Journal of Applied Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/cjas20

Identification of multiple influential observations in logistic regression

A. A.M. Nurunnabi [a], A. H.M. Rahmatullah Imon [b] & M. Nasser [c]

[a] Department of Business Administration, Uttara University, Dhaka, Bangladesh

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

[c] Department of Statistics, University of Rajshahi, Rajshahi, Bangladesh

Published online: 21 Sep 2010.

To cite this article: A. A.M. Nurunnabi, A. H.M. Rahmatullah Imon & M. Nasser (2010) Identification of multiple influential observations in logistic regression, Journal of Applied Statistics, 37:10, 1605-1624, DOI: 10.1080/02664760903104307

To link to this article: http://dx.doi.org/10.1080/02664760903104307

Journal of Applied Statistics
Vol. 37, No. 10, October 2010, 1605–1624

Taylor & Francis
Taylor & Francis Group

Identification of multiple influential observations in logistic regression

A.A.M. Nurunnabi[a]*, A.H.M. Rahmatullah Imon[b] and M. Nasser[c]

[a]*Department of Business Administration, Uttara University, Dhaka, Bangladesh;* [b]*Department of Mathematical Sciences, Ball State University, Muncie, IN 47306, USA;* [c]*Department of Statistics, University of Rajshahi, Rajshahi, Bangladesh*

(*Received 12 May 2008; final version received 8 June 2009*)

The identification of influential observations in logistic regression has drawn a great deal of attention in recent years. Most of the available techniques like Cook's distance and difference of fits (DFFITS) are based on single-case deletion. But there is evidence that these techniques suffer from masking and swamping problems and consequently fail to detect multiple influential observations. In this paper, we have developed a new measure for the identification of multiple influential observations in logistic regression based on a generalized version of DFFITS. The advantage of the proposed method is then investigated through several well-referred data sets and a simulation study.

Keywords: generalized DFFITS; generalized Studentized Pearson residual; generalized weight; high leverage point; influential observation; masking; outlier; swamping

1. Introduction

The use of logistic regression has been exploded in recent years. From its early use in epidemiology, it is now being used in every branch of knowledge. With the increasing use of this method different aspects of inference drawn from logistic regression are being examined. One of the most important issues is the estimation of parameters of a logistic model in the presence of unusual observations. To quote [21],

> The most usual method for estimating the parameters in logistic regression is maximum likelihood method, though the method has good optimality properties in ideal settings, but is extremely sensitive to 'bad' (bad from the point of view of outlying response (Y) and from the point of view of extreme points in the design space (X)) data.

The most popular and commonly used remedy to this problem is the employment of diagnostics. In regression we have an implicit assumption that all observations are equally reliable and should have an equal role in determining the regression equation and the subsequent conclusions [6]. But in the presence of influential observations, this assumption breaks down and the necessity

*Corresponding author. Email: pyal1471@yahoo.com

ISSN 0266-4763 print/ISSN 1360-0532 online

DOI: 10.1080/02664760903104307
http://www.informaworld.com

of regression diagnostics comes to our mind. Diagnostics are certain quantities computed from the data with the purpose of pinpointing influential points after which these influential points can be removed or corrected. Hence we need to identify such observations and study their impact on the model and the subsequent analyses. We first briefly discuss the idea of influential observations along with outliers and high leverage points. We also discuss some commonly used measures of influential observations designed for logistic regression. Then we propose a method based on the generalized difference of fits (GDFFITS) [16] for the identification of multiple influential observations in logistic regression. Finally, we present few examples and a simulation study to show the effectiveness of the newly proposed method in the identification of multiple influential observations in logistic regression.

2. Influential observations in logistic regression

According to Ref. [2], an observation is influential if it is individually or together with several other observations, has a demonstrably larger impact on the calculated values of various estimates than is the case for most of the other observations. In regression diagnostics we often deal with the terms, 'outlier' and 'high leverage point' that have very close ties with an influential observation. However, there is evidence [6] that they are not necessarily inter-related. In logistic regression outlier and/or influential point may occur as altercation (misclassification) between the binary (1, 0) responses. It may occur by meaningful deviation (we also see low leverage) in explanatory variables, which also deviates the response as such, so that the usual pattern (S-curve) of the majority of the data is disrupted.

Influential observations are hard to detect by visual perception, especially when the dimension (number of explanatory variables) is high. The problem of multiple influential observations in any analysis can be quite perplexing because of masking (false-negative) and swamping (false-positive) phenomena. To quote [14],

> Masking occurs when an outlying subset goes undetected because of the presence of another, usually adjacent, subset. Swamping occurs when good observations are incorrectly identified as outliers because of the presence of another, usually remote, subset of observations.

Let us consider the standard logistic regression model:

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

where

$$\pi(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}, \quad 0 \le \pi(X) \le 1. \tag{2}$$

This form gives an S-curve configuration. The well-known 'logit' transformation in terms of $\pi(X)$ is:

$$g(X) = \ln\left[\frac{\pi(X)}{1-\pi(X)}\right] = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p. \tag{3}$$

In matrix notation, $g(X) = X\beta$, where X is an $n \times k$ matrix containing the data for each case with $k = p + 1$, Y is an $n \times 1$ vector of binary responses and $\beta^{\mathrm{T}} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)$ is the vector of regression parameters. For the model $Y = \pi(X) + \varepsilon$, the error term

$$\varepsilon = \begin{cases} 1-\pi(X) & \text{w.p. } \pi(X); \quad \text{if } y = 1 \\ -\pi(X) & \text{w.p. } 1-\pi(X); \quad \text{if } y = 0 \end{cases}$$

has a distribution with mean zero and variance $\pi(X)[1-\pi(X)]$. Here ε violates most of the least squares assumptions and hence the maximum likelihood (ML) method based on iterative

reweighed least squares (IRLS) method is used to estimate the parameters and fit the model. But the ML method is very sensitive to outlying responses and extreme points in design space X. We want to detect the influential observations to remove (if necessary) them from the data to draw more reliable inferences. One of the most popular methods for the detection of influential data points is row deletion technique that examines in turn how the deletion of single row affects the estimates. The observation with the largest residual is not necessarily the most influential [6], however, deletion of observation possessing a small residual, may have a marked effect on the parameter estimates.

A large body of literature is now available [1,7,15,16,23] for the identification of influential observations in linear regression. Most of the theoretical works to extend results of linear regression to logistic regression is available in Ref. [21]. In the logistic regression, the basic building blocks for the identification of outlying and influential points are a residual vector and a projection matrix [21]. Extending this idea and also using the linear regression-like approximations as suggested in Ref. [21], influence diagnostics originally proposed for linear regression and are frequently used in logistic regression. Among a number of available influence measures the Cook's distance [8] and the difference of fits (DFFITS) [2] have become very popular with the practitioners.

For a logistic regression model, the ith Cook's distance is defined [23] as:

$$\mathrm{CD}_i = \frac{\left(\widehat{\beta}^{(-i)} - \widehat{\beta}\right)^{\mathrm{T}} (X^{\mathrm{T}} V X) \left(\widehat{\beta}^{(-i)} - \widehat{\beta}\right)}{k\widehat{\sigma}^2}, \quad i = 1, 2, \ldots, n, \tag{4}$$

where $\widehat{\beta}^{(-i)}$ is the estimated parameter of β with the ith observation deleted. Using the linear regression-like approximations as suggested in Ref. [21], Equation (4) can be expressed as:

$$\mathrm{CD}_i \approx \frac{1}{k} r_{si}^2 \left(\frac{h_{ii}}{1 - h_{ii}}\right), \tag{5}$$

where r_{si} is the ith standardized Pearson residual defined as:

$$r_{si} = \frac{y_i - \widehat{\pi}_i}{\sqrt{v_i(1 - h_{ii})}}, \quad i = 1, 2, \ldots, n, \tag{6}$$

where h_{ii} is the ith leverage value, which is in fact the ith diagonal element of the leverage matrix

$$H = V^{1/2} X (X^{\mathrm{T}} V X)^{-1} X^{\mathrm{T}} V^{1/2} \tag{7}$$

and V is a diagonal matrix with diagonal element v_i defined as:

$$V(y_i|x_i) = v_i = \widehat{\pi}_i (1 - \widehat{\pi}_i). \tag{8}$$

For the cut-off, the Cook's distance index plots are usually applied. But the problem with a graphical display-like index plot is that it is very much subjective and cannot be readily applied for further statistical computations. Some authors emphasize the value of using 'natural gap' in diagnostic statistics while some others propose using 'bench mark' to flag influential and diagnostic measures. In our study, we call an observation influential if its corresponding Cook's distance (CD) value is greater than 1 (see Ref. [9]).

The DFFITS was introduced in Ref. [2], which is defined as:

$$\text{DFFITS}_i = \frac{\widehat{y}_i - \widehat{y}_i^{(-i)}}{\widehat{\sigma}^{(-i)}\sqrt{h_{ii}}}, \quad i = 1, 2, \ldots, n, \tag{9}$$

where $\widehat{y}^{(-i)}$ and $\widehat{\sigma}^{(-i)}$ are respectively the ith fitted response and the estimated standard error with the ith observation deleted. DFFITS values can be expressed in terms of standardized Pearson residuals and leverage values as:

$$\text{DFFITS}_i = r_{si}\sqrt{\frac{h_{ii}}{(1-h_{ii})}\frac{v_i}{v_i^{(-i)}}}, \quad i = 1, 2, \ldots, n. \tag{10}$$

Observations with DFFITS value larger than $c\sqrt{k/n}$, where c is a suitably chosen constant between 2 and 3 or more [2,13,16], is termed as an influential observation. Using Equations (5) and (10), it is easy to establish a simple relationship between CD_i and DFFITS_i as:

$$\text{CD}_i = \frac{v_i^{(-i)^2}}{k v_i^2}\text{DFFITS}_i^2. \tag{11}$$

We observe from Equation (11) that the only essential difference between CD_i and DFFITS_i is the use of the estimated variance. In linear regression, DFFITS is considered as a better choice [16] because, it is more informative about σ^2 than CD_i and it will calculate simultaneous effect on both the parameter estimates and the estimate of variance. Many other influence measures are available in the literature [7,13,15].

3. Identification of multiple influential observations

The diagnostic tools discussed so far are designed for the identification of a single influential observation and are ineffective when masking and/or swamping occur. Therefore, we need detection techniques that are free from these problems. Here we introduce a group-deleted version of the residuals and weights that will be later used to develop effective diagnostics for the identification of multiple influential observations in logistic regression. We assume that d observations among a set of n observations are deleted. 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 are deleted. Without loss of generality, assume that these observations are the last d rows of X, Y and V (variance–covariance matrix) so that

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

Let $\widehat{\beta}^{(-D)}$ be the corresponding vector of estimated coefficients when a group of observations indexed by D is omitted. The fitted values for the entire logistic regression model based on R set are defined as:

$$\widehat{\pi}_i^{(-D)} = \frac{\exp\left(x_i^{\mathrm{T}}\widehat{\beta}^{(-D)}\right)}{1+\exp\left(x_i^{\mathrm{T}}\widehat{\beta}^{(-D)}\right)}, \quad i = 1, 2, \ldots, n. \tag{12}$$

Here we define the ith deletion residual as:

$$\widehat{\varepsilon}_i^{(-D)} = y_i - \widehat{\pi}_i^{(-D)}, \quad i = 1, 2, \ldots, n \tag{13}$$

with the corresponding variance

$$v_i^{(-D)} = \widehat{\pi}_i^{(-D)} \left(1 - \widehat{\pi}_i^{(-D)}\right). \tag{14}$$

Hence the ith diagonal element of the leverage matrix can be re-expressed as:

$$h_{ii}^{(-D)} = \widehat{\pi}_i^{(-D)} \left(1 - \widehat{\pi}_i^{(-D)}\right) x_i^{\mathrm{T}} (X_R^{\mathrm{T}} V_R X_R)^{-1} x_i \quad i = 1, 2, \ldots, n. \tag{15}$$

One can use the generalized Cook-type measure [10] defined as:

$$\mathrm{C}D^{(-D)} = \frac{\left(\widehat{\beta}^{(-D)} - \widehat{\beta}\right)^{\mathrm{T}} (X^T V X) \left(\widehat{\beta}^{(-D)} - \widehat{\beta}\right)}{k\widehat{\sigma}^2}. \tag{16}$$

But the problem is, the aforegiven statistic is defined only for the deletion group and consequently cannot be applied for the identification of the multiple influential cases for the whole data set. To overcome this problem in linear regression, the GDFFITS is used [16] for the entire data set defined as:

$$\mathrm{GDFFITS}_i = \begin{cases} \dfrac{\widehat{y}_i^{(-D)} - \widehat{y}_i^{(-D-i)}}{\widehat{\sigma}^{(-D-i)} \sqrt{h_{ii}^{(-D)}}} & \text{for } i \in R \\ \dfrac{\widehat{y}_i^{(-D+i)} - \widehat{y}_i^{(-D)}}{\widehat{\sigma}^{(-D)} \sqrt{h_{ii}^{(-D+i)}}} & \text{for } i \in D, \end{cases} \tag{17}$$

where

$$h_{ii}^{(-D)} = x_i^{\mathrm{T}} (X_R^{\mathrm{T}} X_R)^{-1} x_i \tag{18}$$

and

$$h_{ii}^{(-D+i)} = x_i^{\mathrm{T}} (X_R^T X_R + x_i x_i^{\mathrm{T}})^{-1} x_i = \frac{h_{ii}^{(-D)}}{1 + h_{ii}^{(-D)}}. \tag{19}$$

Now we would develop the GDFFITS suitable for logistic regression. Using results (17–19) and also using the linear regression-like approximation, we define GDFFITS for the logistic regression model as:

$$\mathrm{GDFFITS}_i = \begin{cases} \dfrac{\widehat{y}_i^{(-D)} - \widehat{y}_i^{(-D-i)}}{\sqrt{v_i^{(-D)}\, h_{ii}^{(-D)}}} & \text{for } i \in R \\ \dfrac{\widehat{y}_i^{(-D+i)} - \widehat{y}_i^{(-D)}}{\sqrt{v_i^{(-D)}\, h_{ii}^{(-D+i)}}} & \text{for } i \in D, \end{cases} \tag{20}$$

where $v_i^{(-D)}$ and $h_{ii}^{(-D)}$ are defined in Equations (14) and (15), respectively.

For the identification of multiple outliers in logistic regression, we can use the generalized standardized Pearson residual (GSPR) [18] defined as:

$$r_{si}^{(-D)} = \begin{cases} \dfrac{y_i - \hat{\pi}_i^{(-D)}}{\sqrt{v_i^{(-D)}(1 - h_{ii}^{(-D)})}} & \text{for } i \in R \\ \dfrac{y_i - \hat{\pi}_i^{(-D)}}{\sqrt{v_i^{(-D)}(1 + h_{ii}^{(-D)})}} & \text{for } i \in D. \end{cases} \tag{21}$$

Observations corresponding to the cases $|\text{GSPR}| > 3$ are considered as outliers. Based on a group of deleted cases indexed by D, let us define the generalized weights (GWs) denoted by $h_{ii}^{*(-D)}$ as:

$$h_{ii}^{*(-D)} = \begin{cases} \dfrac{h_{ii}^{(-D)}}{1 - h_{ii}^{(-D)}} & \text{for } i \in R \\ \dfrac{h_{ii}^{(-D)}}{1 + h_{ii}^{(-D)}} & \text{for } i \in D. \end{cases} \tag{22}$$

The cases having GWs, $h_{ii}^{*(-D)}$ larger than 'Median $(h_{ii}^{*(-D)}) + 3\text{MAD}(h_{ii}^{*(-D)})$' are high leverage points. Using Equations (21) and (22), our proposed GDFFITS quantities can be expressed in terms of GSPR and GW as:

$$\text{GDFFITS}_i = r_{si}^{(-D)} \sqrt{h_{ii}^{*(-D)}}, \quad i = 1, 2, \ldots, n. \tag{23}$$

We suggest considering the observations as influential if

$$|\text{GDFFITS}_i| \geq c \sqrt{\frac{k}{(n-d)}}. \tag{24}$$

We choose c as in linear regression, a suitable constant between 2 and 3 [2,13,16]. Although the expressions for this diagnostics are available for any arbitrary set of deleted cases, D, the choice of such a set is very important as the omission of this group determines the GDFFITS diagnostics for the whole set.

Here we outline a stepwise method for the detection of multiple influential observations in a logistic regression model.

Step 1: In the first step, we try to find out all suspect influential cases. Sometimes graphical displays like index plot and character plot of explanatory and response variables could give us an idea about the influential observations, but these plots are not always helpful for higher dimension of regressors. Another disadvantage of the graphical method is that it heavily depends on the experimenter's own interpretation. We suspect that influential observations are potential outliers or high leverage points or both. Hence we can compute the GSPR [18] and/or some leverage measures suggested in Refs. [15] and [17] to identify the suspect influential cases. In general, we prefer using any suitable robust technique such as the least median of squares (LMS) and the least trimmed squares (LTS) [22], the MM [24], the block-adaptive computationally efficient outlier nominator (BACON) [3] or the best omitted from the ordinary least squares (BOFOLS) [11] local influence measures such as those proposed in Refs. [9] and [19] and the standardized version proposed in Ref. [20] to find all suspect influential cases to form the deletion set D.

After the formation of the deletion set indexed by D we would compute the influence diagnostics GDFFITS as given in Equation (20) for the entire data. For the identification of influential observations we can use the rule as in Equation (24) and finally select 3 for the c, that is,

$$|\text{GDFFITS}_i| \geq 3\sqrt{\frac{k}{(n-d)}}, \tag{25}$$

and observations satisfying this rule will be declared as influential observations.

Step 2: It is now evident [11,16] that the robust methods are too prone to identify influential cases. As the suspect influential cases are mostly based on robust methods, there is a possibility that not all GDFFITS values considered in step 1 should satisfy rule (25). When we delete a group of observations it is possible that some observations may be wrongly detected as suspect influential cases because of their association with other unusual observations [18]. It should not matter much as we observe from Equation (20) that GDFFITS values for both the R and D sets are measured in a similar scale. But we feel that the deletion of harmless observations do not help in producing a better set of residuals and leverages. If not all observations in the D set individually satisfy rule (25), we would like to put back the cases into the estimation subset and observation with the least GDFFITS value will come first. We would recompute the GDFFITS values in every occasion and would continue this process until all GDFFITS values corresponding to the members of the D set individually satisfy rule (25). Observations thus identified will be finally declared as influential observations.

4. Examples

In this section, we consider few examples to investigate the usefulness of our newly proposed method for the identification of multiple influential observations in logistic regression.

Table 1. Imon and Hadi's modified Brown data.

Index	LNI	AP	Index	LNI	AP	Index	LNI	AP
1	0	48	20	0	98	39	0	76
2	0	56	21	0	52	40	0	95
3	0	50	22	0	75	41	0	66
4	0	52	23	1	99	42	1	84
5	0	50	**24**	**0**	**187**	43	1	81
6	0	49	25	1	136	44	1	76
7	0	46	26	1	82	45	1	70
8	0	62	27	0	40	46	1	78
9	1	56	28	0	50	47	1	70
10	0	55	29	0	50	48	1	67
11	0	62	30	0	40	49	1	82
12	0	71	31	0	55	50	1	67
13	0	65	32	0	59	51	1	72
14	1	67	33	1	48	52	1	89
15	0	47	34	1	51	53	1	126
16	0	49	35	1	49	**54**	**0**	**200**
17	0	50	36	0	48	**55**	**0**	**220**
18	0	78	37	0	63			
19	0	83	38	0	102			

Note: LNI, lymph node involvement; AP, alkaline phosphatase.

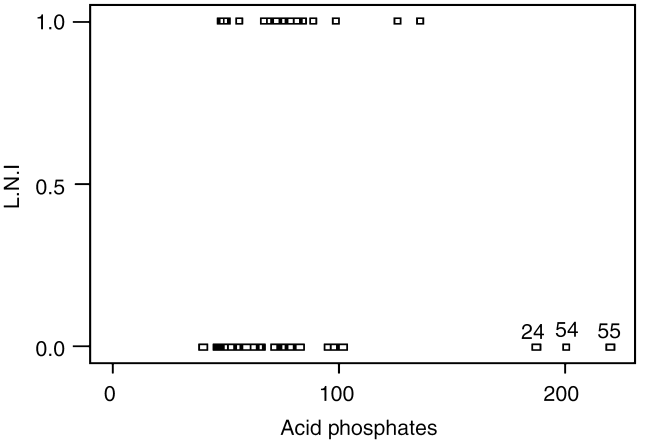

Figure 1. Scatter plot of lymph node involvement against acid phosphates for Imon and Hadi's modified Brown data.

4.1 *Imon and Hadi's modified Brown data*

We first consider the data set given in Ref. [4]. Here the main objective was to see whether an elevated level of acid phosphates (AP) in the blood serum would be of value for predicting whether

Table 2. Influence diagnostics for Imon and Hadi's modified Brown data.

Index	CD (1.00)	|DFFITS| (0.572)	|GDFFITS| (0.588)	Index	CD (1.00)	|DFFITS| (0.572)	|GDFFITS| (0.588)
1	0.01	−0.119	−0.097	29	0.01	−0.117	0.098
2	0.01	−0.110	−0.102	30	0.01	−0.131	−0.089
3	0.01	−0.117	−0.098	31	0.01	−0.111	−0.101
4	0.01	−0.114	−0.100	32	0.01	−0.107	−0.105
5	0.01	−0.117	−0.098	33	0.03	0.232	0.394
6	0.01	−0.118	−0.098	34	0.02	0.222	0.350
7	0.01	−0.122	−0.095	35	0.03	0.229	0.379
8	0.01	−0.105	−0.108	36	0.01	−0.119	−0.097
9	0.02	0.207	0.285	37	0.01	−0.104	−0.110
10	0.01	−0.111	−0.101	38	0.01	−0.136	−0.581
11	0.01	−0.105	−0.108	39	0.01	−0.102	−0.171
12	0.01	−0.101	−0.137	40	0.01	−0.122	−0.439
13	0.01	−0.103	−0.114	41	0.01	−0.103	−0.117
14	0.02	0.186	0.199	42	0.02	0.184	0.193
15	0.01	−0.121	−0.096	43	0.02	0.181	0.190
16	0.01	−0.118	−0.098	44	0.02	0.180	0.186
17	0.01	−0.117	−0.098	45	0.02	0.182	0.190
18	0.01	−0.103	−0.189	46	0.02	0.180	0.188
19	0.01	−0.106	−0.245	47	0.02	0.182	0.190
20	0.01	−0.128	−0.498	48	0.02	0.186	0.199
21	0.01	−0.114	−0.100	49	0.02	0.182	0.191
22	0.01	−0.102	−0.163	50	0.02	0.186	0.199
23	0.02	0.211	0.191	51	0.02	0.181	0.187
24	0.12	−0.498	**−2.121**	52	0.02	0.190	0.196
25	0.06	0.341	0.086	53	0.04	0.300	0.116
26	0.02	0.182	0.191	**54**	0.18	**−0.594**	**−2.376**
27	0.01	−0.131	−0.089	**55**	0.30	**−0.782**	**−2.759**
28	0.01	−0.117	−0.098				

Note: GDFFITS, generalized difference of fits; DFFITS, difference of fits.

or not prostate cancer patients also had lymph node involvement (LNI). It has been reported (see Ref. [23]) that the original data on the 53 patients may contain one influential observation (case 24). This data set was modified by Ref. [18] who put two additional outliers (cases 54 and 55) with potentially high leverages so that they may be considered as influential observations. This modified data is presented in Table 1. Here the dependent variable is nodal involvement, with 1 denoting the presence of nodal involvement and 0 indicating the absence of such involvement.

The scatter plot of LNI against APs (see Figure 1) clearly shows that observations 24, 54 and 55 may severely distort the covariate pattern. They may be considered as outliers and high leverage points at the same time and thus should produce very high influences. Table 2 presents influence diagnostics for the Imon and Hadi's modified Brown data. We observe from this table that the most commonly used diagnostics fail to identify the influential cases. The Cook's distance, using 1 as the cut-off value, fails to identify even a single influential observation although the values for the cases 24, 54 and 55 are substantially larger than the other cases. DFFITS can identify cases 54 and 55 correctly but masks case 24 (see Figure 2b).

Now we apply our newly proposed method to identify influential observations for this data set. When we employ the BACON algorithm we identify cases 24, 54 and 55 as suspect influential observations and form the deletion set $D$with these three observations. We compute GDFFITS for the whole set based on R (without the suspect cases) and the results are presented in Table 2. We observe from this table that the GDFFITS values for the suspected case are much larger than the rest and all of them exceed the cut-off value, 0.588. Similar conclusions may be drawn from the index plot of GDFFITS as shown in Figure 2c. All these three suspected cases are clearly separated from the other data and are correctly identified as influential observations.

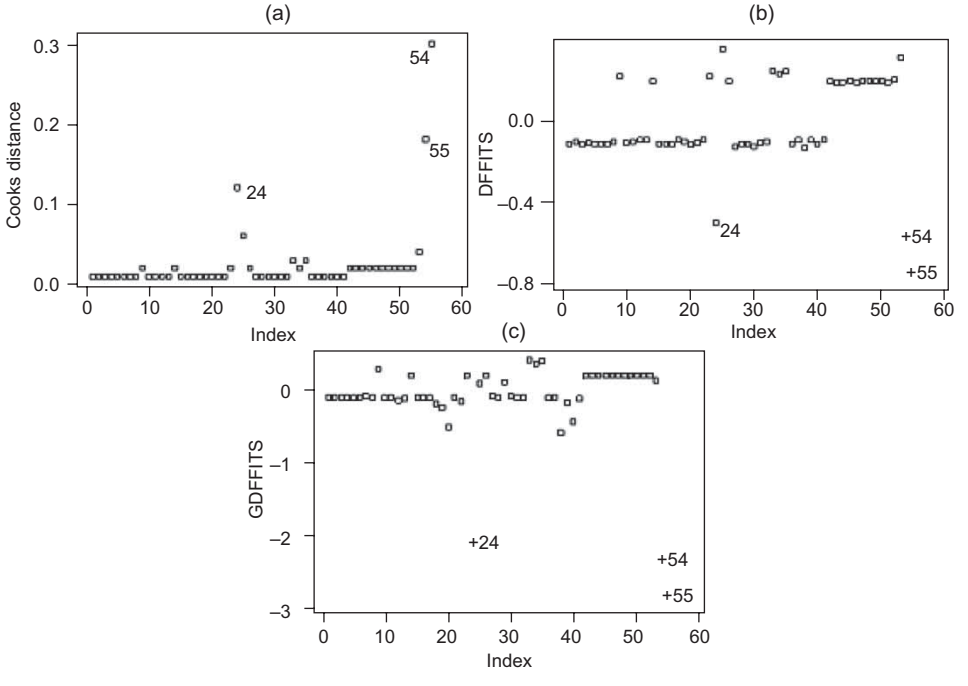

Figure 2. Index plots of: (a) Cook's distance, (b) difference of fits (DFFITS) and (c) generalized DFFITS (GDFFITS) for Imon and Hadi's modified Brown data.

4.2 *Modified Kyphosis data*

Now we consider another data given in Ref. [5]. This data set contains information about 81 children who have had corrective spinal surgery. The response variable tells us whether a post-operative deformity (Kyphosis) is 'present' or 'absent' in the data. It is thought that Kyphosis depends on two variables, the number of vertebrae involved in the operation (number) and the beginning of the range of vertebrae involved in the operation (start). From the original data it seems to us that one observation (case 43) may have a large influence in fitting the model. We now deliberately change six observations (cases 10, 11, 23, 40, 46 and 77) to make them influential and this modified data is presented in Table 3.

Table 4 presents influence diagnostics for the modified Kyphosis data. We observe from this table that Cook's distance (using 1 as cut-off value) fails to identify even a single observation to be influential and DFFITS can identify only one (case 43) as an influential observation. We

Table 3. Modified Kyphosis data.

Index	Kyphosis	Number	Start	Index	Kyphosis	Number	Start
1	0	3	5	42	0	3	13
2	0	3	14	**43**	**0**	**9**	**3**
3	1	4	5	44	0	4	1
4	0	5	1	45	0	3	16
5	0	4	15	**46**	**1**	**3**	**10**
6	0	2	16	47	0	4	15
7	0	2	17	48	0	5	13
8	0	3	16	49	1	3	3
9	0	2	16	50	0	2	14
10	**1**	**6**	**12**	51	0	5	10
11	**1**	**5**	**14**	52	0	2	17
12	0	3	16	53	1	10	6
13	0	5	2	54	0	2	17
14	0	4	12	55	0	4	15
15	0	3	18	56	0	5	15
16	0	3	16	57	0	3	13
17	0	6	15	58	1	5	8
18	0	5	13	59	0	7	9
19	0	5	16	60	0	3	13
20	0	4	9	61	1	4	1
21	0	2	16	62	1	7	8
22	1	6	5	63	0	4	1
23	**1**	**3**	**12**	64	0	3	16
24	0	2	3	65	0	4	16
25	1	7	2	66	0	4	10
26	0	5	13	67	0	2	17
27	0	3	6	68	0	4	13
28	0	3	14	69	0	4	11
29	0	3	16	70	0	5	16
30	0	2	16	71	0	5	14
31	0	3	16	72	0	4	12
32	0	2	11	73	0	4	16
33	0	5	13	74	0	4	10
34	0	3	16	75	0	3	15
35	0	5	11	76	0	4	15
36	0	3	16	**77**	**1**	**3**	**13**
37	0	3	9	78	0	7	13
38	1	5	6	79	0	2	13
39	0	6	9	80	1	7	6
40	**1**	**5**	**12**	81	0	4	13
41	1	5	1				

Table 4. Influence diagnostics for modified Kyphosis data.

Index	CD (1.00)	|DFFITS| (0.577)	|GDFFITS| (0.600)	Index	CD (1.00)	|DFFITS| (0.577)	|GDFFITS| (0.600)
1	0.02	−0.243	−0.106	42	0.00	−0.040	−0.009
2	0.00	−0.030	−0.006	**43**	0.23	**−0.858**	**−1.696**
3	0.04	0.347	0.471	44	0.07	−0.457	−0.521
4	0.07	−0.462	−0.652	45	0.00	−0.008	−0.003
5	0.00	−0.040	−0.010	**46**	0.04	0.350	**1.023**
6	0.00	0.018	−0.002	47	0.00	−0.040	−0.010
7	0.00	0.035	−0.001	48	0.00	−0.088	−0.035
8	0.00	−0.008	−0.003	**49**	0.08	0.505	**0.747**
9	0.00	0.018	−0.002	50	0.00	−0.013	−0.003
10	0.05	0.389	**0.688**	51	0.00	−0.110	−0.072
11	0.05	0.378	**0.992**	52	0.00	0.035	−0.001
12	0.00	−0.008	−0.003	53	0.07	0.450	0.049
13	0.05	−0.396	−0.520	54	0.00	0.035	−0.001
14	0.00	−0.060	−0.023	55	0.00	−0.040	−0.010
15	0.00	0.021	−0.002	56	0.00	−0.077	−0.020
16	0.00	−0.008	−0.003	57	0.00	−0.040	−0.009
17	0.01	−0.140	−0.047	58	0.02	0.256	0.442
18	0.00	−0.088	−0.035	59	0.03	−0.302	−0.503
19	0.00	−0.069	−0.015	60	0.00	−0.040	−0.009
20	0.00	−0.101	−0.050	61	0.06	0.418	0.362
21	0.00	0.018	−0.002	62	0.04	0.361	0.419
22	0.03	0.287	0.246	63	0.07	−0.457	−0.521
23	0.04	0.335	**1.199**	64	0.00	−0.008	−0.003
24	0.05	−0.369	−0.137	65	0.00	−0.032	−0.007
25	0.03	0.284	0.081	66	0.00	−0.083	−0.039
26	0.00	−0.088	−0.035	67	0.00	0.035	−0.001
27	0.01	−0.201	−0.076	68	0.00	−0.053	−0.017
28	0.00	−0.030	−0.006	69	0.00	−0.069	−0.030
29	0.00	−0.008	−0.003	70	0.00	−0.069	−0.015
30	0.00	0.018	−0.002	71	0.00	−0.083	−0.027
31	0.00	−0.008	−0.003	72	0.00	−0.060	−0.023
32	0.00	−0.065	−0.009	73	0.00	−0.032	−0.007
33	0.00	−0.088	−0.035	74	0.00	−0.083	−0.039
34	0.00	−0.008	−0.003	75	0.00	−0.020	−0.005
35	0.00	−0.100	−0.057	76	0.00	−0.040	−0.010
36	0.00	−0.008	−0.003	**77**	0.04	0.343	**1.293**
37	0.00	−0.106	−0.030	78	0.02	−0.255	−0.213
38	0.03	0.277	0.328	79	0.00	−0.029	−0.004
39	0.01	−0.193	−0.198	80	0.03	0.325	0.253
40	0.03	0.295	**0.772**	81	0.00	−0.053	−0.017
41	0.04	0.339	0.190				

Note: GDFFITS, generalized difference of fits; DFFITS, difference of fits.

observe a similar picture when we look at the index plot of Cook's distance and DFFITS as shown in Figure 3a and b.

Now we apply our newly proposed method to this data set. A robust fit based on the BACON algorithm identifies eight observations (cases 10, 11, 23, 40, 43, 46, 49 and 77) as suspects. We compute the GDFFITS values for the entire data set based on a set that excludes these suspect cases and the results are presented in Table 4. This table shows that the GDFFITS values corresponding to all eight suspect cases exceed the cut-off point and hence can be successfully identified as influential observations. It is worth mentioning that we have identified the case 49 as influential which was masked earlier. The index plot of GDFFITS, as given in Figure 3c, shows that all eight influential observations are correctly identified.

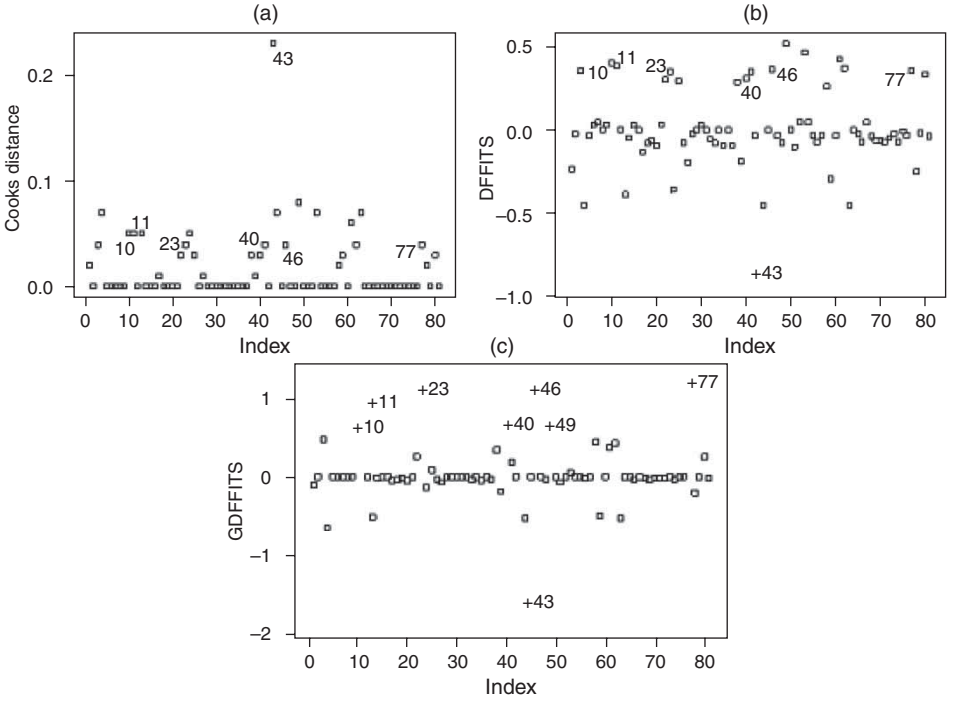

Figure 3. Index plots of: (a) Cook's distance, (b) difference of fits (DFFITS) and (c) generalized DFFITS (GDFFITS) for modified Kyphosis data.

4.3 *Modified Finney data*

We now consider another data set as a real example given in Ref. [12]. The original data set was obtained to study the effect of the rate and volume of air inspired on a transient vasoconstriction in the skin of the digits. The nature of the measurement process was such that only the occurrence and non-occurrence of vasoconstriction could be reliably measured. We modify the data by putting five outliers (cases 3, 10, 11, 20 and 21) where occurrence (1) and non-occurrence (0) are replaced by each other. The modified data set is presented in Table 5 where the suspect cases are indicated by boldfaces.

This data set was analyzed extensively in Ref. [21]. Looking at the pattern of occurrence and non-occurrence with 25% and 75% contours in relation to rate and volume of the original data, it has been reported [21] that this data set might contain two outliers (cases 4 and 18). Now we apply the newly proposed algorithm for the identification of influential observations. Here the deletion set D contains seven cases (3, 4, 10, 11, 18, 20 and 21), suggested by Figure 4. We re-estimate the logistic model without the seven cases and compute GDFFITS values for the whole data set. Table 6 presents influence diagnostics for the modified Finney data. We see from this table that the single-deletion diagnostic Cook's distance and DFFITS fail to detect any one influential observation. On the contrary, our newly proposed GDFFITS successfully identifies all seven influential cases and with three more cases (13, 32 and 39) that were masked before the re-estimation. Figure 5 shows the same evidence from the graphical point of view.

5. Simulation study

In this section, we report a Monte Carlo simulation study that is designed to compare the performance of our newly proposed technique with the existing popular influence measures like Cook's distance and DFFITS.

Table 5. Modified Finney data.

Index	Response	Volume	Rate	Index	Response	Volume	Rate
1	1	3.70	0.820	**21**	**0**	**2.50**	**2.000**
2	1	3.50	1.090	22	0	0.95	1.360
3	**0**	**1.25**	**2.500**	23	0	1.35	1.350
4	**1**	**0.75**	**1.500**	24	0	1.50	1.360
5	1	0.80	3.200	25	1	1.60	1.780
6	1	0.70	3.500	26	0	0.60	1.500
7	0	0.60	0.750	27	1	1.80	1.500
8	0	1.10	1.700	28	0	0.95	1.900
9	0	0.90	0.750	29	1	1.90	0.950
10	**1**	**0.90**	**0.450**	30	0	1.60	0.400
11	**1**	**0.80**	**0.570**	31	1	2.70	0.750
12	0	0.55	2.750	32	0	2.35	0.030
13	0	0.60	3.000	33	0	1.10	1.830
14	1	1.40	2.330	34	1	1.10	2.200
15	1	0.75	3.750	35	1	1.20	2.000
16	1	2.30	1.640	36	1	0.80	3.330
17	1	3.20	1.600	37	0	0.95	1.900
18	**1**	**0.85**	**1.420**	38	0	0.75	1.900
19	0	1.70	1.060	39	1	1.30	1.630
20	**0**	**1.80**	**1.800**				

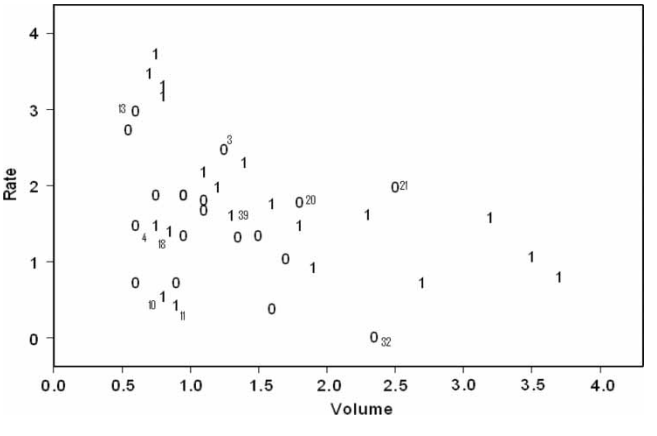

Figure 4. Character plot of occurrence and non-occurrence for modified Finney data.

First, we present an example to show how we design our experiment. Initially, we have simulated 50 observations. For the explanatory variable X, the first 25 observations are generated from Uniform (10, 100) and the last 25 observations are generated from Uniform (50, 500). To generate the original Y values, we set the first 25 values of Y at 0 and the last 25 at 1, so that the general pattern of this design is that the Y values corresponding to bigger X values is more likely to have the value 1 and the data set is presented in Table 7.

This design is synonymous to a real-life experiment where persons with higher cholesterol levels are more susceptible to heart disease. To generate a single influential observation, we change the value of Y corresponding to the largest X value (*i.e.*, the 40th observation) to 0. This changed point should be an influential observation as it appears against the pattern of the majority of the data. We compute the Cook's distance and DFFITS for this data which is presented in Table 8.

Table 6. Influence diagnostics for modified Finney data.

Index	CD (1.00)	|DFFITS| (0.832)	|GDFFITS| (0.918)	Index	CD (1.00)	|DFFITS| (0.832)	|GDFFITS| (0.918)
1	0.0154	0.2147	0.0000	**21**	0.1362	−0.6392	**−12.6585**
2	0.0141	0.2060	0.0000	22	0.0110	−0.1820	−0.0001
3	0.0228	−0.2613	**−3.7293**	23	0.0108	−0.1796	−0.0467
4	0.0431	0.3596	**6.7722**	24	0.0119	−0.1886	−0.2572
5	0.0300	0.3001	0.0065	25	0.0089	0.1634	0.0189
6	0.0400	0.3464	0.0018	26	0.0663	−0.4460	0.0000
7	0.0143	−0.2068	0.0000	27	0.0104	0.1767	0.0126
8	0.0093	−0.1674	−0.0293	28	0.0099	−0.1726	−0.0202
9	0.0151	−0.2128	0.0000	29	0.0186	0.2360	0.3810
10	0.1398	0.6477	**11.3037**	30	0.0245	−0.2711	−0.0003
11	0.1386	0.6449	**11.5619**	31	0.0228	0.2615	0.0000
12	0.0221	−0.2575	−0.1454	**32**	0.0695	−0.4567	**−1.2930**
13	0.0312	−0.3060	**−1.9011**	33	0.0094	−0.1683	−0.0886
14	0.0106	0.1781	0.0023	34	0.0122	0.1913	0.4066
15	0.0434	0.3609	0.0001	35	0.0106	0.1785	0.4288
16	0.0124	0.1927	0.0000	36	0.0325	0.3122	0.0019
17	0.0119	0.1890	0.0000	37	0.0099	−0.1726	−0.0202
18	0.0384	0.3395	**6.3102**	38	0.0110	−0.1813	−0.0010
19	0.0171	−0.2265	−0.3332	**39**	0.0116	0.1865	**1.1413**
20	0.0239	−0.2680	**−4.9304**				

Note: GDFFITS, generalized difference of fits; DFFITS, difference of fits.

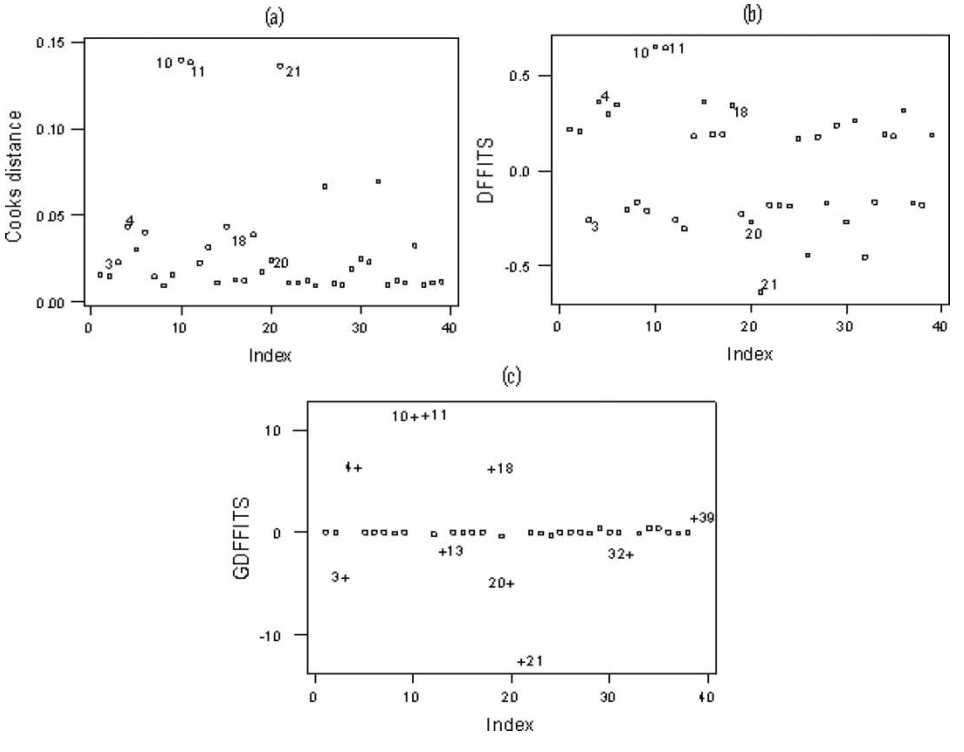

Figure 5. Index plots of: (a) Cooks distance, (b) difference of fits (DFFITS) and (c) generalized DFFITS (GDFFITS) for modified Finney data.

Table 7. Simulated data.

Index	X	Y (original)	Y (single IO)	Y (10% IOs)
1	17.744	0	0	0
2	72.468	0	0	0
3	82.386	0	0	0
4	39.326	0	0	0
5	55.374	0	0	0
6	85.402	0	0	0
7	55.138	0	0	0
8	56.423	0	0	0
9	12.319	0	0	0
10	17.182	0	0	0
11	91.767	0	0	0
12	55.864	0	0	0
13	41.782	0	0	0
14	62.915	0	0	0
15	91.929	0	0	0
16	29.948	0	0	0
17	50.865	0	0	0
18	46.458	0	0	0
19	28.992	0	0	0
20	99.538	0	0	0
21	63.987	0	0	0
22	61.178	0	0	0
23	10.519	0	0	0
24	35.931	0	0	0
25	75.225	0	0	0
26	399.007	1	1	1
27	132.205	1	1	1
28	333.290	1	1	1
29	142.059	1	1	1
30	200.090	1	1	1
31	**69.825**	**1**	1	1
32	278.651	1	1	1
33	153.545	1	1	1
34	**492.174**	**1**	1	**0**
35	**463.957**	**1**	1	**0**
36	267.007	1	1	1
37	167.421	1	1	1
38	**465.084**	**1**	1	**0**
39	255.983	1	1	1
40	**493.902**	**1**	**0**	**0**
41	432.498	1	1	1
42	205.499	1	1	1
43	131.446	1	1	1
44	293.712	1	1	1
45	359.356	1	1	1
46	463.904	1	1	1
47	343.309	1	1	1
48	98.252	1	1	1
49	**487.862**	**1**	1	**0**
50	445.410	1	1	1

Note: Influential observation.

Index plots of Cook's distance and DFFITS as shown in Figure 6a and b suggest that observation 40 is an influential observation. We observe from Table 8 that although the Cook's distance with a cut-off value 1 fails to identify any influential observations, DFFITS can successfully identify the observation number 40 as an influential observation. Our proposed GDFFITS method can also successfully identify the influential case. Now we change the Y values corresponding to the

Table 8. Influence diagnostics for simulated data with a single influential observation.

Index	CD (1.00)	|DFFITS| (0.60)	|GDFFITS| (0.61)	Index	CD (1.00)	|DFFITS| (0.60)	|GDFFITS| (0.61)
1	0.003	−0.073	−0.002	26	0.000	0.025	0.000
2	0.007	−0.118	−0.076	27	0.033	0.263	0.143
3	0.008	−0.125	−0.136	28	0.005	0.101	0.000
4	0.004	−0.093	−0.009	29	0.030	0.249	0.070
5	0.006	−0.106	−0.026	30	0.018	0.191	0.001
6	0.008	−0.127	−0.163	**31**	0.068	0.382	**1.100**
7	0.006	−0.106	−0.026	32	0.010	0.141	0.000
8	0.006	−0.107	−0.028	33	0.027	0.235	0.029
9	0.002	−0.067	−0.001	34	0.013	−0.160	0.000
10	0.003	−0.072	−0.002	35	0.004	−0.093	0.000
11	0.009	−0.131	−0.247	36	0.011	0.149	0.000
12	0.006	−0.106	−0.027	37	0.023	0.220	0.010
13	0.005	−0.095	−0.010	38	0.005	−0.095	0.000
14	0.006	−0.112	−0.043	39	0.012	0.155	0.000
15	0.009	−0.131	−0.250	**40**	0.629	**−1.274**	**−15.100**
16	0.004	−0.084	−0.004	41	0.000	−0.030	0.000
17	0.005	−0.102	−0.019	42	0.017	0.187	0.001
18	0.005	−0.099	−0.014	43	0.033	0.264	0.151
19	0.004	−0.084	−0.004	44	0.009	0.132	0.000
20	0.009	−0.136	−0.438	45	0.003	0.075	0.000
21	0.006	−0.112	−0.046	46	0.004	−0.093	0.000
22	0.006	−0.110	−0.038	47	0.004	0.091	0.000
23	0.002	−0.065	−0.001	48	0.048	0.321	0.559
24	0.004	−0.090	−0.007	49	0.011	−0.149	0.000
25	0.007	−0.120	−0.090	50	0.001	−0.054	0.000

Note: GDFFITS, generalized difference of fits; DFFITS, difference of fits.

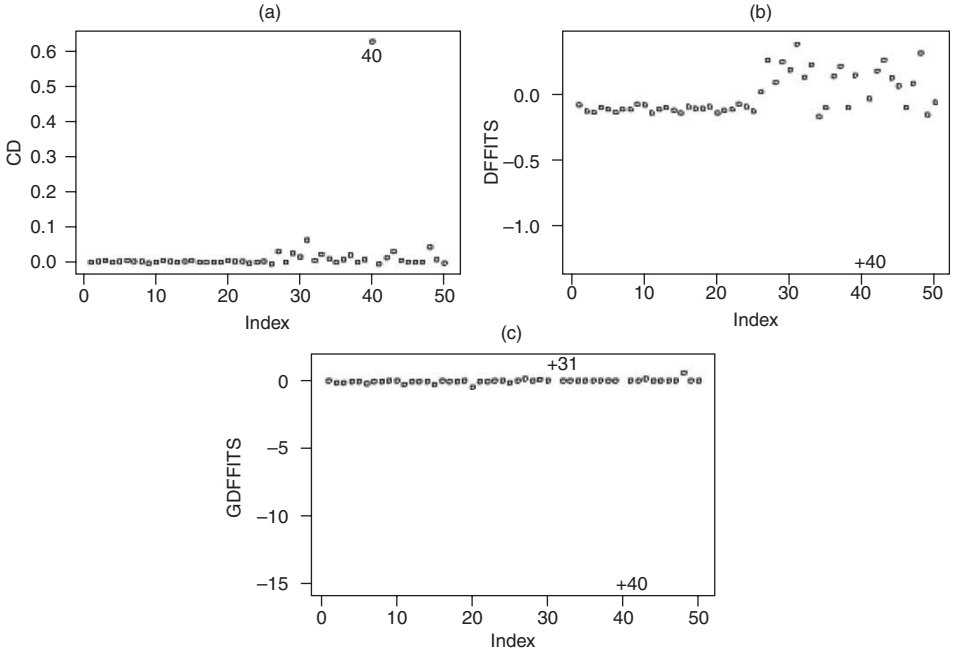

Figure 6. Index plots of: (a) Cook's distance, (b) difference of fits (DFFITS) and (c) generalized DFFITS (GDFFITS) for the simulated data with a single influential observation.

biggest five X values to generate 10% influential cases in the same data. The five influential cases are 34, 35, 38, 40 and 49 as shown in Table 7. We compute the Cook's distance and DFFITS for the new (10% influential cases) data and the results are presented in Table 9.

Although the index plots of Cook's distance and DFFITS as shown in Figure 7a and b suggest that observations 34, 35, 38, 40 and 49 are influential, we observe from Table 9 that the Cook's distance fails to identify even a single observation as influential and DFFITS fails to identify cases 35 and 38 as influential. Our proposed GDFFITS can successfully identify all the five influential cases with one more (case 31). Similar remarks may apply with Figure 7c.

We carry out a simulation study to measure the performance of the newly proposed measure for different sample sizes with different levels of influential observations for logistic regression. We consider $n = 30$, 50 and 100 with $\gamma = 10\%$ and 20% influential cases and our results are based on 10,000 simulations. We generate $100(1-\gamma)\%$ of X from Uniform (10, 30). We compute the values of

$$\pi_i(x) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}, \quad i = 1, 2, \ldots, n,$$

using $\beta_0 = -78.8$ and $\beta_1 = 4$; and $\beta_0 = -12.47$ and $\beta_1 = 1.741$ for halves of X data, respectively. The Y values corresponding to these X values are generated from a Bernoulli (0, 1) variable as $\pi_i(x) < 0.5(= 0)$ and $\pi_i(x) \geq 0.5(= 1)$. The remaining $100\,\gamma\%$ of X values are generated from Uniform (100, 110). According to our previous configuration, the Y values corresponding to our last $100\,\gamma\%$ of X values should be 1 each, but we changed them to 0 to make the last $100\,\gamma\%$ observations influential.

Table 9. Influence diagnostics for simulated data with 10% influential cases.

Index	CD (1.00)	|DFFITS| (0.60)	|GDFFITS| (0.632)	Index	CD (1.00)	|DFFITS| (0.60)	|GDFFITS| (0.632)
1	0.004	−0.090	−0.002	26	0.016	0.180	0.000
2	0.005	−0.101	−0.076	27	0.024	0.220	0.143
3	0.005	−0.103	−0.136	28	0.017	0.182	0.000
4	0.005	−0.095	−0.009	29	0.022	0.212	0.070
5	0.005	−0.098	−0.026	30	0.017	0.184	0.001
6	0.005	−0.104	−0.163	**31**	0.040	0.290	**1.101**
7	0.005	−0.098	−0.026	32	0.016	0.179	0.000
8	0.005	−0.099	−0.028	33	0.020	0.204	0.029
9	0.004	−0.088	−0.001	**34**	0.181	**−0.618**	**−15.039**
10	0.004	−0.090	−0.002	**35**	0.139	−0.539	**−13.985**
11	0.006	−0.105	−0.247	36	0.016	0.178	0.000
12	0.005	−0.099	−0.027	37	0.019	0.197	0.010
13	0.005	−0.096	−0.010	**38**	0.140	−0.542	**−14.027**
14	0.005	−0.100	−0.043	39	0.016	0.178	0.000
15	0.006	−0.105	−0.250	**40**	0.184	**−0.623**	**−15.104**
16	0.004	−0.093	−0.004	41	0.015	0.175	0.000
17	0.005	−0.098	−0.019	42	0.017	0.183	0.001
18	0.005	−0.097	−0.014	43	0.024	0.220	0.151
19	0.004	−0.093	−0.004	44	0.016	0.180	0.000
20	0.006	−0.106	−0.438	45	0.017	0.182	0.000
21	0.005	−0.100	−0.046	46	0.014	0.166	0.000
22	0.005	−0.099	−0.038	47	0.017	0.182	0.000
23	0.004	−0.088	−0.001	48	0.031	0.253	0.560
24	0.005	−0.094	−0.007	**49**	0.174	**−0.605**	**−14.878**
25	0.005	−0.102	−0.090	50	0.015	0.171	0.000

Note: GDFFITS, generalized difference of fits; DFFITS, difference of fits.

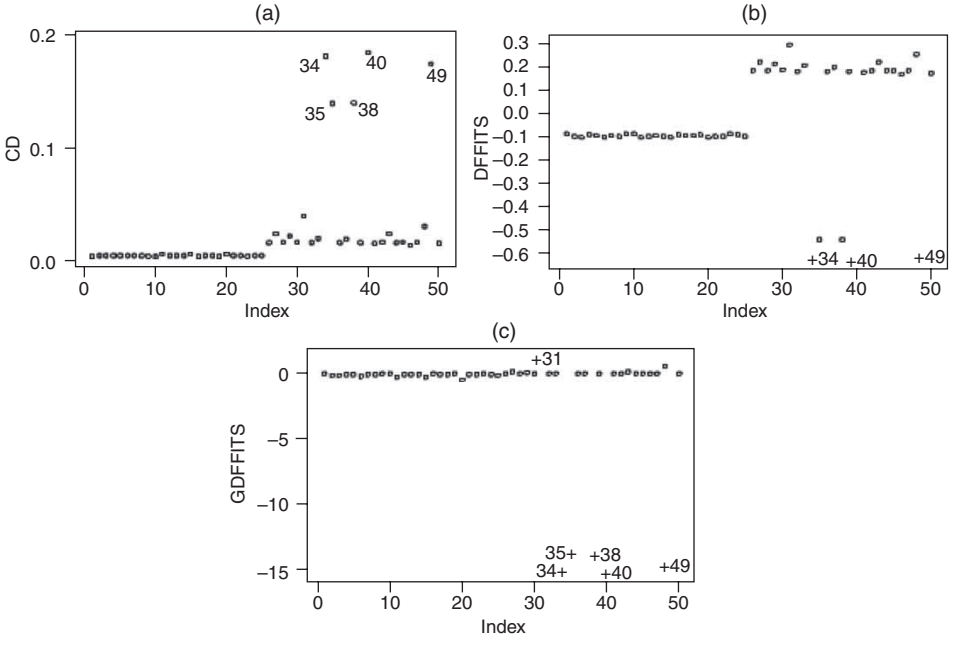

Figure 7. Index plots of: (a) Cook's distance, (b) difference of fits (DFFITS) and (c) generalized DFFITS (GDFFITS) for the simulated data with 10% influential observations.

Table 10 reports the correct identification rate (*i.e.*, total number of influential observations identified/total number of influential observations) and the swamping rate (*i.e.*, total number of non-influential observations identified as influential/total number of non-influential observations) of Cook's distance, DFFITS and GDFFITS for each design.

Table 10. Simulation results.

Sample size	Percentages of influential cases	Diagnostics	Correct identification (in percentages)	Swamping rate (in percentages)
$n = 30$	10	CD	2.0	0.000
		DFFITS	31.7	0.000
		GDFFITS	100.0	3.074
	20	CD	0.0	0.000
		DFFITS	5.2	0.000
		GDFFITS	100.0	0.098
$n = 50$	10	CD	0.0	0.000
		DFFITS	0.0	0.000
		GDFFITS	100.0	2.480
	20	CD	0.0	0.000
		DFFITS	0.0	0.000
		GDFFITS	100.0	0.081
$n = 100$	10	CD	0.0	0.000
		DFFITS	0.0	0.000
		GDFFITS	100.0	2.127
	20	CD	0.0	0.000
		DFFITS	0.0	0.000
		GDFFITS	100.0	0.080

We observe from the results presented in Table 10 that Cook's distance with the cut-off value 1 fails to identify the influential cases on most occasions. The performance of DFFITS is slightly better than Cook's distance but its performance tends to deteriorate with the increase in sample size and level of contamination. Our proposed measure performs very effectively in the identification of multiple influential observations in logistic regression. It can successfully identify all influential cases irrespective of the sample sizes and the proportion of influential cases. However, it has a tendency of swamping non-influential cases, but this swamping rate is very low in every situation and tends to decrease with the increase in sample sizes and interestingly swamps less for higher proportion of the influential cases.

6. Conclusions

In this paper, we propose a new and stepwise group deletion diagnostic method based on the idea of GDFFITS for the identification of multiple influential observations in logistic regression. The numerical examples and simulation study show that the proposed method is able to identify multiple influential observations successfully when the existing commonly used diagnostic methods fail on several occasions.

Acknowledgements

The authors gratefully thank the editor and two anonymous reviewers for their helpful comments and suggestions that helped in the substantial improvement of the present version of this article.

References

[1] A.C. Atkinson and M. Riani, *Robust Diagnostic Regression Analysis*, Springer-Verlag, New York, 2000.
[2] D.A. Belsley, E. Kuh, and R.E. Welsch, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, Wiley, New York, 1980.
[3] N. Billor, A.S. Hadi, and F. Vellman, *BACON: Blocked adaptive computationally efficient outlier nominator*, Computat. Stat. Data Anal. 34 (2000), pp. 279–298.
[4] 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., pp. 3–18, Wiley, New York, 1980.
[5] J. M. Chambers and T. J. Hastie, *Statistical Models in S*, Wadsworth and Brooks, Pacific Grove, CA, 1992.
[6] S. Chatterjee and A.S. Hadi, *Sensitivity Analysis in Linear Regression*, Wiley, New York, 1988.
[7] S. Chatterjee and A.S. Hadi, *Regression Analysis by Examples*, 4th ed., Wiley, New York, 2006.
[8] R.D. Cook, *Detection of influential observations in linear regression*, Technometrics 19 (1977), pp. 15–18.
[9] R.D. Cook, *Assessment of local influence*, J. Roy. Stat. Soc., Ser-B 48 (1986), pp. 131–169.
[10] R.D. Cook and S. Weisberg, *Residuals and Influence in Regression*, Chapman and Hall, London, 1982.
[11] P. Davies, A.H.M.R. Imon, and M.M. Ali, *A conditional expectation method for improved residual estimation and outlier identification in linear regression*, Int. J. Stat. Sci., special issue in honour of Professor M.S. Huq (2004), pp. 191–208.
[12] D.J. Finney, *The estimation from individual records of the relationship between dose and quantile response*, Biometrika 34 (1947), pp. 320–334.
[13] A.S. Hadi, *A new measure of overall potential influence in linear regression*, Comput. Stat. Data Anal. 14 (1992), pp. 1–27.
[14] A.S. Hadi and J.S. Simonoff, *Procedure for the identification of outliers in linear models*, J. Am. Stat. Assoc. 88 (1993), pp. 1264–1272.
[15] D.W. Hosmer and S. Lemeshow, *Applied Logistic Regression*, Wiley, New York, 2000.
[16] A.H.M.R. Imon, *Identifying multiple influential observations in linear regression*, J. Appl. Stat. 32 (2005), pp. 929–946.
[17] A.H.M.R. Imon, *Identification of high leverage points in logistic regression*, Pak. J. Stat. 22 (2006), pp. 147–156.
[18] A.H.M.R. Imon and A.S. Hadi, *Identification of multiple outliers in logistic regression*, Comm. Stat., Theory and Methods 37 (2008), pp. 1697–1709.
[19] A.J. Lawrance, *Local and deletion influence* in *Directions in Robust Statistics and Diagnostics*, Part I, W. Stahel and S. Weisberg, eds., pp. 141–157, Springer, Berlin, 1991.

[20] W.Y. Poon and Y.S. Poon, *Conformal normal curvature and assessment of local influence*, J. Roy. Stat. Soc., Ser-B 61 (1999), pp. 51–56.
[21] D. Pregibon, *Logistic regression diagnostics*, Ann. Stat. 9 (1981), pp. 977–986.
[22] P.J. Rousseeuw and A. Leroy, *Robust Regression and Outlier Detection*, Wiley, New York, 1987.
[23] T.P. Ryan, *Modern Regression Methods*, Wiley, New York, 1997.
[24] V.J. Yohai, *High breakdown point and high efficiency estimates for regression*, Ann. Stat., 15 (1987), pp. 642–656.