行政院國家科學委員會專題研究計畫　成果報告

邏輯斯迴歸之穩健診斷分析

計畫類別：個別型計畫
計畫編號：NSC92-2118-M-004-005-
執行期間：92年08月01日至93年08月31日
執行單位：國立政治大學統計學系

計畫主持人：鄭宗記

報告類型：精簡報告
報告附件：出席國際會議研究心得報告及發表論文
處理方式：本計畫可公開查詢

中　華　民　國 93年11月24日

Robust Diagnostics for the Logistic Regression Model

Tsung-Chi Cheng[*]

Abstract

Current imputation methods, such as the EM types, are easily affected by outliers. They may either impute extra outliers if the existing outliers are ignored, or may move (potential) outliers to the center of the data when only part of them are observed. Atkinson and Cheng (2000) first employ the high breakdown estimator for the detection of outliers in the presence of missing values for the linear regression model. Recently, Atkinson and Riani (2001) adapt the forward search algorithm for the least median of squares (LMS) estimator to deal with outlier problem in the binary data. In this article, we extend their approach to identify multiple outliers for the logistic regression model when part of data are missing. The proposed forward algorithm starts with using the imputation method based on a small subset of the data to fill in those missing observations in the whole data set. Then the subset augments until all observations are included in the analysis. The algorithm monitors the statistics of interest during the progress of the forward search. The LMS criterion is used to judge the performance of each search. Real data are used to illustrate the resulting procedure.

Keywords: EM algorithm; forward search algorithm; high breakdown point; incomplete data; logistic regression model; robust diagnostics.

1 Introduction

Multiple outliers may strongly affect the generalized linear model fitted to data. However, those observations may be difficult to identify. The deletion diagnostics described in Cook and Weisberg (1982) and Atkinson (1985) for linear regression model may fail due to the masking and swamping effects. For the generalized linear models deletion diagnostics was first discussed by Pregibon (1981) and then summarized by McCullagh and Nelder (1989, Chapter 12). Haslett (1999) uses multiple-deletion diagnostics to the problem of multiple outliers, which may likewise fail either owing masking or computational requirement and

[*]Department of Statistics, National Chengchi University, 64 Chih-Nan Road, Section 2, Taipei 11623, Taiwan. E-mail: chengt@nccu.edu.tw.

interpretability if there are too many outliers. When there are several outliers, they may not be obviously revealled from the classical residual plots. Hence, it essentially requires high breakdown estimators for the identification of multiple outliers.

During the last few years, there have been attempts to explore and develop robust methods for the generalized linear models. Stefanski *et al.* (1986) and Künsch *et al.* (1989) propose bounded influence estimators which depend on an auxiliary centering constant and nuisance matrix. Christmann (1994) suggests the least median weighted squares (LMWS) estimator in logistic regression model for large strata data, which first transforms the binomial data by multiplying each observation with a proper weight, followed by the application of the least median of squares (LMS) algorithm to the transform data. Recently, Kordzakhia *et al.* (2001) present robust M-estimates based on influence function approach for the multiple logistic regression model. Christmann and Rousseeuw (2001) and Rousseeuw and Christmann (2003) discuss robustness and outliers in the logistic regression model. Atkinson and Riani (2001) adapt the LMS of the forward search algorithm (Atkinson 1994) to deal with the problem of the detection of multiple outliers in binomial data. Müller and Neykov (2003) apply the trimmed likelihood estimator for the generalized linear model, which is based on trimming the likelihood function rather than directly trimming the data.

On the other hand, data are often collected in which some observations are incomplete. With missing data, a simple and thus popular approach is to do an analysis throwing out all subjects with any missing data. Such a "complete case analysis" is often a wasteful of information, because the omitted units carry information with respect to the relation between the observed covariates and the outcome variable. The complete-case estimates could be inefficient compared to a likelihood method that uses the incomplete data and would be biased if the complete cases are not a random subset of the full dataset. Hence, the detection of outliers only considered in complete cases may be unreliable.

Maximum likelihood procedures for analysis mixed continuous and categorical data with missing values are presented by Little and Schluchter (1985). When the missing covariates are categorical, a useful technique for obtaining parameter estimates is the EM algorithm by the method of weights proposed by Ibrahim (1990). Ibrahim, Chen and Lipsitz (1999) extend this method to continuous or mixed categorical and continuous covariates, and for arbitrary parametric regression models, by adapting a Monte Carlo version of the EM algorithm discussed by Wei and Tanner (1990).

In this article, we propose a robust procedure for the detection of multiple outliers in generalized linear model with incomplete data. The proposed algorithm is basically a combination of two procedures, the forward search algorithm and the EM algorithm. The algorithm starts by firstly choosing a subset of "good" points from the completely observed data, and then fill in the missing data by using the EM algorithm based on the selected data. Once all missing data are imputed, we can obtain the residuals for the whole data. Based on the ordering residuals, observations in the subset are augmented. The procedure ends until when all data points are included.

2 Logistic regression model

Let there be n binomial observations of the form $y_i/m_i, i = 1, 2, \ldots, n$, where $E(y_i) = m_i\pi_i$ and π_i is the success probability corresponding to the ith observation. The binomial distribution for a fixed number of trials is determine by the probability π of success. Both the mean and the variance depend only on π_i and the known number m_i of trials.

For each y_i we know the number of trials m_i, and in addition there is an associated vector of $p+1$ predictors $\boldsymbol{x_i}$. Assume that the probability of success depends on $\boldsymbol{x_i}$, then the probability function of y can be written as

$$\begin{aligned} \pi_i &= f(y; \boldsymbol{\beta}) \\ &= \frac{\exp(\boldsymbol{x_i^T}\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i^T\boldsymbol{\beta})} \ . \end{aligned} \tag{1}$$

Note that $0 \leq \pi_i \leq 1$ for all values of $\boldsymbol{\beta}$ and $\boldsymbol{x}_i$. The log odds ratio,

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_i^T\boldsymbol{\beta} \ ,$$

is linear in the parameters $\boldsymbol{\beta}$. The logistic regression may be viewed as a nonlinear model with heterosceddastic errors. That is,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim bin(m, \pi), \tag{2}$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$, $E(\epsilon_i) = m_i\pi_i$ and $\mathrm{Var}(\epsilon_i) = m_i\pi_i(1-\pi_i)$.

In order to fit a linear logistic model to a given set of data, the $p+1$ unknown parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ have first to be estimated. These parameters are readily estimated using the method of maximum likelihood. The likelihood function is given by

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\begin{array}{c} m_i \\ y_i \end{array} \right) \pi_i^{y_i} (1-\pi_i)^{m_i - y_i} \ .$$

This likelihood depends on the unknown success probabilities π_i, which in turn depend on the $\boldsymbol{\beta}s$ through equation (1), so the likelihood function can be regarded as a function of $\boldsymbol{\beta}$.

The logarithm of the likelihood function is

$$\begin{aligned} \log L(\boldsymbol{\beta}) &= \sum_i \left\{ \log \left(\begin{array}{c} m_i \\ y_i \end{array} \right) + y_i \log \pi_i + (m_i - y_i) \log(1-\pi_i) \right\} \\ &= \sum_i \left\{ \log \left(\begin{array}{c} m_i \\ y_i \end{array} \right) + y_i \log (\frac{\pi_i}{1-\pi_i}) + m_i \log(1-\pi_i) \right\} \\ &= \sum_i \left\{ \log \left(\begin{array}{c} m_i \\ y_i \end{array} \right) + y_i \eta_i - m_i \log(1+e^{\eta_i}) \right\} \ , \end{aligned}$$

where $\eta_i = \sum_{j=0}^{p} \beta_j x_{ji}$ and $x_{0i} = 1$ for all $i = 1, \dots, n$. Taking derivatives of this log-likelihood function with respect to the $p+1$ unknown parameters and equating them to zero, we obtain a set of a $p+1$ non-linear equations. Then the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$ is obtained.

Once the estimate of $\boldsymbol{\beta}$ is obtained, the estimated value of the model is

$$\hat{\eta}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_p x_{pi} \ .$$

The fitted probabilities $\hat{\pi}_i$ can then be found using $\hat{\pi}_i = \exp(\hat{\eta}_i)/[1+\exp(\hat{\eta}_i)]$.

In linear regression, summary measures of the fit, as well as diagnostics for casewise effect on the fit, are functions of a residual. In logistic regression there are several possible ways to measure the difference between the observed and the fitted values, one of which is the deviance residual. The ith deviance residual is defined as

$$d_i = sign(y_i - \hat{y}_i) \sqrt{2 y_i \log \left(\frac{y_i}{\hat{y}_i} \right) + 2(m_i - y_i) \log \left(\frac{m_i - y_i}{m_i - \hat{y}_i} \right)} \ , \qquad (3)$$

where "*sign*" denotes the sign of $(y_j - \hat{y}_j)$. When the response is binary, the deviance residual becomes

$$d_i = -2 sign(y_i - \hat{p}_i) \sqrt{y_i \mathrm{log} \hat{p}_i + (1-y_i) \mathrm{log}(1-\hat{p}_i)}. \qquad (4)$$

The deviance residual provides information about how well the model fits each particular observation. The advantage of this statistics is that a single number is used to summarized considerable information.

2.1 Example: the recumbent cow survival data

This example is a study from 435 recumbent cows, collected at the Ruakura Animal Health Laboratory, New Zealand, during 1983-1984. Some cows become unable to support their own weight, just before or after calving, so they become recumbent. The goal of the study is to determine if any of the physical measurements and blood test are related to survival probabilities for the cows, and how survival probabilities varies with characteristics.

There are eight factors that effect the survival probabilities of the cows, which are defined in Table 1. The response variable is binary, which is coded 1 if survived, 0 if died or killed. None of the explanatory variables are fully observed, because some measurements are only available during the second year.

Table 1. Recumbent cow survival data: variable description.

Variables	*Number of Observed Cases*	*Description*
AST	429	Serum asparate amino transferase, IU/1 at 30 C
Calving	431	0 if condition first occured before calving, 1 if post-calving
CK	413	Serum creatine phosphokinase, IU/1 at 30 C
Daysrec	432	Days recumbent when measurements were taken, rounded down to the nearest day
Inflamat	136	Imflamation: 1 if present, 0 if absent
Myopathy	222	Muscle disorder: 1 if present, 0 if absent
PCV	175	Packed cell volume (hematocrit),%
Urea	266	Serum urea, mmo 1/1

If missing values are omitted to fit a logistic regression including the intercept term, indicator variables for *Calving*, and *Myopathy*, the variable *Daysrec*, and base-two logarithms of *AST* and *CK*, Table 2 shows the estimation result by 216 fully observed cases. The deviance for the fit is 211.268, and the Pearson's χ^2 statistics has the value 207.681 with 210 d.f's. Hence, the logistic regression model fits the observed data very well. Figure 1 shows the deviance residuals plot, in which only six observations have deviance residuals slightly larger than ±2. Cook and Weisberg (1999) conclude that there is no any outlier in complete data analysis .

Table 2. Recumbent cow survival data: fit by available cases.

Variables	*Coefficient*	*Std. Error*	*t value*
(Intercept)	1.23649835	1.6091425	0.7684207
Calving	-0.30331715	0.4068258	-0.7455701
Myopathy	-1.84036905	0.6080790	-3.0265296
Daysrec	-0.05459210	0.1119614	-0.4875977
AST	0.03881005	0.2630031	0.1475650
CK	-0.15257528	0.1440310	-1.0593224

3 The forward search algorithm in logistic regression

Atkinson and Riani (2000) apply the forward search algorithm to obtain the robust diagnostic result for the logistic regression model. The procedure is summarized as follows. If the model contains $p+1$ parameters, the algorithm starts with the selection of a subset of $p+1$ units. Either by sampling 1,000 subsets or by exhaustively evaluating all subsets. Let S_j^{p+1} be a subset of size $p+1$ that is selected at step j of the forward search. Recalled the deviance defined in (3), if $d_{i,s_j^{p+1}}$ is the deviance residual for unit i given that observations in S_j^{p+1} are used in fitting the model, the initial subset is such that

$$d^2_{[med],s_*^{p+1}} = \min_j(d^2_{[med],s_j^{p+1}}), \tag{5}$$

where $d^2_{[l],S_j^{p+1}}$ is the lth ordered squared deviance residuals among $d^2_{i,S_j^p}, i = 1, ...n$ and $l = p + \lfloor(n-p)/2\rfloor$. Observations in this subset are intended to be outlier free. In this resampling algorithm the model is fitted to $m = p+1$ observations, when the remaining $n-(p+1)$ observations can be tested to see if any outliers are present.

Given a subset of size $m \geq p+1$, say $S_*^{(m)}$, the forward search moves to dimension $m+1$ by selecting the $m+1$ units with the smallest squared deviance residuals, the units being chosen by ordering all squared deviance residuals d^2_{i,S_*^m}, $i = 1, \ldots, n$. Repeat until all units are included in the subset.

Suppose at some step in the forward search the set of m observations used in fitting is $S_*^{(m)}$. Fitting to this subset is by least squares yielding the parameter estimates $\hat{\beta}_m^*$. We can calculate a set of n deviance residuals d_m^* from these parameter estimates. There will then be $n-m$ observations not used in fitting that may contain outliers. Monitor the changes in the forward plot as m goes from $p+1$ to n, quantities such as the deviance residuals, and the t statistic, Cook's distance and other diagnostic quantities, which can always be associated with the introduction of a particular group of observations. In most moves from m to $m+1$ just one new unit joins the subset. When the search includes one unit which belongs to a cluster of outliers, at the next step the remaining outliers in the cluster seems less outlying and so several may be included at once.

Through the joint examination of simple plots, which monitor the effect on the deviance residuals of the sequential inclusion of the units, we can obtain great insight into the structure of the data. The method is not sensitive to the selection of an initial subset. Details of forward search algorithm will be state in the following.

3.1 *Step 1: Choice of the initial subset*

Let $U = (X, y)$ be n observations. The fitted model takes the form in (2). If the model contains $p+1$ parameters, forward search algorithm starts with selection of a subset of $p+1$ units. Observations in this subset are intended to be outlier free. The choice of initial subset can be performed by exhaustive enumeration of all distinct $p+1$-tuple $S^{(p+1)}_{i_1,\ldots,i_{p+1}} \equiv u_{i_1}, \ldots, u_{i_{p+1}}$, where $u^T_{i_1}$ is the i_1th row of U, for $1 \leq i_1, \ldots, i_p \leq n$, and $i_j \neq i_{j'}$. Let $\iota' = [i_1, \ldots, i_{p+1}]$ and let $d_{i,S^{(p+1)}_\iota}$ be the least squares deviance residual for unit i given observations in $S_{(p+1)_\iota}$. The initial subset must satisfy

$$d^2_{[med],S^{(p+1)}_*} = \min_\iota(d^2_{[med],S^{(p+1)}_\iota}), \tag{6}$$

where $d^2_{[l],S^{(p+1)}_\iota}$ is the lth order squared residual among $d^2_{i,S^{(p+1)}_\iota}$, $i = 1, \ldots, n$, and med is the integer part of $(n+p+1)/2$. If both n and p are small, the choice can be performed by exhaustive enumeration of S^{P+1}_j, otherwise we evaluate the properties of some large number, usually 1,000 subsamples are drew. For large n or p, a large number of subsamples of p observations is taken.

3.2 *Step 2: Adding observations during the forward search*

The forward search selects the $m+1$ units with the smallest squared deviance residuals, the units being chosen by ordering all squared deviance residuals $d^2_{i,S^m_*}, i = 1, ..., n$. The residuals are computed by (5), where $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimated coefficients of the first m units included in the model.

In most moves, just one new unit joins the subset. When two or more units join $S^{(p+1)}_*$ as one or more leaves, indicating that the search includes one units belongs to a cluster of outliers, then in the next step the remaining potential outliers in the cluster seem less outlying and so several may be included at once. The introduction of influential observations is signaled by sharp changes in the curves that monitor parameter estimates, t test or an other statistic at each step. For monitoring the effect of individual observations on statistics and parameter estimates is helpful to be able to connect particular effects with particular observations. Forward search is accomplished by repeating this step.

3.3 *Step 3: Monitoring the search*

The forward search algorithm repeats Step 2 until all units are included in the subset. One of the most important plots monitors all deviance residuals at each move of the forward search. Large values of the deviance residuals among cases not in the subset indicate the presence of outliers.

4 Forward search for the binary response

Binary data are binomial data with each $m_i = 1$. The analysis is similar to that for binomial data with large m_i. However the binary responses lead to problems that do not arise for binomial data (Atkinson and Riani, 2001). In general the numbers of zero and one responses will not be equal. Suppose that zeroes predominate. Then a fit just to observations with a zero response will exactly fit enough of the data to give a value of zero for the residual to modify the search to avoid including only observations of one kind. This requires modification both of the initial subset and of the progress of the search.

Atkinson and Riani (2000) modified the search as follows: the initial subset is constrained to include at least one observation of each type. During the search, balance the ratio of zeros and ones in the subset so that it is as close as possible to the ratio in the complete set of n observations. Given a subset of size m we fit the model and then separately order the observation with zero response and those with unit responses. From these two lists of observations, we then take the m_0 smallest squared deviance residuals from the zero responses and m_1 smallest squared deviance with unit responses such that $m_0 + m_1 = m + 1$ and the ratio m_0/m_1 is close as possible to the ratio n_0/n_1 in the whole set of observations, where $n_0 + n_1 = n$.

4.1 Example: the recumbent cow survival data (continued)

Applying the algorithm of Atkinson and Riani (2000) to the recumbent cow data in previous section, many potential outliers are identified, that is, observations with deviance residuals larger than 2 in absolute value. LMS of the search occurs at the $129th$ step of the search. At this step, 77 observations has deviance residuals larger than 2.5. Once the potential outliers are included, the remaining potential outliers seems less outlying. Observations identified to be extreme by the classical method are included in these 77 observations.

The last 10 observations included in the search are listed in Table 3, together with the corresponding deviance. As the last column of the table shows, there is a steadily upward trend in the deviance as each observation is added. The large proportion of potential outliers may be caused by the deleting of large amounts of information, that is, omit from a logistic regression analysis any case that has a missing value for any variable. Omitting them from the model will tend to introduce bias in the estimation of the coefficients. There are variations of these estimated parameters during the forward search, and the variations indicate the inclusion of influential observations. Note that, we take the unbalance method in this search. This is because of the singularity problem occurred during the search.

Table 3. *(Recumbent cow survival data) : the last* 10 *stages of the forward search.*

m	*Obs. i*	*Deviance*
207	327	159.2
208	420	162.1
209	209	164.9
210	273	167.5
211	282	177.5
212	432	185.8
213	84	192.9
214	232	199.4
215	413	205.3
216	2	211.3

5 Missing values

In this section we introduce the most commonly used method for imputation of the incomplete data. Little and Schluchter (1985) discuss a model for missing data with mixed normal and categorical variables and provide relatively simple and computationally EM algorithm with missing data. Schafer (1997) tries to simulate posterior draws of multiple imputations of the missing data. To save computation time, we only consider single imputation method instead of multiple imputations. We also assume that the missing data are missing at random (MAR). MAR means that the probability that an observation is missing may depend on Y_{obs} but not on Y_{mis}, where Y_{obs} and Y_{mis} denote the fully-observed data and missing data, respectively (see Rubin (1976)).

Let R be an $n \times p$ matrix of indicator variables whose elements are zero or one depending on whether the corresponding element of Y are missing or observed. The distribution of R is related to Y. Hence we posit a probability model for R, $P(R|Y, \xi)$, which depends on Y as well as some unknown parameters ξ. The MAR assumption is that this distribution does not depend on Y_{mis},

$$P(R|Y_{obs}, Y_{mis}, \xi) = P(R|Y_{obs}, \xi).$$

Throughout, we assume that the missing responses follow a *monotone* pattern. The missingness pattern for a data matrix is said to be monotone if, whenever an element y_{ij} is missing, y_{ik} is also missing for all $k > j$ (Rubin, 1974; Little and Rubin, 1987). Monotone missing pattern will achieve stationary more rapidly than ordinary data argumentation

5.1 Notations

Let $W_1, W_2, \ldots, W_p$ denote a set of categorical variables and $Z_1, Z_2, \ldots, Z_q$ a set of continuous variables. If these variables are recorded for a sample of n units, the result is an $n \times (p+q)$ data matrix $(\boldsymbol{W}, \boldsymbol{Z})$, where $\boldsymbol{W}$ and $\boldsymbol{Z}$ represent the categorical and continuous parts, respectively.

The categorical data $\boldsymbol{W}$ may be summarized by a contingency table. Suppose that W_j takes possible values 1,2,...,d_j, so that each unit can be classified into a cell of a p-dimensional table with total number of cells equal to $D = \Pi_{j=1}^{p} d_j$. The frequencies in the complete-data contingency table will be

$$x = \{x_w : w \subset \boldsymbol{W}\},$$

where x_w is the number of units for which $(W_1, W_2, \ldots, W_p) = w$, and $\boldsymbol{W}$ is the set of all possible w.

The general location model, named by Olkin and Tate (1961), defined in terms of the marginal distribution of W and the conditional distribution of Z given W. The former is described by a multinomial distribution on the cell counts x,

$$x|\pi \sim M(n, \pi),$$

where $\pi = \{\pi_d : d = 1, 2, \ldots, D\}$ is an array of cell probabilities corresponding to x. Given W, the rows of z_1^T,z_2^T,...,z_n^T of Z are then modeled as conditionally multivariate normal. Let E_d denote a $1 \times D$ vector with 1 as the dth entry and 0s elsewhere.

We assume

$$(z_i|w_i = E_d) \sim N(\boldsymbol{\mu}_d, \boldsymbol{\Sigma}), \tag{7}$$

independently for $i = 1, 2 \ldots, n$, where μ_d is a q-vector of means corresponding to cell d, and $\boldsymbol{\Sigma}$ is a $q \times q$ covariance matrix. The means of $Z_1, Z_2, \ldots, Z_q$ are allowed to vary from cell to cell, but a common covariance structure $\boldsymbol{\Sigma}$ is assumed for all cells.

The parameters of the general location model will be written as

$$\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{8}$$

where $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_D)^T$ is a $D \times q$ matrix of means. The number of parameters to be estimated in the model is thus

$$(D - 1) + Dq + q(q + 1)/2.$$

5.2 EM algorithm

The EM algorithm is a general technique for finding maximum-likelihood estimates for parametric models when the data are not fully observed. The fact that missing data contains information relevant to estimating θ, and θ in turn helps us to find likely values of missing data. Dempster, Laird and Rubin (1977) suggest the following scheme for estimating θ in the presence of the observed data alone: "Input" the missing data based on an initial estimate of θ, re-estimate θ based on the observed data and the filled-in data and iterate until the estimates converge, naming the algorithm Expectation-Maximization or EM.

Now suppose some of the Z's and W's are missing. For subject i, let $Z_{obs,i}$ denote the vector of observed continuous variables, $Z_{mis,i}$ denote the vector of missing continuous variables, and S_i denote the set of cells in the contingency table where subject i could lie, given the observed categorical variables. Consider the EM algorithm for ML estimation of $\boldsymbol{\theta}$ given data $\{Z_{obs,i}, S_i : i = 1, \ldots, n\}$.

The EM algorithm consists of two steps as follows.

5.2.1 The E-step

To begin, specify initial estimates of the cell means $\mathbf{\Gamma}^{(0)} = [\boldsymbol{\mu}_1^{(0)}, \ldots, \boldsymbol{\mu}_D^{(0)}]$, cell probabilities $\boldsymbol{\pi}^0 = (\pi_1^{(0)}, \ldots, \pi_D^{(0)})^T$, and covariance matrix $\mathbf{\Omega}^{(0)}$. At iteration t the E step computes the expected values of the complete-data sufficient statistics given data $\{Z_{obs,i}, S_i : i = 1, \ldots, n\}$ and current parameter estimates $\boldsymbol{\theta}^{(t)} = (\mathbf{\Gamma}^{(t)}, \mathbf{\Omega}^{(t)}, \mathbf{\Pi}^{(t)})$. The E step of the algorithm computes

$$\begin{aligned} T_{1i} &= E(Z_i Z_i^T | Z_{i,obs}, S_i, \mathbf{\Gamma}^{(t)}, \mathbf{\Omega}^{(t)}, \boldsymbol{\pi}^{(t)}), \\ T_{2i} &= E(W_i Z_i^{'} | Z_{i,obs}, S_i, \mathbf{\Gamma}^{(t)}, \mathbf{\Omega}^{(t)}, \boldsymbol{\pi}^{(t)}), \\ T_{3i} &= E(W_i | Z_{i,obs}, S_i, \mathbf{\Gamma}^{(t)}, \mathbf{\Omega}^{(t)}, \boldsymbol{\pi}^{(t)}). \end{aligned} \tag{9}$$

5.2.2 The M-step

Computes the complete-data ML estimates with complete-data sufficient statistics replaced by their estimates from the E step:

$$\begin{aligned} \boldsymbol{\pi}^{(t+1)} &= n^{-1} \sum_{i=1}^{n} T_{3i}, \\ \mathbf{\Gamma}^{(t+1)} &= \boldsymbol{D}^{-1} (\sum_{i=1}^{n} T_{2i}), \\ \mathbf{\Omega}^{(t+1)} &= n^{-1} \left[\sum_{i=1}^{n} T_{1i} - \left(\sum_{i=1}^{n} T_{2i} \right)^T \boldsymbol{D}^{-1} \left(\sum_{i=1}^{n} T_{2i} \right) \right]. \end{aligned} \tag{10}$$

where $\boldsymbol{D}$ is matrix with elements of $\sum T_{3i}$ along the main diagonal and 0's elsewhere. The algorithm then returns to the E step to recomputed (10) with the new parameter estimates, and cycle back and forth between E and M steps until convergence.

5.2.3 The I-step

It is convenient to input the missing data for unit i in two stages: first by deciding the cell to which unit i belongs; the cell probabilities are given by (12). After assigning unit i to cell w, we may input the missing continuous variables in $z_{i(mis)}$ according to the multivariate regression on $z_{i(obs)}$. The regression predictor of $z_{i(mis)}$ is

$$z^*_{w,ij} = \mu^*_{w,j} + \sum \sigma^*_{jk} z_{ik},$$

where j indicates the missing covariate in unit i, and $\mu^*_{w,j}$, σ^*_{jk} are current estimates after the sweep operations are done.

We here only consider the single imputation method to simplify the computation.

5.3 Details of the E-step calculations

As shown by Little and Schluchter (1985), the discriminates $\delta^*_{w,i}$ and the parameters of the conditional normal distribution of $z_{i(mis)}$, $\boldsymbol{\theta}^{(t)} = (\boldsymbol{\Gamma}^{(t)}, \boldsymbol{\Omega}^{(t)}, \boldsymbol{\Pi}^{(t)})$ can be obtained by a single application of the sweep operator. Arrange the parameters into a matrix

$$\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\Omega} & \boldsymbol{\Gamma}^T \\ \boldsymbol{\Gamma} & \boldsymbol{P} \end{pmatrix}.$$

where $\boldsymbol{P}$ is a $D \times D$ matrix with elements $P_w = 2\text{log}\pi_w$ on the diagonal and zeros elsewhere. Sweep this $\boldsymbol{\theta}$-matrix on the positions in $\boldsymbol{\Omega}$ corresponding to $z_{i(obs)}$ and obtained

$$\boldsymbol{\theta}^* = \begin{pmatrix} \boldsymbol{\Omega}^* & \boldsymbol{\Gamma}^{*T} \\ \boldsymbol{\Gamma}^* & \boldsymbol{P}^* \end{pmatrix}.$$

Calculation of T_{3i} involves finding $E(w_i|Z_{obs,i}, S_i, \theta^{(t)})$ for each subject. The m_{th} components of this vector will be denoted $\omega_{im} = \Pr(\omega_i = E_m|Z_{obs,i}, S_i, \boldsymbol{\theta}^{(t)})$, the conditional posterior probability that subject i belongs in cell m, given the observed continuous variables. ω_{im} takes the form

$$\omega_{im} = \exp(\delta_m)/\sum_{m \in S_i} \exp(\delta_m), \tag{11}$$

where

$$\delta_m = z_{obs,i}\boldsymbol{\Omega}^{-1}_{obs,i,}\mu^T_{obs,i,m} - \frac{1}{2}\mu_{obs,i,m}\boldsymbol{\Omega}^{-1}_{obs,i}\mu^T_{obs,i,m} + \ln(\pi_m), \tag{12}$$

and $\mu_{obs,i,m}$ and $\Omega_{obs,i}$ are the cell mean and covariance in cell m of the continuous variables.

The predictive probability can be found by the following two steps: (a) Sweep the $\boldsymbol{\theta}$-matrix on position corresponding to $z_{i(obs)}$ to obtain $\boldsymbol{\theta}^*$; (b) Calculate the discriminates given by (12) for all cells from $z_{i(obs)}$ and $\boldsymbol{\theta}^*$.

To calculate T_{1i} and T_{2i}, write the continuous variables for subject i as $\{z_{ij}, j = 1, \ldots, q\}$. Define $\hat{Z}^{(m)}_{ij} = E(Z_{ij}|Z_{obs,i}, w_i = E_m, \theta^{(t)})$ as the predicted value of Z_{ij} if Z_{ij} is missing. The element in the m_{th} row and j_{th} column of T_{2i}, for $m = 1, \ldots, D$ and $j = 1, \ldots, p$, is obtained by z_{ij} or its estimate by the conditional posterior probability that subject i falls in cell m:

$$E(w_{im}Z_{ij}|Z_{obs,i}, S_i, \theta^{(t)}) = \begin{cases} \omega_{im}\hat{Z}^{(m)}_{ij} & \text{if } Z_{ij} \text{ is missing} \\ \omega_{im}Z_{ij} & \text{if } Z_{ij} \text{ is observed} \end{cases}. \tag{13}$$

The parts of $\hat{Z}_{ij}^{(m)}$ is the predicted values from the multivariate regression of $Z_{i(mis)}$ on $Z_{i(obs)}$ with cell w,

$$\hat{Z}_{ij}^{(m)} = \mu_{w,j}^* + \sum \sigma_{jk}^* Z_{i(obs)}$$

where σ_{jk}^* is the (j,k)th element of $\mathbf{\Omega}^*$.

Given that $w_i = E_m$, let $\sigma_{jk,obs,i}$ denote the conditional covariance of Z_{ij} and Z_{ik} given $Z_{obs,i}$. Then the jk_{th} element of T_{1i}, for $k = 1, \ldots, q$ is

$$E(Z_{ij}Z_{ik}|Z_{obs,i}, S_i, \theta^{(t)}) = \begin{cases} Z_{ij}Z_{ik} & \text{both } Z_{ij}, Z_{ik} \text{ observed} \\ Z_{ik}\sum_{m\in S_i}\omega_{im}\hat{Z}_{ij}^{(m)} & Z_{ij} \text{ missing, } Z_{ik} \text{ observed} \\ Z_{ij}\sum_{m\in S_i}\omega_{im}\hat{Z}_{ik}^{(m)} & Z_{ij} \text{ observed, } Z_{ik} \text{ missing} \\ \sigma_{ik,obs,i} + \sum_{m\in S_i}\omega_{im}\hat{Z}_{ij}^{(m)}\hat{Z}_{ik}^{(m)} & \text{both } Z_{ij}, Z_{ik} \text{ missing} \end{cases}.$$

The computations of the EM algorithm are easily performed by sweep operations, which was first introduced by Beaton (1964).

6 Robust diagnostics and missing values

In this section, we extend the forward search algorithm to incomplete data, that is, integrate the forward search algorithm and the EM algorithm.

6.1 Combination of EM algorithm and the forward search

To start the forward search with missing data, we first obtain an initial subset from the fully observed data. This can be done by the forward search method for the generalized linear model as shown in Section 3. Once the initial subset is obtained, those missing values can be imputed with the EM algorithm based on the selected observations. To be specific, the calculation of equations (10) is only involved with those selected observations.

6.1.1 *Step 1: Choice of the initial subset*

Let $(\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{Y})$ be a data containing n observations, where $\boldsymbol{W}$ represents the p categorical variables, $\boldsymbol{Z}$ is a set of q continuous variables, and $\boldsymbol{Y}$ is responses. Hence $\boldsymbol{W}$ and $\boldsymbol{Z}$ together yields the covariates of the data, say $\boldsymbol{X}$. In order to estimate the unknown parameters in the EM algorithm, the forward search algorithm starts with selection of a subset of m units, where m must be large enough in case to estimate the unknown parameters in the EM algorithm.

The initial subset used here is obtained by fitting a logistic regression to the fully observed data. The choice of initial subset can be performed by exhaustive enumeration of all distinct m-tuple, $S^{(m)}_{i_1,\ldots,i_p} \equiv x_{i_1},\ldots,x_{i_m}$, where $x^T_{i_1}$ is the i_1th row of $\boldsymbol{X}$, for $1 \leq i_1,\ldots,i_p \leq n$, and $i_j \neq i_{j'}$. Let $\iota' = [i_1,\ldots,i_p]$ and let $r_{i,S^{(m)}_\iota}$ be the least squared deviance residual for unit i given observations in $S^{(m)}_\iota$. After trying some possible subset, say 100, the initial subset must satisfies

$$d^2_{[med],S^{(m)}_*} = \min_\iota(r^2_{[med],S^{(m)}_\iota}) \tag{14}$$

where $d^2_{[l],S^{(m)}_\iota}$ is the lth order squared deviance residual among $d^2_{i,S^{(m)}_\iota}$, $i = 1,\ldots,n$, and med is the integer part of $(n + number\ of\ variables + 1)/2$. Observations in this subset are intended to be outlier free.

6.1.2 *Step 2: Fill-in the missing values based on the chosen subset*

Denote $\mathcal{M}$ as the set of m observations selected in the subset. Once the initial subset is obtained, apply the EM algorithm to estimate the unknown parameters $\boldsymbol{\theta} = (\boldsymbol{\Gamma},\boldsymbol{\Omega},\boldsymbol{\Pi})$. The missing part of the data can be imputed with these estimates. Specify initial estimates of the cell means $\boldsymbol{\Gamma}^{(0)}_{\mathcal{M}} = [\boldsymbol{\mu}^{(0)}_1,\ldots,\boldsymbol{\mu}^{(0)}_D]$, cell probabilities $\boldsymbol{\pi}^0_{\mathcal{M}} = (\pi^{(0)}_1,\ldots,\pi^{(0)}_D)^T$, and covariance matrix $\boldsymbol{\Omega}^{(0)}_{\mathcal{M}}$ to begin the EM algorithm. Note that these estimates are based on the selected observations.

At iteration t the E step computes the expected values of the selected complete-data sufficient statistics given data $\{X_{obs,i}, S^{(\mathcal{M})}_* : i = i_1,\ldots,i_m\}$ and current parameter estimates $\boldsymbol{\theta}^{(t)}_{\mathcal{M}} = (\boldsymbol{\Gamma}^{(t)}_{\mathcal{M}},\boldsymbol{\Omega}^{(t)}_{\mathcal{M}},\boldsymbol{\Pi}^{(t)}_{\mathcal{M}})$, where $\boldsymbol{\theta}^{(t)}_{\mathcal{M}}$ indicates the parameter estimates base on the selected m observations. The E step of the algorithm computes

$$\begin{aligned} T_{1i,S^{(m)}_*} &= E(Z_iZ_i^{'}|X_{i,obs},S^{(m)}_*,\boldsymbol{\Gamma}^{(t)}_{\mathcal{M}},\boldsymbol{\Omega}^{(t)}_{\mathcal{M}},\boldsymbol{\pi}^{(t)}_{\mathcal{M}}),\\ T_{2i,S^{(m)}_*} &= E(W_iZ_i^{'}|X_{i,obs},S^{(m)}_*,\boldsymbol{\Gamma}^{(t)}_{\mathcal{M}},\boldsymbol{\Omega}^{(t)}_{\mathcal{M}},\boldsymbol{\pi}^{(t)}_{\mathcal{M}}),\\ T_{3i,S^{(m)}_*} &= E(W_i|X_{i,obs},S^{(m)}_*,\boldsymbol{\Gamma}^{(t)}_{\mathcal{M}},\boldsymbol{\Omega}^{(t)}_{\mathcal{M}},\boldsymbol{\pi}^{(t)}_{\mathcal{M}}). \end{aligned} \tag{15}$$

Compute the complete-data ML estimates with selected complete-data sufficient statistics replaced by their estimates from the E step:

$$\pi^{(t+1)}_{\mathcal{M}} = n^{-1}\sum_{i=i_1}^{i_m} T_{3i}\ ,$$

$$\begin{aligned}\Gamma_{\mathcal{M}}^{(t+1)} &= \boldsymbol{D}^{-1}\left(\sum_{i=i_1}^{i_m} T_{2i}\right) , \\ \Omega_{\mathcal{M}}^{(t+1)} &= n^{-1}\left[\sum_{i=i_1}^{i_m} T_{1i} - \left(\sum_{i=i_1}^{i_m} T_{2i})^T \boldsymbol{D}^{-1} (\sum_{i=i_1}^{i_m} T_{2i}\right)\right] . \end{aligned} \tag{16}$$

where $\boldsymbol{D}$ is matrix with elements of $\sum T_{3i}$ along the main diagonal and 0's elsewhere. The algorithm then returns to the E step to recompute (15) with the new parameter estimates, and cycle back and forth between E and M steps until convergence.

The parameter estimates $\boldsymbol{\theta}_{\mathcal{M}}^{(t)} = (\boldsymbol{\Gamma}_{\mathcal{M}}^{(t)}, \boldsymbol{\Omega}_{\mathcal{M}}^{(t)}, \boldsymbol{\Pi}_{\mathcal{M}}^{(t)})$ of the EM algorithm are then used to fill-in the missing part of the data, that is, arrange the parameters into a matrix

$$\boldsymbol{\theta}_{\mathcal{M}} = \left(\begin{array}{cc} \boldsymbol{\Omega}_{\mathcal{M}} & \boldsymbol{\Gamma}_{\mathcal{M}}^T \\ \boldsymbol{\Gamma}_{\mathcal{M}} & \boldsymbol{P}_{\mathcal{M}} \end{array} \right). \tag{17}$$

Missing values then can be obtained by application of the sweep operator.

6.1.3 *Step 3: Adding observations during the forward search*

The forward search selects the $m + 1$ units with the smallest squared residuals of the fitted data, the units being chosen by ordering all squared deviance residuals $r^2_{i,S_*^{\mathcal{M}}}, i = 1, ..., n$. The residuals are computed by (3), where $\boldsymbol{\theta}$ is the maximum likelihood estimated coefficients of the previous m units included in the model.

6.1.4 *Step 4: Recursion of EM Algorithm and LMS*

The algorithm repeats Step 2 and 3 until that all observations are included in the subset. We need apply the EM algorithm based on the selected observations in step 3, re-estimate the unknown parameters, $\boldsymbol{\theta} = (\boldsymbol{\Gamma}, \boldsymbol{\Omega}, \boldsymbol{\Pi})$, and impute those missing observations. We also fit a logistic regression to the currently selected observations, and order the deviance residuals to decide the observation to be selected for the following step. In each step,

$$d^2_{[med],S_t^{(m)}}$$

is retained, where *med* is the integer part of $(n + number\ of\ variables + 1)/2$ and m is the number of units in the subset.

6.1.5 *Step 5: Monitoring the search*

Monitor the squared deviance at each step,

$$d^2_{[med],S_\iota^{(m)}}, d^2_{[med],S_\iota^{(m+1)}}, \ldots, d^2_{[med],S_\iota^{(n)}},$$

Choose the smallest deviance residuals of the set, $\min(r^2_{[med],S_\iota^{(p)}})$, the corresponding subset is then used to estimated the unknown parameters of EM algorithm, and fit the logistic regression. The final result is deviance residuals of all the observations based on corresponding subset. Absolute deviance residuals larger than 2 are the potential outliers.

6.2 Example: Recumbent cow data (continued)

The recumbent cow data introduced in Section 3.5, one objective of this study was to understand the effects of varies characteristic on survival probability of cows just before or just after calving. We consider five covariates. The continuous variables measured on the ith cow are *Daysrec*, days recumbent when measurements were taken; *AST*, seurm asparate amino transferase; *CK*, seurm creatine phosphokinase. Thus there are $q = 3$ continuous variables and $p = 3$ categorical variables. These 3 categorical variables form a $2 \times 2 \times 2$ contingency table with 8 cells. The proportion of missing values for each variable is shown in Table 4.

Table 4. *Proportion of missing values for recumbent cows data*

Variables	*% missing*	*number*
Outcome	0	0
Calving	0.009	4
Myopathy	0.489	213
Dayserc	0.006	3
AST	0.013	6
CK	0.050	22

There are $(8-1) + 8 \times 3 + 3(3+1)/2 = 37$ unknown parameters. The initial subset of the forward search algorithm then must start with any $m \geq 37$ cases. We here use $m = 100$. Select 100 observations from the 216 complete-observed data.

To find this initial subset involves the LMS method for generalized linear models which was described in Section 3. Arbitrary try 100 possible subsets sampled from the 216 complete-observed observations. Initial subsets are allowed to be different from each, because this does not effect the result; that is, different initial subsets leads to the same conclusion. The chosen units are used to estimate the initial covariance matrix $\mathbf{\Omega}$ and cell means $\mathbf{\Gamma}$ of EM algorithm, which are used to start the algorithm.

When applying the proposed algorithm, the LMS solution occurs at step 303 of the search. It reveals that 131 observations have absolute deviance residuals larger than 2. It is obvious that there is serious loss of information of deleting missing values from the analysis in comparison with the previous results.

References

Atkinson, A. C. (1994), "Fast Very Robust Methods for the Detection of Multiple Outliers," *Journal of the American Statistical Association*, **89**, 1329-1339.

Atkinson, A. C. and Cheng, T.-C. (2000) "On Robust Linear Regression with Incomplete Data", *Computational Statistics and Data Analysis*, **33**, 361-380.

Atkinson, A. C. and Riana, M. (2000), *Robust Diagnostic and Regression Analysis*, New York: Springer-Verlag.

Atkinson, A. C. and Riana, M. (2001), "Regression Diagnostics for Binomial Data from the Forward Search", *The Statistician*, **50**, 63-78.

Christmann, A. (1994) "Least Median of Weighted Squares in Logistic Regression with Large Strata", *Biometrika*, **81**, 413-417.

Christmann, A. and Rousseeuw, P. J. (2001) "Measuring Overlap in Binary Regression," *Computational Statistics and Data Analysis*, **37**, 65-75.

Collett, D. (1991) *Modelling Binary Data*, London: Chapman & Hall.

Cook, R. D. and Weisberg, S. (1999) *Applied Regression Including Computing and Graphics*, New York: John Wiley & Sons.

Croux, C., Flandre, C. and Haesbroeck, G. (2002) "The breakdown behavior of the maximum likelihood estimator in the logistic regression", *Statistics & Probability Letters*, **60**, 377-386.

Dempster, A. P., Laird, M. and Rubin, D. B. (1977) "Maximum Likelihood from Incomplete Data via the EM algorithm", *Journal of the Royal Statistical Society*, Ser. B, **39**, 1-38. % item Fuchs, C. (1982), "Maximum Likelihood Estimation and Model Selection in

Contingency Tables with Missing Data," *Journal of the American Statistical Association*, **77**, 270-278.

Haslett, J. (1999) "A Simple Derivation of Deletion Diagnostic Results for the Generalized Model With Correlated Errors," *Journal of the Royal Statistical Society*, Ser. B, **61**, 603-609.

Ibrahim, J. G. (1990), "Incomplete Data in Generalized Linear Models," *Journal of the American Statistical Association*, **85**, 765-769.

Ibrahim, J. G. and Chen, M. H., Lipsitz, S. R., (1999) "Monte Carlo EM for missing covariates in parametric regression models", *Biometrics*, **55**, 591-596.

Kordzakhia, N. Mishra, G. D. and Reiersølmoen, L. (2001) "Robust Estimation in the Logistic Regression Model," *Journal of Statistical Planning and Inference*, **98**, 211-223.

Künsch, H. R., Stefanski, L. A. and Carrol, R. J. (1989) "Conditionally Unbiased Bounded-Influence Estimation in Generalized Regression Models With Applications to Generalized Linear Models," *Journal of the American Statistical Association*, **84**, 460-466.

Little, R. J. A. and Rubin, D. B. (1987), *Statistical Analysis with Missing Data*, New York: John Wiley.

Little, R. J. A. and Schluchter, M. D. (1985), "Maximum Likelihood Estimation for Mixed Continuous and Categorical Data With Missing Values," *Biometrika*, **72**, 497-512.

McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*, 2nd ed., London: Chapman and Hall.

Müller, C. H. and Neykov, N. (2003) "Breakdown Points of Trimmed Likelihood Estimators and Related Estimators in Generalized Linear Models," *Journal of Statistical Planning and Inference*, **116**, 503-519.

Olkin, I. and Tate, R. F. (1961) "Multivariate Correlation Models with Mixed Discrete and Continuous Variables," *Annals of Mathematical Statistics*, **32**, 448-465.

Pregibon, D. (1981) "Logistic Regression Diagnostics," *The annals of Statistics*, **9**, 705-724.

Rousseeuw, P. J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association*, **79**, 871-880.

Rousseeuw, P. J. and Christmann, A. (2003) "Robustness Against Separation and Outliers in Logistic Regression," *Computational Statistics and Data Analysis*, **43**, 315-332.

Rousseeuw, P. J. and Leroy, A. M. (1987), *Robust Regression and Outlier Detection*, New York: John Wiley.

Rubin, D. B. (1976) "Inference and missing data", *Biometrika*, **63**, 581-592.

Schafer, J. L. (1997) *Analysis of Incomplete Multivariate Data*, London: Chapman & Hall.

Stefanski, L. A., Carrol, R. J. and Ruppert, D. (1986), "Optimally Bounded Score Functions for Generalized Linear Models With Applications to Logistic Regression," *Biometrika*, **73**, 413-424.

Schafer, J. L. (1997) *Analysis of Incomplete Multivariate Data*, London: Chapman & Hall.

Wei, G. C. and Tanner, M. A. (1990). "A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm", *Journal of the American Statistical Association*, **85**, 699-704.