	2	1.657	1.811	**11296.97**
	3	1.534	1.657	**4391.16**
	4	1.420	1.516	**1707.79**
E	1	2.161	2.323	**4563.54**
	2	1.984	2.117	**1777.49**
	39	-1.978	-2.111	**-1954.29**
	40	-2.154	-2.317	**-5017.08**

We observe from the results presented in Table 4 that the data set A does not contain any outlier and neither of the three sets of residuals, PR, SPR and GSPR contains any residual (in absolute term) bigger than 3. The data set B contains a single outlier and all three techniques can correctly identify it. But when we have more than one outlier such as the examples C, D and E we see that both Pearson and standardized Pearson residuals fail to identify even a single outlier. But the generalized standardized Pearson residuals can correctly identify each and every outlier and hence are considered as the most effective diagnostics.

5 Conclusions

In this paper we suggest a method to identify outliers in logistic regression that are responsible for the disruption of the usual design. We use the generalized standardized Pearson residuals as diagnostics in this method. The numerical examples show that our proposed method may be very effective to identify multiple outliers under a variety of situations where the traditional methods fail to do so.

References

[1] Atkinson, A.C. (1994). Fast very robust methods for the detection of multiple outliers, *Journal of the American Statistical Association*, **89**, 1329 - 1239.

[2] Barnett, V. and Lewis, T.B. (1994). *Outliers in Statistical Data*, 3rd edition, Wiley, New York.

[3] Brown, B.W., Jr. (1980). 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.

[4] Hadi, A.S. and Simonoff, J.S. (1993). Procedure for the identification of outliers in linear models, *Journal of the American Statistical Association*, **88**, 1264 - 1272.

[5] Hosmer, D.W. and Lemeshow, S. (2000). *Applied Logistic Regression*, 2nd Ed., Wiley, New York.

[6] Imon, A. H. M. R. (2005). Identifying multiple influential observations in linear regression, *Journal of Applied Statistics*, **32**, 929 - 946.

[7] Imon, A. H. M. R. and Hadi, A.S. (2008). Identification of multiple outliers in logistic regression, *Communications in Statistics - Theory and Methods*, **37**, 1697 - 1709.

[8] Montgomery, D.C. (2005). *Design and Analysis of Experiments*, 6th ed., Wiley, New York.

[9] Munier, S. (1999). Multiple outlier detection in logistic regression, *Student*, **3**, 117 - 126.

[10] Pregibon, D. (1981). Logistic regression diagnostics, *Annals of Statistics*, **9**, 977 - 986.

[11] Ryan, T.P. (1997). *Modern Regression Methods*, Wiley, New York.