The Korean Communications
in Statistics Vol. 9, No. 1, 2002
pp. 197-212

Influential Points in GLMs via Backwards Stepping

Kwang Mo Jeong[1], Hae Young Oh[2]

Abstract

When assessing goodness-of-fit of a model, a small subset of deviating observations can give rise to a significant lack of fit. It is therefore important to identify such observations and to assess their effects on various aspects of analysis. A Cook's distance measure is usually used to detect influential observation. But it sometimes is not fully effective in identifying truly influential set of observations because there may exist masking or swamping effects. In this paper we confine our attention to influential subset in GLMs such as logistic regression models and loglinear models.

We modify a backwards stepping algorithm, which was originally suggested for detecting outlying cells in contingency tables, to detect influential observations in GLMs. The algorithm consists of two steps, the identification step and the testing step. In identification step we identify influential observations based on influencial measures such as Cook's distances. On the other hand in testing step we test the subset of identified observations to be significant or not. Finally we explain the proposed method through two types of dataset related to logistic regression model and loglinear model, respectively.

1. INTRODUCTION

When assessing goodness-of-fit of a model, a small subset of deviating observations can give rise to a significant lack of fit. It is therefore important to identify such observations and to assess their effects on various aspects of analysis. Many statisticians have been interested in identifying outliers which are usually identified via residuals. On the other hand the observations which greatly affect parameter estimates in regression models are called influential points, as was firstly discussed by Cook(1977). A Cook's distance measure is usually used to detect influential observation. But it sometimes is not fully effective in identifying truly influential set of observations because there may exist masking or swamping effects. Influential points may be masked by other points, or any points which are in fact not

[1] Professor, Department of Statistics, Research Institute of Computer, Information and Communication, Pusan National University, 609-735
E-mail : kmjung@pusan.ac.kr

[2] Graduate, Department of Statistics, Pusan National University, 609-735

influential may be regarded as influential because of swamping effect. In both cases we are confronted with difficulty in detecting truly influential points. Many researchers, for example, Hoaglin and Welsch(1978), Belsley, Kuh and Welsch (1980), Cook and Weisberg(1980), Atkinson(1981), and Cook and Weisberg(1982) studied this problem. We omit detailed discussions in linear models.

For general discussions on generalized linear model (GLM) we refer to, in particular, McCullagh and Nelder(1983) and Agresti(1990), and here we only confine our attention to influential subset in GLMs. Two types of the most popular GLMs are the logistic regression model and the loglinear model. In logistic regression model Pregibon(1981) discussed detecting influential points using perturbation. On the other hand Simonoff(1988) suggested a method, which is known as backwards stepping procedure, to detect outliers in contingency tables. The backwards stepping algorithm of Simonoff(1988) consists of two steps, the identification step and the testing step. In this paper we slightly modify the algorithm to be suitable for detection of influential points and hence we introduce appropriate measures to identify influential points in the identification step. Chi-squared test statistics can be used as deviance measures between successive steps in the testing step. The likelihood ratio test (LRT) statistic and the Pearson statistic are popular ones. In Section 2 we briefly overview GLMs with some discussions on the estimation of parameters. Diagnostic measures in GLMs such as influential measures, residuals, and goodness-of-fit statistics will be discussed in Section 3. In Section 4 we introduce a modified backwards stepping algorithm for detecting influential observations. Finally we explain the proposed method through two types of dataset related to logistic regression model and loglinear model, respectively. The calculations in this paper were performed via Splus software programming.

2. Generalized Linear Model

First we briefly review GLM and its estimation techniques. Let $X_1,\cdots,X_p$ be p explanatory variables, and let y be a response variable with mean $\mathrm{E}(y)=\mu$. We assume a relationship between $\eta(\mu)$, a function of μ, and a linear predictor of explanatory variables $\{x_1,\cdots,x_p\}$ of the form

$$\eta(\mu)=\beta_0+\beta_1 x_1+\cdots+\beta_p x_p. \tag{2.1}$$

This type of linear model is called a GLM in the sense that the relationship is linear in terms of $\eta(\mu)$. The function $\eta(\mu)$ is called a link function and the right hand side of (2.1) denotes a systematic component of GLM. To simplify the notation we denote the linear predictor in (2.1) as

$$\mathbf{x}\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p,$$

where $\mathbf{x}$ is a vector $(1, x_1, \cdots, x_p)'$ of explanatory variables and $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)'$. Logistic regression models and loglinear models are typical ones of GLMs.

As a special case of GLM we first consider a logistic model for binary responses taking values of one and zero with respective probabilities π and $1-\pi$. Hereafter in logistic regression model we use the notation π instead of μ. The link function of the form

$$\mathrm{logit}(\pi) = \mathbf{x}\boldsymbol{\beta} \tag{2.2}$$

is called a logit link with an abbreviated notation $\mathrm{logit}(\pi) = \log(\pi/(1-\pi))$. On the other hand a loglinear model of the form

$$\log(\mu) = \mathbf{x}\boldsymbol{\beta} \tag{2.3}$$

is defined if the response variable is distributed as Poisson with mean μ. The link function of (2.3) is called a log link. We may consider various kinds of GLM according to link functions. We refer to McCullagh and Nelder(1983) for more detailed discussions on GLMs.

Next we consider the estimation of parameters based on likelihood function. Let y be a response with probability function

$$f(y; \eta) = \exp\{y\eta - a(\eta) + b(y)\}. \tag{2.4}$$

We note that $\eta = \mathrm{logit}(\pi)$, $a(\eta) = \log(1 + e^{\eta})$ and $b(y) = 0$ for a binary response variable. On the other hand if y is a Poisson variable with mean μ, then $\eta = \log(\mu)$ and $a(\eta) = \exp(\eta)$ with $b(y) = -\log(y!)$. Hence the log-likelihood function can be written as

$$l(\mathbf{X}\boldsymbol{\beta}; \mathbf{y}) = \sum_i \{y_i \mathbf{x}_i \boldsymbol{\beta} - a(\mathbf{x}_i \boldsymbol{\beta}) + b(y_i)\}, \tag{2.5}$$

where $\mathbf{X}$ is a design matrix having $\mathbf{x}_i$ as its ith column. The maximum likelihood estimates(MLE) are found by solving the likelihood equation given by $\partial l(\mathbf{X}\boldsymbol{\beta}; \mathbf{y})/\partial\boldsymbol{\beta} = \mathbf{0}$. But the MLEs of GLM cannot be obtained in explicit form in general except some simple models. When there does not exist direct estimates we use the Newton-Raphson iterative algorithm to find MLEs from the likelihood equations. Detailed iterative steps are routine and we omit them here.

After the initial value $\boldsymbol{\beta}^{(0)}$ is assigned we can express $\boldsymbol{\beta}^{(t)}$, the value at tth iteration, in the following recursive form

$$\beta^{(t+1)} = \beta^{(t)} + [\,\mathbf{X}^T \widehat{\mathbf{W}}^{(t)} \mathbf{X}\,]^{-1} \mathbf{X}^T (\mathbf{y} - \hat{\mu}^{(t)}). \tag{2.6}$$

We note that the $\widehat{\mathbf{W}}^{(t)}$ is an $n \times n$ diagonal matrix with diagonal elements $w_i^{(t)}$; $w_i^{(t)} = \hat{\pi}^{(t)}(1-\hat{\pi}^{(t)})$ for a logistic model, and $w_i^{(t)} = \hat{\mu}_i^{(t)}$ for a loglinear model. The MLE $\hat{\beta}$ is obtained when a certain convergence criterion is attained and we also obtain the covariance of $\hat{\beta}$, given by $[\,\mathbf{X}^T \widehat{\mathbf{W}} \mathbf{X}\,]^{-1}$, where $\widehat{\mathbf{W}}$ is a diagonal matrix with estimated values of $w_i^{(t)}$ at the final stage of iteration. The matrix $\mathbf{X}^T \widehat{\mathbf{W}}\, \mathbf{X}$ is called the Fisher information matrix.

It sometimes is useful to view the iterative process outlined above in the sense of iteratively reweighted least squares estimation (IRWLSE). We briefly review the discussion on IRWLSE by Pregibon(1981). Let a pseudo-observation vector $\mathbf{z}^{(t)}$ be defined by

$$\mathbf{z}^{(t)} = \mathbf{X}\beta^{(t)} + \widehat{W}^{(t)\,-1}(\mathbf{y} - \hat{\mu}^{(t)}).$$

Then $\beta^{(t+1)}$ in (2.6) can be rewritten in the form of IRWLSE as

$$\beta^{(t+1)} = (\mathbf{X}^T \widehat{\mathbf{W}}^{(t)} \mathbf{X})^{-1} \mathbf{X}^T \widehat{\mathbf{W}}^{(t)} \mathbf{z}^{(t)}.$$

3. Diagnostic Measures

3.1 Influence Measures

Residuals, which are differences between observations and estimated values in usual sense, are most commonly used for identifying outliers, but we are not certain whether the residuals can also be effective for detecting influential observations. Influential observations are the ones which greatly affect the estimate $\hat{\beta}$. In linear regression models diagonal elements of hat matrix defined by

$$\mathbf{H} = \mathbf{X}[\,\mathbf{X}^T\mathbf{X}\,]^{-1}\mathbf{X}^T$$

is a useful measure for leverage points. We omit here detailed discussions on the properties of hat matrix. An analogue of hat matrix in GLMs can be given in the form

$$\mathbf{H} = \widehat{\mathbf{W}}^{\frac{1}{2}} \mathbf{X}[\,\mathbf{X}^T \widehat{\mathbf{W}} \mathbf{X}\,]^{-1} \mathbf{X}^T \widehat{\mathbf{W}}^{\frac{1}{2}}. \tag{3.1}$$

Diagonal elements of $\mathbf{H}$, denoted by h_i, are useful in detecting high leverage points. Influencial points tend to have large values of h_i but large residuals are seldom associated with high leverage points, whereas small residuals are typically of the opposite character.

Even though the quantities of residuals and diagonal elements of hat matrix are useful for detecting extreme points, but not for assessing their impact on parameter estimates or fitted values. Let $\hat{\beta}_{(i)}$ be an MLE with the ith observation deleted. Cook(1977) firstly proposed a statistic, which is the so called Cook's distance, of the form

$$c_i = \frac{(\hat{\beta} - \hat{\beta}_{(i)})^T \mathbf{X}^T \mathbf{X} (\hat{\beta} - \hat{\beta}_{(i)})}{(p+1)\,\hat{\phi}} \tag{3.2}$$

as an influence measure for the ith observation, where $\hat{\phi}$ is an estimate of dispersion parameter. The statistic c_i measures an effect on the estimate $\hat{\beta}$ when the ith observation is deleted. As discussed in Pregibon(1981) an analogue of Cook's statistic in GLMs can be written in a generalized form

$$c_i = \frac{(\hat{\beta} - \hat{\beta}_{(i)})^T \mathbf{X}^T \widehat{\mathbf{W}} \mathbf{X} (\hat{\beta} - \hat{\beta}_{(i)})}{(p+1)\,\hat{\phi}} \ . \tag{3.3}$$

In (3.3) the one-step approximation of $\hat{\beta}_{(i)}$ is usually taken to alleviate computational burden.

Hereafter we follow the expressions given in Chambers and Hastie(1993) to represent Cook's distances in terms of norm. If we let $\hat{\eta} = \mathbf{X}\hat{\beta}$ and $\hat{\eta}_{(i)} = \mathbf{X}\hat{\beta}_{(i)}$ the Cook's distance can be written using a norm weighted by $\widehat{\mathbf{W}}$ as

$$c_i = \frac{\| \hat{\eta} - \hat{\eta}_{(i)} \|^2}{(p+1)\,\hat{\phi}} \, . \tag{3.4}$$

Possibly a more useful diagnostic measure is the Cook's distance confined to a subset of parameters of interest, in particular those subset belonging to an individual term in the model. It shows why some of the observations have large Cook's distances for both of the quadratic and linear terms. This type of Cook's distance measure, which is usually called an index influence measure for each parameter, can be expressed as

$$c_i^{\,j} = \frac{\| \hat{f}_j - \hat{f}_{j(i)} \|^2}{\sum_{i=1}^{n} \hat{w}_i \mathrm{Var}(\hat{f}_{ji})} \tag{3.5}$$

where $\hat{f}_j = \mathbf{X}_j\hat{\beta}_j$, with $\mathbf{X}_j$ denoting a subset of model matrix corresponding to a subset of parameters. The corresponding estimate $\hat{\beta}_j$ is defined in a similar way. The estimate $\hat{f}_{j(i)} = \mathbf{X}_j\hat{\beta}_{j(i)}$ is an approximation of $\hat{f}_j$ with the ith observation omitted. The equation (3.5) can be applied to each single parameter of the models in (2.2) and (2.3).

3.2 Goodness-of-Fit Statistics and Residuals

After we obtain fitted values under the assumed model we are to assess goodness-of-fit of a model. The widely used statistics for testing the assumed model are the Pearson statistic and the LRT statistic. The Pearson statistic is of the form

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} \tag{3.6}$$

and the LRT is

$$G^2 = 2\sum_{i=1}^{n} y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right), \tag{3.7}$$

where $\hat{\mu}_i$ denotes the predicted value of y_i under the assumed model. When the assumed model is true these two statistics are known to have the same asymptotic chi-square distribution with an appropriate degree of freedom.

A few observations, which are the so called outliers, can cause a significant lack of fit when we fit a model. Many statisticians have been interested in identifying outliers in various ways. Outliers can usually be identified via residuals. We briefly review three types of residuals in relation to Cook's distance measure. The Pearson residual which is defined by

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{\mathrm{Var}(\hat{\mu}_i)}} \tag{3.8}$$

is directly related to the Pearson statistic in the sense that the residual sum of squares is equal to the Pearson statistic of (3.6). The Cook's distance c_i given in (3.3) may be represented in terms of Pearson residual and leverage value as follows

$$c_i = \frac{(r_i^P)^2 h_i}{(p+1)\hat{\phi}(1-h_i)^2}\ , \tag{3.9}$$

where h_i is a diagonal element of hat matrix given in (3.1). The adjusted residual can be represented as

$$r_i^A = \frac{r_i^P}{\sqrt{(1-h_i)\widehat{\phi}}}\ .$$

On the other hand the deviance residual is defined as

$$r_i^D = \sqrt{2}\, sign(y_i - \widehat{\mu}_i)[y_i \log(\frac{y_i}{\widehat{\mu}_i}) - y_i + \widehat{\mu}_i]^{\frac{1}{2}}, \tag{3.10}$$

where sign(δ) is 1 if δ is positive and -1 otherwise. We also note that the sum of squares of deviance residuals is equal to G^2 given in (3.7).

4. Backwards Stepping for Influential Observations

A backwards stepping algorithm, originally suggested by Simonoff(1988) to detect outlying cells in contingency tables, is based on residuals and goodness-of-fit statistics. We modify the algorithm so that it can be applied to detect influential observations in GLMs. The algorithm consists of two steps, the identification step and the testing step. In identification step we identify influential observations based on the influencial measure of Cook's distances. On the other hand in testing step we test the subset of influential observations detected in identification step to be significant or not. The drop in a goodness-of-fit statistic with the observation of interest deleted can be used as a test statistic in testing steps.

First of all we temporarily determine the number of influential observations, hereafter denoted by K_0, which will be used in identifying observations until K_0 suspected observations are detected. Determining K_0 is a little more difficult problem. We suggest a method of determining K_0 from a scree plot, which is usually used in factor analysis to find the number of common factors by plotting eigenvalues of sample correlation matrix against the order of their magnitudes. In a similar way we plot Cook's distances in the order of their magnitudes and choose the value of K_0 at the number in which the relative magnitudes of Cook's distances decrease steeply.

Identification Step

Step 1. Set the identification step number i equal to 0.

Step 2. Fit an assumed model to the whole dataset of observations.

Step 3. Identify the most influential observation based on Cook's distance measure c_i defined in (3.3).

Step 4. Repeat until $i = K_0$, then we stop the identification step.

After a subset of K_0 observations is temporarily detected to be influential we nextly perform testing steps to guarantee the significance of identified influential observations. As a test statistic we take the drop in LRT statistic after the observation of interest is deleted. Given the overall significance level α we choose $\alpha_0 = \alpha/K_0$ as a common significance level for each testing step. This fact stems from the Bonferroni bound, which is known to be very conservative in controlling Type I error of test. As pointed out by Rosner(1975) this strategy matches fairly well with the anti-conservativeness of backwards stepping algorithm.

Testing Step

Step 1. Set the testing step number i equal to K_0.

Step 2. Construct a test statistic T_i, at the ith step, to be the change in G^2 after the observation of interest is deleted.

Table 4.1. Finney's Data on vaso-constriction

Volume	Rate	Response*	Volume	Rate	Response
3.7	0.825	1	0.4	2	0
3.5	1.09	1	0.95	1.36	0
1.25	2.5	1	1.35	1.35	0
0.75	1.5	1	1.5	1.36	0
0.8	3.2	1	1.6	1.78	1
0.7	3.5	1	0.6	1.5	0
0.6	0.75	0	1.8	1.5	1
1.1	1.7	0	0.95	1.9	0
0.9	0.75	0	1.9	0.95	1
0.9	0.45	0	1.6	0.4	0
0.8	0.57	0	2.7	0.75	1
0.55	2.75	0	2.35	0.30	0
0.6	3.0	0	1.1	1.83	0
1.4	2.33	1	1.1	2.2	1
0.75	3.75	1	1.2	2.0	1
2.3	1.64	1	0.8	3.33	1
3.2	1.6	1	0.95	1.9	0
0.85	1.415	1	0.75	1.9	0
1.7	1.06	0	1.3	1.625	1
1.8	1.8	1			

* binary response indicates the occurrence (1) or nonoccurrence (0)

Step 3. Compare T_i with an appropriate critical value of chi-squared statistic with significance $\alpha_0 = \alpha / K_0$.

Step 4. If T_i is not significant and $i > 1$, decrease i to $i - 1$ and go to Step 2, otherwise no values are significant and hence we have no influential observations.

In testing Step 2 an alternative choice of test statistic is possible by taking X^2 instead of G^2. But we prefer G^2 statistic in the respect that G^2 can be partitioned into chi-squared components.

Example 4.1. To explain the modified backwards stepping algorithm adapted to detect influential observations in logistic regression model we introduce an example from Finney(1947), which had also been explained in Pregibon(1981). The data listed in Table 4.1 were obtained in a controlled study on the effect of the rate and volume of air inspired on a transient vaso-constriction in the skin of the digits.

The nature of the measurements process was such that only the occurrence or nonoccurrence of vaso-constriction could be reliably measured. Three subjects were involved in the study: the first 9 responses, the second contributed 8 responses, and the third contributed 22 responses.

We assume a logistic model of the form

$$\mathrm{logit}(\pi) = \beta_0 + \beta_1 \log(\mathrm{Rate}) + \beta_2 \log(\mathrm{Volume}),$$

where a logarithmic transformation is taken on each explanatory variable. The estimated parameters are

$$\widehat{\beta}_0 = -2.924(1.246),\ \widehat{\beta}_1 = 4.631(1.731),\ \widehat{\beta}_2 = 5.220(1.798)$$

with the estimated standard errors in parentheses. The deviance for the fit is 29.26 on 36 degrees of freedom (df) with P-value 0.220 from the asymptotic chi-squared distribution, and hence there is no gross inadequacies of the assumed model. Influence measures for the whole dataset are listed in Table 4.2. We may temporarily choose $K_0 = 2$ observations from the scree plot of Cook's distances as shown in Figure 4.1. The $4th$ observation has the largest c_i value and the $18th$ observation has the next largest c_i value.

After the $4th$ observation is deleted we fit the same logistic model for the remaining 38 observations. The G^2 value for this subset is 22.42 with 35 df as shown in Table 4.4. Cook's distances in identification steps are given in Table 4.3. As was expected the $18th$ observation has the largest c_i value among the remaining 38 observations after the $4th$ observation is omitted.

Table 4.2. Influential measures for the data of Table 4.1.

Obs	c_i	h_i	r_i^A	c_i^j		
				β_0	β_1	β_2
1	0.002	0.094	0.231	0.000	0.000	0.003
2	0.000	0.043	0.136	0.000	0.000	0.000
3	0.002	0.059	0.295	0.002	0.004	0.003
4	0.410	0.083	3.691	1.098	0.809	0.986
5	0.013	0.111	0.553	0.005	0.017	0.003
6	0.024	0.146	0.650	0.006	0.025	0.000
7	0.000	0.007	-0.030	0.000	0.000	0.000
8	0.021	0.055	-1.045	0.020	0.003	0.003
9	0.000	0.031	-0.090	0.000	0.000	0.000
10	0.000	0.007	-0.027	0.000	0.000	0.000
11	0.000	0.009	-0.034	0.000	0.000	0.000
12	0.017	0.149	-0.547	0.014	0.004	0.028
13	0.047	0.163	-0.849	0.011	0.000	0.034
14	0.001	0.053	0.258	0.002	0.003	0.003
15	0.010	0.126	0.456	0.007	0.017	0.003
16	0.000	0.040	0.157	0.000	0.001	0.000
17	0.000	0.017	0.069	0.000	0.000	0.000
18	0.311	0.090	3.066	0.871	0.635	0.677
19	0.061	0.125	-1.133	0.056	0.032	0.000
20	0.001	0.052	0.243	0.001	0.002	0.003
21	0.000	0.037	-0.105	0.000	0.000	0.000
22	0.007	0.096	-0.431	0.018	0.013	0.012
23	0.029	0.073	-1.056	0.034	0.013	0.000
24	0.050	0.070	-1.412	0.014	0.001	0.012
25	0.002	0.058	0.341	0.002	0.003	0.005
26	0.000	0.052	-0.158	0.001	0.001	0.001
27	0.003	0.066	0.375	0.001	0.003	0.006
28	0.020	0.064	-0.926	0.020	0.003	0.010
29	0.062	0.159	0.992	0.048	0.031	0.002
30	0.000	0.045	-0.094	0.000	0.000	0.000
31	0.054	0.239	0.721	0.010	0.007	0.021
32	0.000	0.091	-0.136	0.001	0.002	0.000
33	0.027	0.051	-1.237	0.008	0.000	0.000
34	0.007	0.059	0.558	0.002	0.007	0.004
35	0.006	0.055	0.553	0.001	0.006	0.005
36	0.011	0.112	0.504	0.006	0.017	0.003
37	0.020	0.064	-0.926	0.020	0.003	0.010
38	0.009	0.098	-0.507	0.019	0.010	0.018
39	0.010	0.053	0.726	0.001	0.001	0.003

From Table 4.4 the G^2 value is 7.36 after these two observations are deleted. The upper $\alpha_0 = \alpha/K_0 = 0.05/2 = 0.025$ percentile for the chi-square distribution with 1 df corresponds to $\chi^2_{0.025}(1) = 3.170$. Hence the G^2 change of 15.07 in the testing step is very significant(P=0.000) compared to the critical value 3.170. Since the first step results in significance we stop the testing procedure at this step and conclude that both *4th* and *18th* observations are influential under the assumed logistic model for Finney's data. This fact exactly coincides with that of Pregibon(1981).

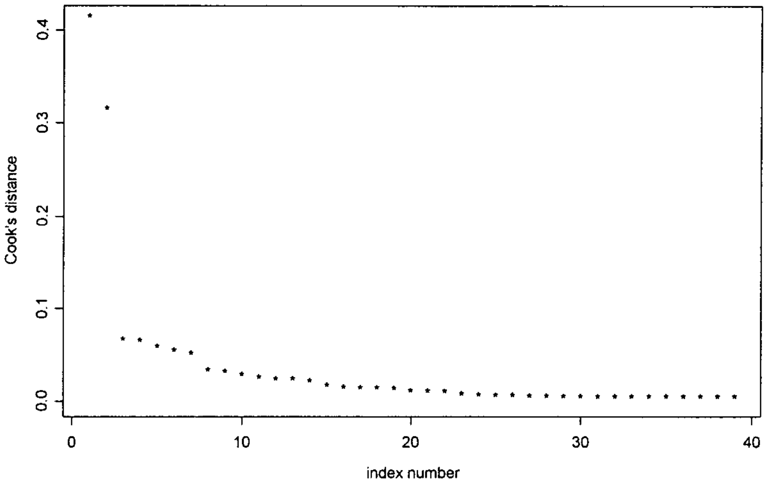

Figure 4.1 Scree plot of Cook's distances for the data of Finney

Table 4.3. Cook's distances in identification steps

obs	steps		obs	steps	
	1st	2nd		1st	2nd
1	0.002	0.000	21	0.000	0.000
2	0.000	0.000	22	0.007	0.000
3	0.002	0.000	23	0.029	0.027
4	0.410	-	24	0.050	0.061
5	0.013	0.013	25	0.002	0.001
6	0.024	0.003	26	0.000	0.000
7	0.000	0.000	27	0.003	0.002
8	0.021	0.021	28	0.020	0.018
9	0.000	0.000	29	0.062	0.128
10	0.000	0.000	30	0.000	0.000
11	0.000	0.000	31	0.054	0.071
12	0.017	0.004	32	0.000	0.000
13	0.047	0.029	33	0.027	0.030
14	0.001	0.003	34	0.007	0.007
15	0.010	0.007	35	0.006	0.006
16	0.000	0.000	36	0.011	0.009
17	0.000	0.000	37	0.020	0.018
18	0.311	1.052	38	0.009	0.002
19	0.061	0.058	39	0.010	0.014
20	0.001	0.000			

Table 4.4. Testing steps for the data of Table 4.1.

step	identified observations	G^2	G^2 changes	df
1	#4, #18	7.36	15.07 *	34
2	#4	22.42	7.23	35
3	-	29.26	-	36

* means the significance at $\alpha_0 = 0.025$

Example 4.2. We nextly illustrate the application of backwards stepping procedure to loglinear models through an artificial dataset of Table 4.5, which was originally designed by Simonoff(1988) to detect outlying cells using backwards stepping procedure. As shown in Table 4.5 the counts in cells (1,2), (1,3) and (2,1) are much larger than those in other cells which are approximately equal to twenty. So these cells are suspected to be outlying cells as explained by Simonoff(1988). Here we are going to detect influencial cells under the independence loglinear model. First we assume the independence loglinear model and find influence measures such as Cook's distances and index distances for each parameter as given in Table 4.6. We temporarily determine K_0 = 4 from the scree plot of Cook's distances as shown in Figure 4.2. Firstly, the (1,1) cell is identified as suspected influential in the first step for the whole dataset. Secondly we identify (2,1) cell as influential by fitting the assumed model to the remaining $5\times5-1$ cells.

In this way we repeat the identification step until K_0 = 4 cells are identified. The identified four cells are (1,1), (2,1), (1,5), (1,4) in the order of identification as shown in Table 4.7. We note that these cells are different from the outlying cells (2,1), (1,3) and (1,2) identified using deleted residuals in Simonoff(1988). Now we are going to test the significance of identified influential cells using drops in G^2 statistics. The significance level at each testing step of Table 4.8 is taken to be a common value of $\alpha_0 = \alpha/K_0 = 0.05/4 = 0.0125$, and hence the critical value is $\chi^2_{0.0125} = 6.25$ with 1 df.

Table 4.5 Hypothetical data by Simonoff

cell	1	2	3	4	5
1	18	41	41	20	21
2	39	20	20	22	22
3	24	20	20	16	18
4	20	20	19	19	19
5	23	18	20	17	20

Table 4.6 Influence measures for the data of Table 4.5

cell	c_i	h_i	r_i^A	c_i^j		
				β_0	β_1	β_2
(1,1)	0.633	0.418	-2.386	0.350	0.567	0.665
(1,2)	0.418	0.412	1.980	0.240	0.357	0.436
(1,3)	0.399	0.413	1.926	0.240	0.357	0.436
(1,4)	0.053	0.378	-0.777	0.049	0.039	0.065
(1,5)	0.067	0.386	-0.856	0.056	0.052	0.080
(2,1)	0.471	0.393	2.218	0.336	0.471	0.467
(2,2)	0.137	0.387	-1.224	0.109	0.141	0.146
(2,3)	0.147	0.388	-1.262	0.109	0.141	0.146
(2,4)	0.005	0.352	0.272	0.007	0.005	0.007
(2,5)	0.000	0.360	-0.017	0.000	0.000	0.000
(3,1)	0.017	0.359	0.467	0.018	0.020	0.015
(3,2)	0.003	0.351	-0.204	0.004	0.005	0.004
(3,3)	0.004	0.353	-0.242	0.004	0.005	0.004
(3,4)	0.001	0.314	-0.132	0.001	0.000	0.001
(3,5)	0.000	0.323	0.096	0.001	0.000	0.001
				0.009	0.010	0.007
(4,1)	0.009	0.357	-0.342			
(4,2)	0.001	0.350	-0.158	0.003	0.003	0.002
(4,3)	0.012	0.351	-0.415	0.013	0.014	0.011
(4,4)	0.025	0.313	0.650	0.037	0.026	0.027
(4,5)	0.009	0.322	0.379	0.013	0.010	0.009
(5,1)	0.005	0.359	0.253	0.004	0.004	0.003
(5,2)	0.030	0.351	-0.641	0.020	0.021	0.017
(5,3)	0.004	0.353	-0.242	0.006	0.007	0.005
(5,4)	0.000	0.314	0.113	0.001	0.000	0.000
(5,5)	0.020	0.027	0.021	0.024	0.019	0.019

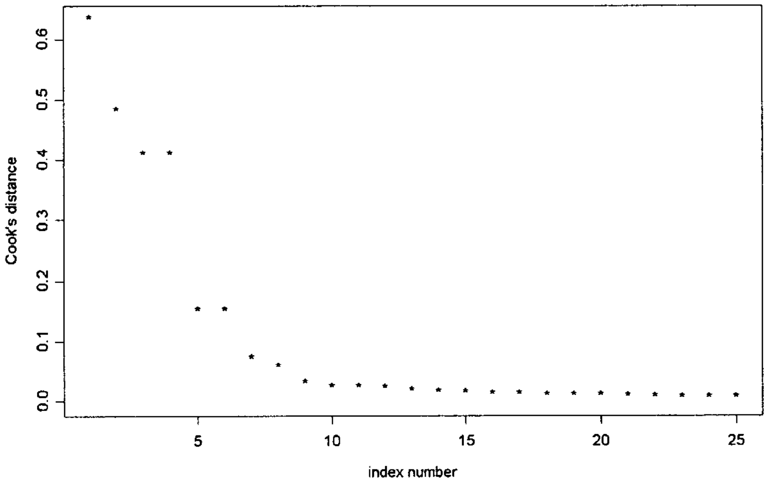

Figure 4.2 Scree plot of Cook's distances for the data of Simonoff

Table 4.7 Cook's distances in identification steps

cell	steps			
	1	2	3	4
(1,1)	0.633	-	-	-
(1,2)	0.418	0.314	0.349	0.242
(1,3)	0.399	0.290	0.322	0.211
(1,4)	0.053	0.270	0.300	0.831
(1,5)	0.067	0.329	0.365	-
(2,1)	0.471	0.371	-	-
(2,2)	0.137	0.120	0.064	0.045
(2,3)	0.147	0.130	0.072	0.053
(2,4)	0.005	0.020	0.091	0.146
(2,5)	0.000	0.003	0.044	0.003
(3,1)	0.017	0.005	0.026	0.031
(3,2)	0.003	0.000	0.003	0.000
(3,3)	0.004	0.000	0.004	0.000
(3,4)	0.001	0.000	0.000	0.000
(3,5)	0.000	0.005	0.001	0.013
(4,1)	0.009	0.135	0.049	0.059
(4,2)	0.001	0.000	0.001	0.000
(4,3)	0.012	0.004	0.016	0.007
(4,4)	0.025	0.050	0.041	0.069
(4,5)	0.009	0.024	0.015	0.000
(5,1)	0.005	0.023	0.003	0.004
(5,2)	0.030	0.019	0.041	0.029
(5,3)	0.004	0.000	0.004	0.000
(5,4)	0.000	0.006	0.001	0.007
(5,5)	0.020	0.045	0.034	0.004

The value 5.529 of G^2 changes is not significant compared to $\chi^2_{0.0125}$ = 6.25 at testing Step 1. This means that the cell (1,4) is omitted from the set of identified cells and we go to the next step. This process is repeated until the set of remaining identified cells are significant. Finally we conclude that (1,1) cell is the only one influential cell detected by backwards stepping algorithm.

Table 4.8 Testing steps for the data of Table 4.5

step	identified cells	G^2	G^2 changes	df
1	(1,1)(2,1)(1,5)(1,4)	1.347	5.529	12
2	(1,1)(2,1)(1,5)	6.876	3.601	13
3	(1,1)(2,1)	10.477	3.542	14
4	(1,1)	14.019	10.659 *	15
5	-	24.678	-	16

* means the significance at α_0=0.0125

5. Conclusion and Future Research

In this paper we discussed a modified backwards stepping algorithm detecting influential observations in GLMs such as logistic regression model and loglinear model. Backwards stepping algorithm, which was originally suggested by Simonoff(1988) for the purpose of detecting outlying cells in contingency tables, consists of two steps, the identification step and the testing step. We slightly modify the identification step using Cook's distance measure to be used for identifying influencial observations in GLMs. As an alternative measure of identification we may choose an index measure for each single parameter of interest by confining the Cook's distance to a subset of parameters.

In identification steps the temporal number of influential observations are predetermined from a scree plot of Cook's distances. Under the assumed model the observation with the largest Cook's distance measure is deleted and we refit the assumed model for the remaining data. This process is repeated until all K_0 observations are identified. In testing step we test the significance of identified observations using the changes in G^2 statistics after the observation of interest is deleted. The usual goodness-of-fit statistic such as Pearson chi-squared statistic or LRT statistic can be used as a test statistic in the testing step, but we prefer the LRT statistic because it can be partitioned into independent chi-squared components. For the given significance level α we take $\alpha_0 = \alpha / K_0$ to be a common significance level at each testing step. This fact stems from the Bonferroni bound which is known to be very conservative in controlling Type I error of test.

We explain the proposed procedure through two types of dataset, one for the logistic regression model and the other for the loglinear model. The first example of Finney(1947), which had also been explained by Pregibon(1981), reveals the same results as before in detecting influential observations. In the second example with artificial dataset of Simonoff(1988) we obtain an influential subset which is quitely different from that of outlying cells detected by Simonoff(1988). We don't have any other benchmark to compare the result in the sense of influential observations.

Finally we remain it as a future research to do a Monte Carlo simulation to study some properties of the proposed procedure such as the power of detecting influential observations and the error rate of incorrectly detecting the observations which are not influential.

References

[1] Agresti, A. (1990), *Categorical Data Analysis*, John Wiley, New York.

[2] Atkinson, A. C. (1981), Two Graphical Displays for Outlying and Influential Observations in Regression, *Biometika*, 68, 13-20.

[3] Belsley, D. A., Kuh, E. and Welsch, R. E.(1980), *Regression Diagnostics: Identifying Influential Data Sources of Collinearity*, John Wiley, New York.

[4] Chambers, J. M. and Hastie, T. J. (1993), *Statistical Models in S*, Chapman and Hall, New York.

[5] Cook, R. D.(1977), Detection of Influential Observations in Linear Regression,*Technometrics*, 19, 15-18.

[6] Cook, R. D, and Weisberg, S.(1980), Characterizations of an Empirical Influence Function for Detecting Influential Cases in Regression, *Technometrics*, 22, 495-508.

[7] Cook, R. D, and Weisberg, S.(1982), *Residuals and Influence in Regression*, Chapman and Hall, New York.

[8] Finney, D. J.(1947), The Estimation from Individual Records of the Relationship Between Dose and Quantal Respose, *Biometrika*, 34, 320-334.

[9] Hoaglin, D. C. and Welsch, R. E.(1978), The Hat Matrix in Regression and ANOVA, *American Statistician*, 32, 17-22.

[10] McCullagh, P. and Nelder, J. A. (1983), *Generalized Linear Models*, Chapman and Hall, New York.

[11] Pregibon, D. (1981), Logistic Regression Diagnostics, The *Annals of Statistics*, 9, 705-724.

[12] Rosner, B. (1975), On the Detection of Many Outliers, *Technometrics,* 17, 221-227.

[13] Simonoff, J. S.(1988), Detecting Outlying Cells in Two-Way Contingency Tables via Backwards Stepping, *Technometrics*, 10, 339-345.

[2001년 8월 접수, 2002년 3월 채택]