Journal of the Korean Statistical Society (2007), **36***: 4, pp 457–469*

MULTIPLE OUTLIER DETECTION IN LOGISTIC REGRESSION BY USING INFLUENCE MATRIX[†]

GWI HYUN LEE[1] AND SUNG HYUN PARK[2]

ABSTRACT

Many procedures are available to identify a single outlier or an isolated influential point in linear regression and logistic regression. But the detection of influential points or multiple outliers is more difficult, owing to masking and swamping problems. The multiple outlier detection methods for logistic regression have not been studied from the points of direct procedure yet. In this paper we consider the direct methods for logistic regression by extending the Peña and Yohai (1995) influence matrix algorithm. We define the influence matrix in logistic regression by using Cook's distance in logistic regression, and test multiple outliers by using the mean shift model. To show accuracy of the proposed multiple outlier detection algorithm, we simulate artificial data including multiple outliers with masking and swamping.

AMS 2000 subject classifications. Primary 62J20; Secondary 62J12.
Keywords. Influence matrix, logistic regression, multiple outliers.

1. INTRODUCTION

Many procedures are available to identify a single outlier or an isolated influential point in linear regression and logistic regression. Beckman and Cook (1983) and Chatterjee and Hadi (1986) surveyed some of procedures in linear regression. Pregibon (1981) developed diagnostic measures to aid the analyst in detecting such observations. But the detection of influential subsets or multiple outliers is more difficult, owing to masking and swamping problems. The multiple outlier detection methods for linear regression have already been studied from two points of view, direct procedure and indirect procedure. The direct methods

Received July 2006; accepted July 2007.
[†]This paper was supported by the second stage Brain Korea 21 Project.

[1]Department of Statistics, Seoul National University, Seoul 151-747, Korea
[2]Corresponding author. Department of Statistics, Seoul National University, Seoul 151-747, Korea (e-mail: parksh@plaza.snu.ac.kr)

use algorithm to isolate outliers and the indirect methods use the results from robust regression estimates. But the multiple outlier detection methods for logistic regression have not been studied from the point of direct procedure yet. In this paper we consider the direct methods for logistic regression by extending the influence matrix algorithm of Peña and Yohai (1995). Peña and Yohai (1995) suggested a new method to identify influential subsets by looking at the eigenvalues of an influence matrix. The matrix is defined as the uncentred covariance of a set of vectors which represents the effect on the fit of the deletion of each data point. The matrix is normalized to have the univariate Cook's distance (Cook, 1979) on the diagonal. We define the influence matrix in logistic regression by using Cook's distance in logistic regression, and identify the influential subset of observations. And then we test outliers by using the mean shift model.

2. Review of Detection of Influential Subsets

In this chapter, we will review the meaning of eigenvector of influence matrix suggested by Peña and Yohai (1995) briefly. Peña and Yohai (1995) suggest a new method to identify influential subsets in linear regression problems. The procedure uses the eigenstructure of the influence matrix which is defined as the matrix of uncentred covariances of the effect on the whole data set of deleting each observation, and normalized to include the univariate Cook statistics on the diagonal. It is shown that the eigenstructure of the influence matrix is useful to identify influential subsets and a procedure for detecting influential sets is proposed. We consider a linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\mathbf{y}$ is the response vector of dimension n, $\mathbf{X}$ is the $n \times p$ matrix of regressor variables with intercept, $\boldsymbol{\epsilon}$ is the column vector of n random errors with identical distribution of $N(0, \sigma^2)$. Let $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ be the least squares estimate (LSE) and let $\hat{\boldsymbol{\beta}}_{(i)}$ be the LSE when the i^{th} data point is deleted. Then, the vector $\mathbf{t}_i = \hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)}$ summarizes the effect on the fit of deleting the observation i and is given by $\mathbf{t}_i = \{e_i/(1 - h_{ii})\}\mathbf{h}_i$, where $\hat{\mathbf{y}}_{(i)}$ to be the new fitted value using $\hat{\boldsymbol{\beta}}_{(i)}$ and $\mathbf{h}_i$ is the i^{th} column of the $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Cook's distance is given by $\mathbf{t}_i'\mathbf{t}_i/p\sigma^2$. Let us call $\mathbf{T}$ the $n \times n$ matrix $\mathbf{T} = (\mathbf{t_1}, \ldots, \mathbf{t_n})$ whose columns are the vectors $\mathbf{t}_i$. Then we can define the $n \times n$ influence matrix $\mathbf{M}$ as

$$\mathbf{M} = \frac{1}{ps^2}\mathbf{T}'\mathbf{T}, \tag{2.1}$$

where $s^2 = \sum_{i=1}^n e_i^2/(n-p)$.
Let r_{ij} be the uncentred correlation coefficient between $\mathbf{t}_i$ and $\mathbf{t}_j$. Then,

$$r_{ij} = \frac{m_{ij}}{m_{ii}^{1/2} m_{jj}^{1/2}}, \tag{2.2}$$

where m_{ij} is the (i,j) element of $\mathbf{M}$.

Suppose that there are k groups of influential observations $\mathbf{I}_1, \ldots, \mathbf{I}_k$. Then

(a) if $i, j \in \mathbf{I}_h$, then $|r_{ij}| = 1$
(this means that the effects on the least squares fit produced by the deletion of two points in the same set $\mathbf{I}_h$ have correlation 1 or -1),

(b) if $i \in \mathbf{I}_{h1}$ and $j \in \mathbf{I}_{h2}$ with $h1 \neq h2$, then $r_{ij} = 0$
(this means that the effects produced on the least squares fit by the observations i and j belonging to different sets are uncorrelated) and

(c) if i does not belong to any $\mathbf{I}_h$, then $m_{ij} = 0$ for all j
(this means that data points outside these groups have no influence on the fit).

Now, according to (a) we can split each set $\mathbf{I}_h$ into $\mathbf{I}_h^1$ and $\mathbf{I}_h^2$ such that

(i) if $i, j \in \mathbf{I}_h^q$, then $r_{ij} = 1$ and

(ii) if $i \in \mathbf{I}_h^1$ and $j \in \mathbf{I}_h^2$, then $r_{ij} = -1$.

Let $\mathbf{v}_1 = (v_{11}, \ldots, v_{1n})', \ldots, \mathbf{v}_k = (v_{k1}, \ldots, v_{kn})'$ be defined as $v_{hj} = +m_{jj}^{1/2}$ if $j \in \mathbf{I}_h^1$, $v_{hj} = -m_{jj}^{1/2}$ if $j \in \mathbf{I}_h^2$, $v_{hj} = +m_{jj}^{1/2}$ if $j \notin \mathbf{I}_h$. Then, if (a)–(c) hold, by the equation (2.2),

$$m_{ij} = r_{ij} m_{ii}^{1/2} m_{jj}^{1/2} = \begin{cases} 1 v_{hi} v_{hj}, & \text{if } i, j \in \mathbf{I}_h^1, \\ -1 v_{hi}(-v_{hj}), & \text{if } i \in \mathbf{I}_h^1 \text{ and } j \in \mathbf{I}_h^2, \\ 0 v_{hi} 0, & \text{if } i \in \mathbf{I}_h \text{ and } j \notin \mathbf{I}_h. \end{cases}$$

That is,

$$m_{ij} = \begin{cases} v_{hi} v_{hj}, & \text{if } i, j \in \mathbf{I}_h, \\ 0, & \text{if } i \text{ or } j \notin \mathbf{I}_h. \end{cases}$$

Therefore, the matrix $\mathbf{M}$ is

$$\mathbf{M} = \sum_{i=1}^{k} \mathbf{v_i v_i'}$$

and since the $\mathbf{v}_i$ are orthogonal, and the eigenvectors of $\mathbf{M}$ are $\mathbf{v_1}, \ldots, \mathbf{v_k}$, the corresponding eigenvalues $\lambda_1, \ldots, \lambda_k$ are given by

$$\lambda_h = \sum_{i \in \mathbf{I}_h} m_{ii}.$$

For real data sets, the conditions (a)–(c) do not hold exactly. However, the masking effects typically due to the presence in the sample of blocks of influential observations having similar or opposite effects. These blocks are likely to produce the matrix $\mathbf{M}$ with a structure close to that described by (a)–(c).

This suggests the following procedure to identify influential sets:

(a) find the eigenvectors corresponding to the p non-null eigenvalues of the influence matrix $\mathbf{M}$, and

(b) consider the eigenvectors corresponding to the large eigenvalues, and define the sets $\mathbf{I}_j^1$ and $\mathbf{I}_j^2$ by those components with large positive and negative weights, respectively.

3. MULTIPLE OUTLIER DETECTION IN LOGISTIC REGRESSION

3.1. Cook's distance in logistic regression

In this subsection, we show background for the logistic regression model and define Cook's distance in the logistic regression.

Given a sample of n independent binomial responses $y_i \sim B(n_i, p_i)$, the loglikelihood funcion for the sample is the sum of individual loglikelihood contributions;

$$l(\boldsymbol{\theta}; \mathbf{y}) = \sum_{i=1}^{n} l(\theta_i; y_i) = \sum_{i=1}^{n} \{y_i\theta_i - a(\theta_i) + b(y_i)\}.$$

The likelihood function $l(\boldsymbol{\theta}; \mathbf{y})$ is over-specified in $\boldsymbol{\theta}$, since there are as many parameters as observations. Given a set of m explanatory variables $X_1, X_2, \ldots, X_m$, the logistic regression model utilizes the relationship

$$\boldsymbol{\theta} = \text{logit}(\mathbf{p}) = \mathbf{X}\boldsymbol{\beta},$$

as the description of the systematic component of the response $\mathbf{y}$, where $\mathbf{p}=(p_1,\ldots,p_n)$. In terms of the m-dimensional parameter $\boldsymbol{\beta}$, we have the loglikelihood function

$$l(\mathbf{X}\boldsymbol{\beta};\mathbf{y})=\sum_{i=1}^{n} l(\mathbf{x}_i'\boldsymbol{\beta};y_i)=\sum_{i=1}^{n}\{y_i\mathbf{x}_i'\boldsymbol{\beta}-a(\mathbf{x}_i'\boldsymbol{\beta})+b(y_i)\}. \tag{3.1}$$

MLE maximizes the equation (3.1) and is a solution to $(\partial/\partial\hat{\boldsymbol{\beta}})l(\mathbf{X}\hat{\boldsymbol{\beta}};\mathbf{y})=0$. In particular, $\hat{\boldsymbol{\beta}}$ satisfies the system of equations

$$\sum_{i=1}^{n} x_{ij}\left(y_i-\dot{a}(\mathbf{x}_i'\hat{\boldsymbol{\beta}})\right)=0, \qquad j=1,\ldots,m.$$

Writing $\mathbf{s}=\mathbf{y}-\dot{a}(\mathbf{X}\hat{\boldsymbol{\beta}})=(\mathbf{y}-\mathbf{n}\hat{\mathbf{p}})$, the matrix formulation of the likelihood equations is

$$\mathbf{X}'\mathbf{s}=\mathbf{X}'(\mathbf{y}-\hat{\mathbf{y}})=0.$$

These equations, although very similar to their normal theory counterparts, are nonlinear in $\hat{\boldsymbol{\beta}}$, and iterative methods are required to solve them. Typically, when second derivatives are easy to compute (in the present case $-(\partial/\partial\hat{\boldsymbol{\beta}})\mathbf{X}'\mathbf{s}=\mathbf{X}'\mathbf{V}\mathbf{X}$ with $\mathbf{V}=\text{diag}\{\ddot{a}(\mathbf{x}_i\hat{\boldsymbol{\beta}})\}$), the Newton-Raphson method is employed. This leads to the iterative scheme

$$\boldsymbol{\beta}^{t+1}=\boldsymbol{\beta}^{t}+(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{X}'\mathbf{s}, \qquad t=0,1,\ldots,*$$

where both $\mathbf{V}$ and $\mathbf{s}$ are evaluated at $\boldsymbol{\beta}^t$. At convergence ($t=*$), we take $\hat{\boldsymbol{\beta}}=\boldsymbol{\beta}^*$, and denote the fitted values $n_i\hat{p}_i$ by $\hat{y}_i$. The estimated variance of y_i is $\nu_{ii}=n_i\hat{p}_i(1-\hat{p}_i)$.

A most useful way to view the iterative process outlined above is the method of iteratively reweighted least-squares (IRLS). This is obtained by employing the pseudo-observation vector $\mathbf{z}^t=\mathbf{X}\boldsymbol{\beta}^t+\mathbf{V}^{-1}\mathbf{s}$, upon which the above equation becomes

$$\boldsymbol{\beta}^{t+1}=(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{z}^t.$$

At convergence, we have $\mathbf{z}=\mathbf{X}\hat{\boldsymbol{\beta}}+\mathbf{V}^{-1}\mathbf{s}$. Thus we may write the MLE of $\boldsymbol{\beta}$ as $\hat{\boldsymbol{\beta}}=(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{z}$. The Cook's distance for GLM (Pregibon, 1981; Williams, 1987) is

$$C_i=p^{-1}h_{ii}(1-h_{ii})^{-1}r_{pi}^2, \tag{3.2}$$

where h_{ii} is i^{th} diagonal element of $\mathbf{H}=\mathbf{V}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{1/2}$ and r_{pi}^2 is Pearson residuals.

3.2. Influence matrix in logistic regression

We defined the influence matrix in a linear regression at Section 2. The influence matrix is an expansion of Cook's distance and is uncentred covariance matrix of the $\mathbf{t}_i$. By using Cook's distance for GLM, we can define $\mathbf{t}_i$ as

$$\mathbf{t}_i = \frac{r_{pi}}{\sqrt{1-h_{ii}}}\mathbf{h}_i,$$

where $\mathbf{h}_i$ is i^{th} column of the $\mathbf{H} = \mathbf{V}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{1/2}$.

One of the most important types of masking situations occurs when several observations have similar effects. We shall also say that two observations of the i^{th} and j^{th} have similar effects when $\mathbf{t}_i \approx \lambda\mathbf{t}_j$ in logistic regression. To detect possible sets of influential observations having similar or opposite effects on the fit, it seems plausible to look at the influence matrix. Let us define $\mathbf{T}$ as the $n \times n$ matrix $\mathbf{T} = (\mathbf{t}_1, \ldots, \mathbf{t}_n)$. Then we define the $n \times n$ influence matrix $\mathbf{M}$ as

$$\mathbf{M} = \frac{1}{p}\mathbf{T}'\mathbf{T}.$$

3.3. Modified multiple outlier detection algorithm

3.3.1. Identifying sets of outlier candidates: Step 1. A set of candidate outliers is obtained by analysing the eigenvectors corresponding to the non-null eigenvalues of the influence matrix $\mathbf{M}$, and by searching in each eigenvector for a set of co-ordinates with relatively large weights and the same sign. First of all, we explain the Peña and Yohai (1995) algorithm in linear regression. Then, we suggest our algorithm for logistic regression by extending it. Their algorithm can be summarized as follows.

(1) Order the co-ordinates of the eigenvector $\mathbf{v}_l$, obtaining $v_{i_{(1)}} \leq \cdots \leq v_{i_{(n)}}$, and let us call $i_{(1)}, \ldots, i_{(n)}$ the indices of the ordered co-ordinates of the eigenvector.

(2) Compute the ratios $a_j = v_{i_{(j)}}/v_{i_{(j-1)}}$ for $j = n, \ldots, n-c_1$ and $b_j = v_{i_{(j)}}/v_{i_{(j+1)}}$ for $j = 1, \ldots, c_2$. The constants c_1 and c_2 are smaller than $n/2$ and will be discussed below.

(3) Look for the first j_0 such that $|a_j| > k$ and i_0 such that $|b_j| > k$.

(4) If $i_0 > 1$ and/or $j_0 > 1$, consider the sets $\mathbf{I}_0 = \{i_{(1)}, i_{(2)}, \ldots, i_{(i_0)}\}$ and $\mathbf{J}_0 = \{i_{(n)}, i_{(n-1)}, \ldots, i_{(j_0)}\}$ as outlier candidates.

The choice of c_1 and c_2 is related to the desired breakdown point of the procedure that will be smaller than $\min(c_1/n, c_2/n)$. In practice, Peña and Yohai (1995) suggest c_1 and c_2 be close to $n/4$. The power of the procedure for the detection of outliers depends on the choice of k. They suggest $k = 2.5$ through their experience with real and simulated data. This method, however, is conservative and $k = 2.5$ is not appropriate to logistic regression. So we suggest the first $j_0 = \arg\max_j a_j$ for $j = n-3, n-4, \ldots, n-c_1$ and $i_0 = \arg\max_j b_j$ for $j = 4, 5, \ldots, c_2$. Our algorithm can be summarized as follows.

(1*) and (2*) The same as (1) and (2) in the Peña and Yohai (1995) algorithm.

(3*) Look for the first $j_0 = \arg\max_j a_j$ for $j = n-3, n-4, \ldots, n-c_1$ and $i_0 = \arg\max_j b_j$ for $j = 4, 5, \ldots, c_2$.

(4*) Consider the sets $\mathbf{J}_0 = \{i_{(n)}, i_{(n-1)}, i_{(n-2)}, i_{(n-3)}, \ldots, i_{(j_0)}\}$ and $\mathbf{I}_0 = \{i_{(1)}, i_{(2)}, i_{(3)}, i_{(4)}, \ldots, i_{(i_0)}\}$ as outlier candidates.

3.3.2. Checking for outliers: Step 2. We use the mean shift model for checking outliers. The mean shift model provides a simple method of finding the effect of multiple deletion,

$$\theta = \text{logit}(\mathbf{p}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{D}\boldsymbol{\phi}, \tag{3.3}$$

where $\mathbf{D}$ is the matrix that has a single one in each of its columns, which are otherwise zero, and m rows with one nonzero element. These m entries specify the observations that are to have individual parameters or, equivalently, are to be deleted.

1. $\mathbf{J}_0$ (outlier candidates) are tested by using the mean shift model (3.3).
   $\Rightarrow \mathbf{S}_1$: new outlier candidates.
   $\mathbf{I}_0$ (outlier candidates) are tested by using the mean shift model (3.3).
   $\Rightarrow \mathbf{S}_2$: new outlier candidates.

2. Let n_1 and n_2 be the sizes of the new outlier candidates, $\mathbf{S}_1$ and $\mathbf{S}_2$.

   (a) If $n_1 \geq n_2$, then declare all observations in $\mathbf{S}_1$ as outliers, substitute the observations except $\mathbf{S}_2$ for full data, and go to Step 1.
   $\Rightarrow \mathbf{S}_1$: outliers, $\mathbf{I}_0^1$ and $\mathbf{J}_0^1$: outlier candidates.

(b) $\mathbf{I}_0^1$ and $\mathbf{J}_0^1$ are tested by using the mean shift model (3.3).
If the t-statistics for outlier candidates are significant, then declare all observations satisfying significance test as outliers and stop.

3. (a) If $n_1 < n_2$, then declare all observations in $\mathbf{S}_2$ as outliers, substitute the observations except $\mathbf{S}_2$ for full data, and go to Step 1.
$\Rightarrow \mathbf{S}_2$: outliers, and $\mathbf{I}_0^2$ and $\mathbf{J}_0^2$: outlier candidates.
(b) $\mathbf{I}_0^2$ and $\mathbf{J}_0^2$ are tested by using the mean shift model (3.3).
If the t-statistics for outlier candidates are significant, then declare all observations satisfying significance test as outliers and stop.

4. If n_1 and n_2 become 0, $\mathbf{J}_0$ and $\mathbf{I}_0$ are tested by using the mean shift model at once. If the t-statistics for outlier candidates are significant, then declare all observations satisfying significance test as outliers and stop.

Peña and Yohai (1995) suggest that $\mathbf{J}_0$ and $\mathbf{I}_0$ can be tested by the mean shift model at once. In logistic regression, their method brings on swamping effect or decreasing accuracy. In particular, if multiple outliers exist with attaching weight to one side ($y = 0$ or $y = 1$), their method brings on a worse result. So we suggest that $\mathbf{J}_0$ and $\mathbf{I}_0$ each are tested by the mean shift model. Then, we define that $\mathbf{S}_1$ and $\mathbf{S}_2$ are new outlier candidates. If multiple outliers exist with attaching weight to one side $y = 0$, $\mathbf{S}_2$ is caused by swamping. In logistic regression, if we exclude $\mathbf{S}_1$ from data and repeat Step 1 and Step 2, swamping effect is reduced and accuracy is increased.

3.4. Examples of multiple outlier detection

3.4.1. Example 1. The first example is designed to show the interpretation of the eigenvectors of the influential matrix in simple masking scheme. We use here the artificial data generated by Ryan (1996). The model contains 50 data points in two dimensions (one response and one explanatory variables). We change from the response variable $y = 1$ to the response variable $y = 0$ in the observations $2, 6, 11, 23, 33$ and 46 to give multiple outliers. Table 3.1 presents the eigenvector corresponding to the largest eigenvalue of the influence matrix ($\lambda_1 = 3.00325 \times 10^{-4}$). In this case, outliers $(2, 6, 11, 23, 33, 46)$ have the largest negative weight. The smallest eigenvector has $b_1 = 1.108788$, which corresponds to $j_{(1)} = 2$, and $b_6 = 1.503226$ with $j_{(1)} = 46$. This $b_6 = 1.503226$ with $j_{(1)} = 46$

TABLE 3.1 *Eigenvectors corresponding to the largest eigenvalue for Example 1*

observation	2	11	33	6	23	46	29	37
eigenvector	−0.269	−0.243	−0.243	−0.219	−0.219	−0.197	−0.131	−0.131
observation	49	7	14	30	8	38	13	28
eigenvector	−0.131	−0.1154	−0.119	−0.119	−0.107	−0.107	−0.0962	−0.0962
observation	47	12	15	22	34	25	43	40
eigenvector	−0.096	−0.087	−0.087	−0.087	−0.087	−0.078	0.078	−0.070
observation	21	41	39	19	20	32	3	16
eigenvector	−0.063	−0.063	−0.057	−0.051	−0.051	0.083	0.093	0.093
observation	35	1	45	27	44	4	9	26
eigenvector	0.093	0.093	0.103	0.114	0.114	0.127	0.127	0.127
observation	36	24	31	48	10	17	18	5
eigenvector	0.127	0.155	0.155	0.155	0.190	0.190	0.190	0.211
observation	50	42						
eigenvector	0.211	0.234						

TABLE 3.2 *t-statistics for Example 1*

observation	*t-statistics for observation*
2	−2.0251
6	−1.9049
11	−1.9796
23	−1.9049
33	−1.9796
46	−1.7774

is $\arg\max_j b_j$ for $j = 4, 5, \ldots, c_2$. Therefore, it has a clear cut-off point at the set $\mathbf{I}_0 = \{2, 11, 33, 6, 23, 46\}$.

Table 3.2 presents the t-statistics for these points when they are removed from the least squares fit. This shows that the t-statistic identifies clearly outliers. In summary, the components of the eigenvector corresponding to the largest eigenvalues show the relevant structure of the data set, and the relevant set is automatically selected by the procedure suggested in Section 3.3.

3.4.2. Example 2. In this subsection, we will test the performance of the multiple outlier detection procedure which we suggested and compare the result of our procedure with that of robust estimation. In multiple outlier detection for logistic regression, there is not classic data. So we design an example to have multiple outliers with masking and swamping. We consider the true model following $p = 2$

TABLE 3.3 $\hat{\beta}$ *and relative efficiency for Example 2*

method	β_0	β_1	β_2	*Relative efficiency*
True model	12.138	−5.342	7.702	
Classical ML method	6.974	−3.443	3.645	
Our procedure by using influence matrix	6.872	−3.390	3.587	3.877
Cantoni and Ronchetti's method	0.983	−0.662	0.648	1.008
Croux and Haesbroeck's method	1.002	−0.669	0.654	1.017

independent variables and $n = 50$. There are multiple outliers in the response variable $y = 0$ and $y = 1$, but the number of outliers in $y = 0$ is more than that of outliers in $y = 1$. The true model is defined as

$$\log\left(\frac{\pi}{1-\pi}\right) = 6.973 - 3.442\mathrm{x}_1 + 3.645\mathrm{x}_2.$$

Then, we change from the response variable $y = 0$ to the response variable $y = 1$ in the observations $10, 15, 21, 26$ and 35 and from the response variable $y = 1$ to the response variable $y = 0$ in the observations $7, 16$ and 32 to give multiple outliers. The automatic procedure suggested in this paper detects multiple outliers correctly.

For comparison of various methods, we compute the finite sample relative efficiency (RE), defined as

$$\mathrm{RE} = \frac{\text{MSE of ML estimator}}{\text{MSE of robust estimator}}. \tag{3.4}$$

We include the robust estimators from Cantoni and Ronchetti (2001) and Croux and Haesbroeck (2003). Cantoni and Ronchetti (2001) proposed a robust approach to inference based on robust deviances that are natural generalization on quasi-likelihood functions. Croux and Haesbroeck (2003) complemented Bianco and Yohai (1996) who proposed a highly robust procedure for estimation of the logistic regression model.

In Figure 3.1, "True model GLM fitted" represents fit by logistic model with non-contaminated data. "The Model fitted with outliers" represents the fit by four methods for estimating $\boldsymbol{\beta}$ with contaminated data. The four methods are GLM, our procedure, the robust estimate by Cantoni and Ronchetti (2001) and

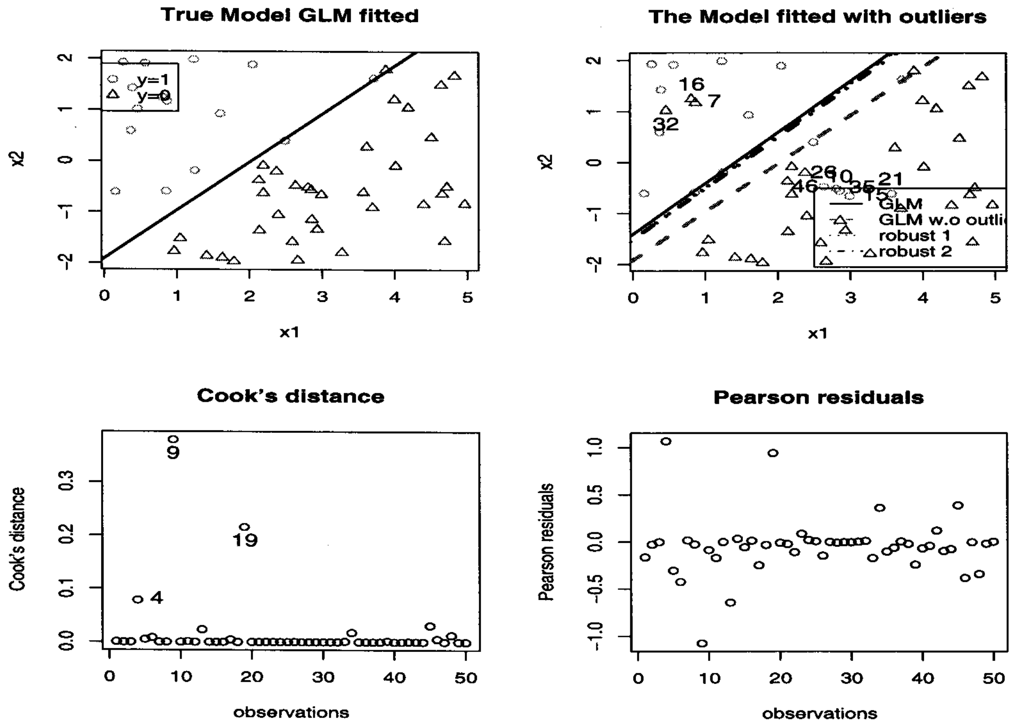

FIGURE 3.1 *Plots for Example 2.*

the robust estimate by Croux and Haesbroeck (2003). In GLM without outliers, we detected multiple outliers by using our procedure and we estimated $\hat{\beta}$ from contaminated data except detected multiple outliers. "Cook's distance" and "Pearson Residuals" show that the contaminated data set has masking effect. Table 3.3 presents β and $\hat{\beta}$ by using the four methods and relative efficiency. As shown in Figure 3.1, the fitting by using our procedure is the most similar one to the true model. In Table 3.3, our procedure maintains high efficiency for the contaminated data.

4. CONCLUDING REMARKS

In this paper, we suggested the necessity of multiple outlier detection and multiple outlier detection algorithm. Then, we compared our algorithm with robust statistical models by using relative efficiency.

Our algorithm was based on influence matrix. We defined influence matrix in logistic regression as extending Cook's distance in logistic regression (Pregibon, 1981). This means that, the influence matrix is expansion of Cook's distance and uncentred covariance matrix of the vector $\mathbf{t}_i = \hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)}$. The eigenvalue of influence matrix explains influence of the arbitrary group and the eigenvector corresponding to the eigenvalue explains how far observations are included in that group. We obtained the set of candidate outliers by using relative value of these eigenvalues and by testing mean shift model.

To show accuracy of multiple outlier detection algorithm in Section 3, we simulated artificial data including multiple outliers with masking and swamping. In this result, our procedure detected multiple outliers correctly. Also we compared our procedure with the robust modelling for inference as Cantoni and Ronchetti (2001) and Croux and Haesbroeck (2003). We showed that our procedure maintains high efficiency for the contaminated data. In some cases, we noted that the robust modelling for inference is worse than the classical ML method with contaminated data.

In GLM, multiple outlier detection leaves much to be studied further. Also, our algorithm is needed to complement many points. First, our algorithm has to be studied for a real data set. Secondly, we have to show how to change the accuracy of multiple outlier detection, as the independent variable number p increases. Thirdly, our study is limited to multiple outlier detection in logistic regression. So we have to extend our study to multiple outlier detection in GLM.

REFERENCES

BECKMAN, R. J. AND COOK, R. D. (1983). "Outlier . . . s", *Technometrics*, **25**, 119–163.

BIANCO, A. M. AND YOHAI, V. J. (1996). "Robust estimation in the logistic regression model", In *Robust Statistics, Data Analysis, and Computer Intensive Methods; Lecture Notes in Statistics 109* (Rieder, H. ed.), 17–34, Springer-Verlag, New York.

CANTONI, E. AND RONCHETTI, E. (2001). "Robust inference for generalized linear models", *Journal of the American Statistical Association*, **96**, 1022–1030.

CHATTERJEE, S. AND HADI, A. S. (1986). "Influential observations, high leverage points, and outliers in linear regression", *Statistical Science*, **1**, 379–416.

COOK, R. D. (1979). "Influential observations in linear regression", *Journal of the American Statistical Association*, **74**, 169–174.

CROUX, C. AND HAESBROECK, G. (2003). "Implementing the Bianco and Yohai estimator for logistic regression", *Computational Statistics & Data Analysis*, **44**, 273–295.

PREGIBON, D. (1981). "Logistic regression diagnostics", *The Annals of Statistics*, **9**, 705–724.

PEÑA, D. AND YOHAI, V. J. (1995). "The detection of influential subsets in linear regression by using an influence matrix", *Journal of the Royal Statistical Society*, Ser. B, **57**, 145–156.

RYAN, T. P. (1996). *Modern Regression Methods,* John Wiley & Sons, New York.

WILLIAMS, D. A. (1987). "Generalized linear model diagnostics using the deviance and single case deletions", *Journal of the Royal Statistical Society,* Ser. C, **36**, 181–191.