Monte Carlo EM algorithm in logistic linear models involving non-ignorable missing data

Jeong-Soo Park a,*, Guoqi Q. Qian b, Yuna Jun c

a Department of Statistics, Chonnam National University, Gwangju, Republic of Korea
b Department of Statistics, La Trobe University, Victoria 3086, Australia
c Samsung Tesco LTD, SambuBldg 676 Yeoksamdong, Seoul, 135-979, Korea

Abstract

Many data sets obtained from surveys or medical trials often include missing observations. Since ignoring the missing information usually cause bias and inefficiency, an algorithm for estimating parameters is proposed based on the likelihood function of which the missing information is taken account. A binomial response and normal exploratory model for the missing data are assumed. We fit the model using the Monte Carlo EM (Expectation and Maximization) algorithm. The E-step is derived by Metropolis–Hastings algorithm to generate a sample for missing data, and the M-step is done by Newton–Raphson to maximize the likelihood function. Asymptotic variances and the standard errors of the MLE (maximum likelihood estimates) of parameters are derived using the observed Fisher information.

Keywords: Conditional expectation; Fisher information matrix; Maximum likelihood estimation; Metropolis–Hastings algorithm; Newton–Raphson iteration; Standard error

1. Introduction

Many data sets obtained from surveys or medical trials often include missing observations [1]. When these data sets are analyzed, it is general to use only complete cases with data all observed after removing missing data. However, this may cause some problems if the missing data is related to the values of the missing item [2]. The estimate of parameter could be biased and be inefficient [3]. So we need some method for utilizing the partial information involved in the missing data instead of ignoring them. Little and Rubin [3] described many statistical methods dealing with the missing data. Baker and Laird [4] used the EM (Expectation and Maximization) algorithm to obtain maximum likelihood estimates (MLE) of parameters from the incomplete data. Ibrahim and Lipsitz [5,6] presented Bayesian methods for estimation in generalized linear models.

* This work was supported by Korean Science and Engineering Foundation (KOSEF) grant funded by the Korea government (MOST) (R01-2006-000-11087-0).
* Corresponding author.
E-mail addresses: jspark@chonnam.ac.kr (J.-S. Park), g.qian@latrobe.edu.au (G.Q. Qian), umumu@hanmail.net (Y. Jun).
Our proposed method stems from [5,6], and can be thought as an extended and modified version for different model.

There are two types of missing data: Ignoable and non-ignorable [3]. Missing data is called ignorable (non-ignorable) if the probability of observing a data item is independent of (dependent on) the value of that data item. The data that is missing at random is ignorable, while non-ignorable missing data is not at random.

In this paper, we propose a method for estimating parameters in logistic linear models involving non-ignorable missing data. A binomial response and normal covariate model for the missing data is assumed. The Monte Carlo EM algorithm is used to estimate parameters [7]. Metropolis–Hastings algorithm to generate a sample for missing data is used in the E-step. Newton–Raphson iteration to solve the score equation is used to maximize the conditional expectation of likelihood function in the M-step. The standard errors of these estimates are calculated by the observed Fisher information matrix.

The rest of this paper is organized as follows. In Section 2, notation and model are stated. In Section 3 we derive the E- and M-steps of the Monte Carlo EM algorithm including Metropolis–Hastings algorithm and Newton–Raphson iteration. Calculation for standard error is described in Section 4. In Section 5 we illustrate our method with one example. A summary is given at the last section. Details of derivatives for Newton–Raphson iteration and formulas for elements of observed Fisher information matrix are given in the Appendix.

2. Notation and model

Suppose that \( y_1, \ldots, y_n \) are independent observations, where each \( y_i \) has a binomial distribution with sample size \( m_i \) and success probability \( \pi_i \). Let \( X_i = (X_{1i}, X_{2i})' \) is a 2 \( \times \) 1 random vector of covariates, where \( X_{1i} \) and \( X_{2i} \) are independent observations and follow normal distributions with means \( \mu_1, \mu_2 \) and variances \( \sigma_1^2, \sigma_2^2 \), respectively. Further, let \( \beta = (\beta_0, \beta_1, \beta_2) \) are regression coefficients assuming to include an intercept. It is also assumed that

\[
\text{logit}(\pi_i) = \log \frac{\pi_i}{1 - \pi_i} = X_i' \beta, \quad \text{and} \quad p(y_i|X_i, \beta) = \frac{\exp(y_iX_i'\beta)}{1 + \exp(X_i'\beta)}.
\] (1)

We assume that \( X_{1i} \) is completely observed, and \( Y_i \) and \( X_{2i} \) are partially missing. Our objective is to estimate \( \beta, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2 \) (using maximum likelihood estimation) and their standard errors from the given data with non-ignorable missing values.

Missing value indicators are introduced [3] as

\[
r_i = \begin{cases} 0 & \text{if } y_i \text{ is observed}, \\ 1 & \text{if } y_i \text{ is missing}, \end{cases} \quad \text{and} \quad s_i = \begin{cases} 0 & \text{if } x_{2i} \text{ is observed}, \\ 1 & \text{if } x_{2i} \text{ is missing}, \end{cases}
\] (2)

with probabilities \( P(r_i) = \psi_i, P(s_i) = \phi_i \). Following [6], the non-ignorable missing-data mechanism is defined as

\[
\begin{align*}
\text{logit}(\psi_i) &= \delta_1 X_{1i} + \delta_2 X_{2i} + y_i \omega, \\
\text{logit}(\phi_i) &= \alpha_1 X_{1i} + \alpha_2 X_{2i} + y_i \tau,
\end{align*}
\] (3)

where \( \delta = (\delta_1, \delta_2)' \), \( \alpha = (\alpha_1, \alpha_2)' \), \( \omega \) and \( \tau \) are parameters determining the missing mechanism. The conditional probability functions for \( r_i \) and \( s_i \) are derived by Eqs. (1) and (3) as

\[
p(r_i|X_i, y_i, \delta, \omega) = \frac{\exp(r_i(X_i'\delta + y_i \omega))}{1 + \exp(X_i'\delta + y_i \omega)},
\] (4)

\[
p(s_i|X_i, y_i, \alpha, \tau) = \frac{\exp(s_i(X_i'\alpha + y_i \tau))}{1 + \exp(X_i'\alpha + y_i \tau)}.
\] (5)
Now we derive the joint probability function of \( y_i, x_{2i}, r_i, s_i \) as
\[
p(y_i, x_{2i}, r_i, s_i | x_{1i}) = p(r_i | y_i, X_i, \delta, \omega)p(s_i | y_i, X_i, \alpha, \tau)p(y_i | X_i, \beta)p(x_{2i} | x_{1i})
\]
\[
\propto \frac{\exp\{r_i(X_i'\delta + y_i\omega)\}}{1 + \exp(X_i'\delta + y_i\omega)} \times \frac{\exp\{s_i(X_i'\alpha + y_i\tau)\}}{1 + \exp(X_i'\alpha + y_i\tau)} \times \exp\{X_i'\beta y_i\} \times (1 + \exp\{X_i'\beta\})^{-m},
\]
\[
\times (2\pi\sigma_2^2)^{-1/2} \times \exp\left\{-\frac{(x_2 - \mu_2)^2}{2\sigma_2^2}\right\}.
\]
Therefore, we can write down the complete-data log-likelihood \( l(\theta) \) by
\[
\log L(\theta | y_i, X_i, r_i, s_i) = \sum_{i=1}^{n} \log \left( \frac{\exp\{r_i(X_i'\delta + y_i\omega)\}}{1 + \exp(X_i'\delta + y_i\omega)} \right) + \sum_{i=1}^{n} \log \left( \frac{\exp\{s_i(X_i'\alpha + y_i\tau)\}}{1 + \exp(X_i'\alpha + y_i\tau)} \right) + \sum_{i=1}^{n} X_i'\beta y_i
\]
\[
- \sum_{i=1}^{n} m_i \log(1 + \exp(X_i'\beta)) - \frac{n}{2} \log(2\pi\sigma_2^2) - \sum_{i=1}^{n} \frac{(x_2 - \mu_2)^2}{2\sigma_2^2},
\]
where \( \theta = (\beta, \delta, \omega, \alpha, \tau, \mu_2, \sigma_2^2) \) is the parameters related to develop EM algorithm. The complete-data log-likelihood specifies a model for the joint characterization of the observed data and the associated missing-data mechanism.

3. Monte Carlo EM algorithm

3.1. Algorithm formulation

The MLE of \( \beta \) and other components of \( \theta \) are the ones maximizing the observed-data likelihood \( L(\theta | (y, X)_{\text{obs}}, r_i, s_i) \) that has a quite intractable analytical form, where \( (y, X)_{\text{obs}} \) is the observed components of \( (y, X) \). Rather than directly differentiating \( L(\theta | (y, X)_{\text{obs}}, r_i, s_i) \) with respect to \( \theta \), we compute the MLE of \( \theta \) using an EM algorithm [8] which involves iterative evaluation and maximization of conditional expectation of the complete-data log-likelihood \( l(\theta) \). If the conditional expectation involved is difficult to be evaluated exactly, a Monte Carlo EM (MCEM) algorithm [9] can be used where a Gibbs sampler or Metropolis–Hastings algorithm [10] is used to approximate the conditional expectation.

Specifically, let \( \theta' \) be the current estimate of \( \theta \) and define the conditional expectation of \( l(\theta) \) – with respect to the conditional distribution of the missing data \( (y, X)_{\text{mis}} \) given the observed data \( (y_i, X_i, r_i, s_i) \) and the value \( \theta' \) – as the following:
\[
Q(\theta, \theta') = E[l(\theta) | (y, X)_{\text{obs}}, r, s, \theta'].
\]
The feasibility of calculating the conditional expectation in \( Q(\theta, \theta') \) is dependent on the complexity of the conditional distribution of the missing data.

The EM algorithm is composed of expectation (E-step) and maximization (M-step) iterations. Now for the expectation of the complete-data log-likelihood in the E-step of EM algorithm, we consider four possible cases: response variable \( y_i \) is missing, a covariate \( x_{2i} \) is missing, both of them are missing, and no missing values. Then the expected log-likelihood is written by
\[
E[l(\theta) | X_i, y_i, r_i, s_i] = \begin{cases} 
\sum_{y_i=0}^{m_i} l(\theta)p(y_i | X_i, r_i, s_i) & \text{(if } y_i \text{ has missing components)}, \\
\int l(\theta)p(x_{2i} | x_{1i}, y_i, r_i, s_i)dx_{2i, \text{mis}} & \text{(if } x_{2i} \text{ has missing components)}, \\
\sum_{y_i=0}^{m_i} \int l(\theta)p(y_i, x_{2i} | x_{1i}, r_i, s_i)dx_{2i, \text{mis}} & \text{(if } y_i \text{ and } x_{2i} \text{ have missing components)}, \\
l(\theta) & \text{(for no missing values)}, 
\end{cases}
\]
where \( x_{2i, \text{mis}} \) is the missing components of \( x_{2i} \). Eqs. (8) and (9) lead to the conditional expectation of \( h(\theta) \) which is our target quantity as

\[
Q(\theta, \theta') = \sum_{i=1}^{n_1} l(\theta) + \sum_{i=n_1+1}^{n_2} \sum_{y_i=0}^{n_3} l(\theta)p(y_{i,\text{mis}}|X_i, r_i, s_i, \theta') + \sum_{n_2+1}^{n_3} \int l(\theta)p(x_{2i,\text{mis}}|X_{i,\text{obs}}, y_i, r_i, s_i, \theta')dx_{2i,\text{mis}} \\
+ \sum_{n_2+1}^{n_3} \sum_{y_i=0}^{n_3} \int l(\theta)p(y_{i,\text{mis}}, x_{2i,\text{mis}}|X_{i,\text{obs}}, r_i, s_i, \theta')dx_{2i,\text{mis}},
\]

where \( n_1, n_2, n_3 \) are corresponding sample sizes, \( \theta' \) is the \( r \)th iteration estimate of \( \theta \), \( y_{i,\text{mis}} \) is the missing components of \( y_i \), \( X_{i,\text{obs}} \) is the observed component of \( X_i \), and \( p(y_{i,\text{mis}}|X_i, r_i, s_i) \), \( p(x_{2i,\text{mis}}|X_{i,\text{obs}}, y_i, r_i, s_i) \) and \( p(y_{i,\text{mis}}|X_{i,\text{obs}}, r_i, s_i) \) are the conditional probabilities of the missing data given the observed data. These conditional probabilities are regarded as the weights in \( Q(\theta, \theta') \). The weights have the form respectively as following:

\[
p(y_{i,\text{mis}}, x_{2i,\text{mis}}|X_{i,\text{obs}}, r_i, s_i, \theta') = \frac{p(y_i|X_i, \theta')p(x_{2i}|x_{1i})p(r_i|y_i, X_i, \theta')p(s_i|y_i, X_i, \theta')}{\sum_{y_i=0}^{n_3} p(y_i|X_i, \theta')p(x_{2i}|x_{1i})p(r_i|y_i, X_i, \theta')p(s_i|y_i, X_i, \theta')},
\]
(11)

\[
p(x_{2i,\text{mis}}|X_{i,\text{obs}}, y_i, r_i, s_i, \theta') = \frac{p(x_{2i}|x_{1i}, \theta')p(s_i|y_i, X_i, \theta')}{p(x_{2i}|x_{1i}, \theta')} \times p(x_{2i}|x_{1i}, \theta')p(s_i|y_i, X_i, \theta') \\
\times \exp \left\{ \frac{S_i(X_i'z + y_i\tau)}{1 + \exp\{S_i(X_i'z + y_i\tau)\}} \times \left( \frac{2\pi\sigma_2^2}{} \right)^{-1/2} \times \exp \left\{ -\frac{(x_{2i} - \mu_2)^2}{2\sigma_2^2} \right\} \right\},
\]
(12)

\[
p(y_{i,\text{mis}}|X_i, r_i, s_i, \theta') = \frac{p(y_i|X_i, \theta')p(r_i|y_i, X_i, \theta')}{\sum_{y_i=0}^{n_3} p(y_i|X_i, \theta')p(r_i|y_i, X_i, \theta')} \times p(y_i|X_i, \theta')p(r_i|y_i, X_i, \theta').
\]
(13)

Since \( X_2 \) is continuous random variable, the number of possible values that \( (y_i, X_i')_{\text{mis}} \) can take is almost infinity. Thus the above weights (11)–(13) are not explicitly computed and so is the exact computing of \( Q(\theta, \theta') \). In this situation, we can use a Gibbs sampler or Metropolis–Hastings algorithm to simulate a sample of \( (y_i, X_i')_{\text{mis}} \) values and use the associated empirical distribution to approximate the weights. This means that the conditional expectation \( Q(\theta, \theta') \) is to be calculated by a Monte Carlo approximation, which leads a Monte Carlo EM (MCEM) algorithm.

3.2. Metropolis–Hastings algorithm

For generation of random samples from weights functions (11)–(13), it is particularly convenient to use a Metropolis–Hastings (MH) algorithm [10] where we choose the multivariate normal distribution as the operating transition density and an easily verified acceptance–rejection condition. When some of \( y_i \) and \( x_{2i} \) are missing, the algorithm uses the following steps to generate a sample of \( (y_i, X_{2i})_{\text{mis}} \) [11]. When \( \theta' \), \( m_i \), \( x_{1i} \) and \( x_{2i} \) are given, it is easy to generate \( y_i \) from the binomial distribution \( B(m_i, \pi_i) \) where

\[
\pi_i = \exp\{x_i'\beta'\}/(1 + \exp\{x_i'\beta'\}).
\]
(14)

So we will present a MH algorithm of generating \( \{x_{2i}\}_{\text{mis}} \) for given \( x_{i,\text{obs}}, y_i \) and \( \theta' \).

Step (1) Set the initialization of parameters \( \theta', x_{1i,\text{obs}}, r_i, s_i \).
Step (2) Repeat the following steps for \( k = 0, 1, \ldots, n - 1 \):

1. Generate \( (y_{i,\text{mis}}^{(k)}, x_{2i,\text{mis}}^{(k)}) \) from (11).
2. Generate \( x_{2i}^{(k)} \) and \( y^{(k)} \) from their distributions where \( x_2 \sim N(\mu_2, \sigma_2^2) \) and \( y \sim B(m_k, \pi_k) \), where \( \pi_k \) is given by (14) for the \( x_k \) sample generated.
3. Compute the acceptance probability

\[ \alpha^{(k)} = \alpha(x_{2,\text{mis}}^{(k)}, \pi_k), \]

\[ = \min \left\{ \frac{\pi_k P(x_{2,\text{mis}}^{(k)} | x_2^{(k)}, y^{(k)})}{\pi(x_{2,\text{mis}}^{(k)}) P(x_2^{(k)}, y^{(k)} | x_{2,\text{mis}}^{(k)})}, 1 \right\}, \]

where \( \pi(x_{2,\text{mis}}^{(k)}) \) is calculated using \( x_{2,\text{mis}}^{(k)} \) according to the Eq. (14).

4. Take

\[ x_{2,\text{mis}}^{(k+1)} = \begin{cases} x_2^{(k)} & \text{with probability } \alpha(k), \\ x_{2,\text{mis}}^{(k)} & \text{with probability } 1 - \alpha(k). \end{cases} \]

Step (3) Obtain the sample \( \{x_{2,\text{mis}}^{(1)}, \ldots, x_{2,\text{mis}}^{(n)}\} \).

### 3.3. M-step and convergence

Now turn to the M-step in the MCEM algorithm where we need to find a value of \( \theta \), say \( \theta' \), at which \( Q(\theta, \theta') \) will attain the maximum. This can be done by solving the score equation which sets to 0 the derivative of \( Q(\theta, \theta') \) with respect to \( \theta \). The Newton–Raphson method will be used to solve the score equation. The parameters \( \theta^{r+1} = (\beta, \delta, \omega, \alpha, \tau) \) in the M-step at the \( (r+1) \)st EM iteration and the \( (t+1) \)st Newton–Raphson iteration take the form (for \( \beta \) for example):

\[ \beta^{r+1} = \beta' \left( -\frac{\partial^2 Q(\theta, \theta')}{\partial \beta \partial \beta'} \right)^{-1} \left( \frac{\partial Q(\theta, \theta')}{\partial \beta} \right)|_{\beta = \beta'}. \]

The details of the derivatives used in the iteration are given in the Appendix. The \( (r+1) \)st estimates of \( \mu_2, \sigma_2^2 \) are obtained to by solving the score equations:

\[ \frac{\partial}{\partial \mu} Q(\theta, \theta') = \sum_{i=1}^{n} E(x_{2i} | x_{1i}, y_{i}, r_i, s_i) - n \mu_2 = 0, \]

\[ \frac{\partial}{\partial \sigma_2} Q(\theta, \theta') = \sum_{i=1}^{n} E((x_{2i} - \mu_2)^2 | x_{1i}, y_{i}, r_i, s_i) - n \sigma_2^2 = 0. \]

Therefore, we take \( \mu_2^{r+1}, \sigma_2^{2(r+1)} \) by

\[ \mu_2^{r+1} = \frac{1}{n} E(x_{2i} | x_{1i}, y_{i}, r_i, s_i), \]

\[ \sigma_2^{2(r+1)} = \frac{1}{n} E((x_{2i} - \mu_2)^2 | x_{1i}, y_{i}, r_i, s_i), \]

which are approximated by the sample averages of simulated and given observations.

Since MCEM algorithm needs iterations, convergence check is required to get reliable result. The sequence \( \{Q(\theta, \theta')\} \) often exhibits an increasing trend, and then fluctuate around the value of \( Q(\theta, \theta) \) once \( r \) becomes sufficiently large. The sequence \( \{\theta'\} \) would also fluctuate the the MLE \( \hat{\theta} \) when \( r \) is sufficiently large. To monitor the convergence of MCEM algorithm we can plot \( \{Q(\theta, \theta')\} \) as well as \( \{\theta'\} \) against iteration number. We terminate the algorithm when the sequence of \( \{Q(\theta, \theta')\} \) become stationary. Otherwise, we continue by increasing the Monte Carlo precision in the E-step provided that the required calculation is computationally feasible.

### 4. Standard errors of estimates

It is well known that the distribution of maximum likelihood estimates \( \hat{\theta} \) asymptotically follows a normal distribution \( \text{MVN}(\theta, V(\theta)) \) under some regularity condition. The expected Fisher information
matrix $I(\hat{\theta})$ which gives the inverse of variance matrix of $\hat{\theta}$ is approximated by the observed information matrix $J_\theta(Y)$:

$$
V(\hat{\theta})^{-1} = nE \left[ -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} \right]_{\theta = \hat{\theta}} \propto n \int \left[ -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} \right] d\theta \approx \sum_{i=1}^n \left[ -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} \right]_{\theta = \hat{\theta}} \approx nJ(\hat{\theta}).
$$

(22)

We apply the result of [12] on the information of $\theta$:

observed information = complete information – missing information,

so that we have

$$
I(\hat{\theta}) \approx J_\theta(Y) = -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} = \left[ -\frac{\partial^2 Q(\theta, \theta)}{\partial \theta^2} - \text{Var}_\theta \left( \sum_{i=1}^n \frac{\partial \log L(\theta)}{\partial \theta} \right) \right]_{\theta = \hat{\theta}},
$$

(24)

where Var$_\theta(\cdot)$ is the conditional variance given $(y, X)_{\text{obs}}, r, s$ and $\theta'$. The details are to be provided in the Appendix.

5. An illustration

In this section, we show an example to illustrate MCEM algorithm method with missing response variable and a covariate in logistic regression model. At first, we generate covariate $x_1$, $x_2$, independently at random. Each $x_1$, $x_2$, has normal distribution $N(\mu_1, \sigma_1^2)$, $N(\mu_2, \sigma_2^2)$, respectively. The response variable $y_i$ is generated from binomial distribution with sample size $m_i$, probability $p_i$ by generating $x_{1i}$, $x_{2i}$ where $p_i = \exp(x'\beta)/(1 + \exp(x'\beta))$. We apply missing data mechanism to generate missing data for variable $y_i$, $x_{1i}$, where $y_i$ and $x_i$ are defined in (3). Then each $r_i$, $s_i$ is generate from Bernoulli distribution with success probability $\psi_i$, $\phi_i$. The data set with missing observations generated by the above procedure is presented in Table 1, where ‘–’ is missing, and ‘0’ is observed. At second, we use Metropolis–Hastings algorithm to generate samples from weights (11)–(13) in the E-step. We illustrate the algorithm by each condition.

5.1. Both $y$ and $x_2$ are missing

1. The probability density is defined as Eq. (25) and the conjugate density is normal distribution with mean 0 and variance 1.

$$
f(X) = p(y, x_1, x_2, y_{\text{mis}}, r, s, 0') \propto p(y, x_2, r, s | x_1, 0')
$$

$$
\propto \exp \left\{ r(x'\delta + y, \omega) \right\} \times \exp \left\{ s(x'x + y, \tau) \right\} \times \exp \left\{ x'\beta \right\}
$$

$$
\times (1 + \exp(x'\beta))^{-m_i} \times (2\pi \sigma_2^2)^{-1/2} \times \exp \left\{ \frac{-(x_2 - \mu_2)^2}{2\sigma_2^2} \right\},
$$

(25)

$$
g(X) = \exp \left\{ \frac{(x_2 - \mu_2)^2}{2} \right\}.
$$

(26)

2. To do the Metropolis–Hastings algorithm, we generate a initial sample $x_{2,0} = (x_{2,01}, \ldots, x_{2,0j})$, $j = 1, 2, \ldots, J$ from normal distribution $N(\mu_2, \sigma_2^2)$. Then we generate a sample $x_2$ from $N(\mu_2, \sigma_2^2)$ and $u$ from uniform

Table 1
A sample data set with missing observations (‘-’ is missing, and ‘0’ is observed) generated by the method described in the Section 5

<table>
<thead>
<tr>
<th>N</th>
<th>1</th>
<th>2</th>
<th>···</th>
<th>n$_1$</th>
<th>n$_1+1$</th>
<th>···</th>
<th>n$_2$</th>
<th>n$_2+1$</th>
<th>···</th>
<th>n$_3$</th>
<th>n$_3+1$</th>
<th>···</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>x$_1$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>x$_2$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
distribution. After computing $\pi(x_{j-1}, \bar{x}_j)$, if $u$ is lower than $\pi$ then $x_{2i} = \bar{x}_j$, otherwise $x_{2i} = x_{2i-1}$. We generate $k$ sample sets repeating the step $k = 1, 2, \ldots, K$. Now we can generate $y_i$ from binomial distribution with the computed $\pi_i = \exp\{X_i'\beta\}/(1 + \exp\{X_i'\beta\})$ using the $x_{2i}$ sample generated. And

$$\pi(x_{j-1}, \bar{x}_j) = \min\left\{\frac{f(x_{j-1})g(\bar{x})}{f(\bar{x})g(x_{j-1})}, 1\right\}. \quad (27)$$

5.2. Only $x_2$ is missing

The probability density is defined as Eq. (28) and the conjugate density is normal distribution with mean 0 and variance 1. Also, the method is same as in case when $y_i$ and $x_{2j}$ are missing

$$f(X) = p(x_{2i}, m|x_{i, \text{obs}}, y_i, r_i, s_i, \theta') \propto p(x_{2i}|x_{i, \text{obs}}, \theta')p(r_i|y_i, X_i, \theta')$$

$$\propto \exp\{s_i(X_i'x + y_i')\}
\times \exp\left\{-\frac{(x_{2i} - \mu_2)^2}{2\sigma_2^2}\right\}, \quad (28)$$

$$g(X) = \exp\left\{-\frac{(x_{2i} - \mu_2)^2}{2}\right\}. \quad (29)$$

5.3. Only $y$ is missing

The probability density is defined as Eq. (30) and the conjugate density is binomial distribution

$$f(X) = p(y_i, m|x_i, r_i, s_i, \theta') \propto p(y_i|X_i, \theta')p(r_i|y_i, X_i, \theta')$$

$$\propto \exp\{r_i(X_i'\delta + y_i')\}
\times \exp\{X_i'\beta y_i\}
\times (1 + \exp\{X_i'\beta\})^{-m_i}, \quad (30)$$

$$g(X) = \exp\{X_i'\beta y_i\}
\times (1 + \exp\{X_i'\beta\})^{-m_i}. \quad (31)$$

6. Summary

An algorithm for estimating parameters in logistic linear models is proposed for incomplete data. When some of the response and a covariate observations are missing non-ignorably, the maximum likelihood estimation (MLE) is considered. Metropolis–Hastings (MH) algorithm to compute the conditional expectation of log-likelihood function is implemented in the proposed Monte Carlo EM (Expectation and Maximization) algorithm. Newton–Raphson iteration is used in the M-step of the algorithm. For the standard error of MLE is also considered using the observed Fisher information matrix. Details of MH algorithm, derivatives needed in M-step, and formulas of observed Fisher information matrix are given with an illustration.

Acknowledgements

This paper was started while the third author was visiting the Department of Statistics, La Trobe University, Australia. She thanks Dr. Richard Huggins and the staff in the department for their hospitality and support.

Appendix 1. Derivatives of $Q(\theta, \theta')$ for M-step

Iterating E-step and M-step, the $(r+1)$st Newton–Raphson step estimates of $\theta^{r+1} = (\beta, \delta, \omega, x, \tau)$ can be obtained using these derivatives
\[
\frac{\partial}{\partial \beta} Q(\theta, \theta') = \sum_{i=1}^{n} x'_i [y_i + \sum_{i=1}^{n} E(x'_i | x_i, \theta') + \sum_{i=1}^{n} E(x'_i | x_{obs}, y_i, \theta')] \\
+ \sum_{i=n+1}^{n} E(x'_i | x_{obs}, \theta') + \sum_{i=1}^{n} x'_i \pi_i - \sum_{i=1}^{n} E(x'_i | x_{obs}, \theta'), \\
+ \sum_{i=n+1}^{n} E(x'_i | x_{obs}, y_i, \theta') + \sum_{i=1}^{n} E(x'_i | x_{obs}, \theta'), \\
\frac{\partial^2}{\partial \beta \partial \beta} Q(\theta, \theta') = \sum_{i=1}^{n} x'_i \pi_i (\pi_i - 1) x_i + \sum_{i=1}^{n} E(x'_i \pi_i | x_{obs}, y_i, \theta') \\
+ \sum_{i=n+1}^{n} E(x'_i \pi_i | x_{obs}, y_i, \theta') + \sum_{i=1}^{n} E(x'_i \pi_i | x_{obs}, \theta'), \\
\text{where } \pi_i = \exp \{x'_i \beta \}/(1 + \exp \{x'_i \beta \}), \\
\frac{\partial}{\partial \phi} Q(\theta, \theta') = \sum_{i=1}^{n} x'_i (r_i - \psi_i) + \sum_{i=1}^{n} E(x'_i | r_i - \psi_i | x_{obs}, \theta') \\
+ \sum_{i=n+1}^{n} E(x'_i | r_i - \psi_i | x_{obs}, y_i, \theta') + \sum_{i=1}^{n} E(x'_i | r_i - \psi_i | x_{obs}, \theta'), \\
\frac{\partial^2}{\partial \phi \partial \phi} Q(\theta, \theta') = \sum_{i=1}^{n} x'_i \phi_i (1 - \psi_i) + \sum_{i=1}^{n} E(x'_i \phi_i | x_{obs}, y_i, \theta') \\
+ \sum_{i=n+1}^{n} E(x'_i \phi_i | x_{obs}, y_i, \theta') + \sum_{i=1}^{n} E(x'_i \phi_i | x_{obs}, \theta'), \\
\text{where } \psi_i = \exp \{x'_i \delta + y_i \omega \}/(1 + \exp \{x'_i \delta + y_i \omega \}), \\
\frac{\partial}{\partial x} Q(\theta, \theta') = \sum_{i=1}^{n} x'_i (s_i - \phi_i) + \sum_{i=1}^{n} E(x'_i | s_i - \phi_i | x_{obs}, \theta') \\
+ \sum_{i=n+1}^{n} E(x'_i | s_i - \phi_i | x_{obs}, y_i, \theta') + \sum_{i=1}^{n} E(x'_i | s_i - \phi_i | x_{obs}, \theta'), \\
\frac{\partial^2}{\partial x \partial x} Q(\theta, \theta') = \sum_{i=1}^{n} x'_i \phi_i (1 - \phi_i) + \sum_{i=1}^{n} E(x'_i \phi_i | x_{obs}, \theta') \\
+ \sum_{i=n+1}^{n} E(x'_i \phi_i | x_{obs}, \theta') + \sum_{i=1}^{n} E(x'_i \phi_i | x_{obs}, \theta'), \\
\frac{\partial}{\partial \tau} Q(\theta, \theta') = \sum_{i=1}^{n} y'_i (s_i - \phi_i) + \sum_{i=1}^{n} E(y'_i | s_i - \phi_i | x_{obs}, \theta') \\
+ \sum_{i=n+1}^{n} E(y'_i | s_i - \phi_i | x_{obs}, y_i, \theta') + \sum_{i=1}^{n} E(y'_i | s_i - \phi_i | x_{obs}, \theta'), \\
\text{where } \phi_i = \exp \{x'_i \phi + y_i \tau \}/(1 + \exp \{x'_i \phi + y_i \tau \}).
\[
\frac{\partial^2}{\partial^2 \theta} Q(\theta, \theta') = \sum_{i=1}^{n_1} y_i s_i \phi_i (1 - \phi_i) + \sum_{i=n_1+1}^{n_2} E\{y_i s_i \phi_i (1 - \phi_i) | x_{\text{obs}}, \theta'\} + \sum_{i=n_2+1}^{n_3} E\{y_i s_i \phi_i (1 - \phi_i) | x_{\text{obs}}, y_i, \theta'\} + \sum_{i=n_3+1}^{n} E\{y_i s_i \phi_i (1 - \phi_i) | x_{\text{obs}}, \theta'\},
\]
where \( \phi_i = \exp\{x_i'x + y_i\tau\}/(1 + \exp\{x_i'x + y_i\tau\}) \).

**Appendix 2. Observed Fisher information matrix**

The expected Fisher information matrix \( I(\hat{\theta}) \) which gives the inverse of variance matrix of \( \hat{\theta} \) is approximated by the observed information matrix \( J_\theta(Y) \)

\[
V(\hat{\theta})^{-1} = nE\left[ -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} \right]_{\theta = \hat{\theta}} \approx n \int \left[ -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} \right] dx \approx \sum_{i=1}^{n} \left[ -\frac{\partial^2 \log L(\theta)}{\partial \theta} \right]_{\theta = \hat{\theta}} \approx n I(\hat{\theta}),
\]

\[
I(\hat{\theta}) \approx J_\theta(Y) = \left[ -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} \right]_{\theta = \hat{\theta}} = \left[ -\frac{\partial^2 Q(\theta, \theta')}{\partial \theta^2} - \text{Var}_\theta \left( \sum_{i=1}^{n} \frac{\partial \log L(\theta)}{\partial \theta} \right) \right]_{\theta = \hat{\theta}},
\]
where \( \text{Var}_\theta (\cdot) \) is the conditional variance given a \( (y, X)_{\text{obs}}, r, s \) and \( \theta' \).

\[
J_\theta(Y) = \begin{pmatrix}
1 & 2 & 3 & 4 & 5 & 6 & 7 \\
1 & -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} & -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} & -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} & -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} & -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} & -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} \\
2 & -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} & -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} & -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} & -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} & -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} & -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} \\
3 & -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} & -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} & -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} & -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} & -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} & -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} \\
4 & -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} & -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} & -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} & -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} & -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} & -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} \\
5 & -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} & -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} & -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} & -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} & -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} & -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} \\
6 & -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} & -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} & -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} & -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} & -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} & -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} \\
7 & -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} & -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} & -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} & -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} & -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} & -\frac{\partial^2 \log L(\theta)}{\partial \theta^2}
\end{pmatrix}
\]

and

\[
-\frac{\partial^2 Q(\theta, \theta')}{\partial \theta^2} = \begin{pmatrix}
1 & 2 & 3 & 4 & 5 & 6 & 7 \\
1 & -\frac{\partial^2 Q(\theta, \theta')}{\partial \theta^2} & 0 & 0 & 0 & 0 & 0 \\
2 & 0 & -\frac{\partial^2 Q(\theta, \theta')}{\partial \theta^2} & -\frac{\partial^2 Q(\theta, \theta')}{\partial \theta^2} & 0 & 0 & 0 \\
3 & 0 & -\frac{\partial^2 Q(\theta, \theta')}{\partial \theta^2} & -\frac{\partial^2 Q(\theta, \theta')}{\partial \theta^2} & 0 & 0 & 0 \\
4 & 0 & 0 & -\frac{\partial^2 Q(\theta, \theta')}{\partial \theta^2} & -\frac{\partial^2 Q(\theta, \theta')}{\partial \theta^2} & 0 & 0 \\
5 & 0 & 0 & 0 & -\frac{\partial^2 Q(\theta, \theta')}{\partial \theta^2} & -\frac{\partial^2 Q(\theta, \theta')}{\partial \theta^2} & 0 \\
6 & 0 & 0 & 0 & 0 & -\frac{\partial^2 Q(\theta, \theta')}{\partial \theta^2} & -\frac{\partial^2 Q(\theta, \theta')}{\partial \theta^2} \\
7 & 0 & 0 & 0 & 0 & 0 & -\frac{\partial^2 Q(\theta, \theta')}{\partial \theta^2}
\end{pmatrix}.
\]
Now we can estimate the $J_\theta(Y)$ as follows:

\[
J_\theta(Y)_{11} = -\frac{\partial^2 \log L(\theta)}{\partial \theta \partial \theta'} - \frac{1}{k} \sum_{i=1}^{n} \sum_{k=1}^{k} (X'_i y)(X'_i y)'', \\
J_\theta(Y)_{12} = -\frac{1}{k} \sum_{i=1}^{n} \sum_{k=1}^{k} (X'_i y)(X'_i (r_i - \psi_i))', \\
J_\theta(Y)_{13} = -\frac{1}{k} \sum_{i=1}^{n} \sum_{k=1}^{k} (X'_i y)(y_i (r_i - \psi_i))', \\
J_\theta(Y)_{14} = -\frac{1}{k} \sum_{i=1}^{n} \sum_{k=1}^{k} (X'_i y)(X'_i (s_i - \phi_i))', \\
J_\theta(Y)_{15} = -\frac{1}{k} \sum_{i=1}^{n} \sum_{k=1}^{k} (X'_i y)(y_i (s_i - \phi_i))', \\
J_\theta(Y)_{16} = -\frac{1}{k} \sum_{i=1}^{n} \sum_{k=1}^{k} (X'_i y)((x_2 - \mu_2)/\sigma_2^2), \\
J_\theta(Y)_{17} = -\frac{1}{k} \sum_{i=1}^{n} \sum_{k=1}^{k} (X'_i y)(-1/\sigma_2^2 + (x_2 - \mu_2)/\sigma_2^2), \\
J_\theta(Y)_{22} = -\frac{\partial^2 \log L(\theta)}{\partial \theta \partial \theta'} - \frac{1}{k} \sum_{i=1}^{n} \sum_{k=1}^{k} (X'_i(r_i - \psi_i))(X'_i(r_i - \psi_i))'', \\
J_\theta(Y)_{23} = -\frac{1}{k} \sum_{i=1}^{n} \sum_{k=1}^{k} (X'_i(r_i - \psi_i))(y_i (r_i - \psi_i))', \\
J_\theta(Y)_{24} = -\frac{1}{k} \sum_{i=1}^{n} \sum_{k=1}^{k} (X'_i(r_i - \psi_i))(X'_i(s_i - \phi_i))', \\
J_\theta(Y)_{25} = -\frac{1}{k} \sum_{i=1}^{n} \sum_{k=1}^{k} (X'_i(r_i - \psi_i))(y_i (s_i - \phi_i))', \\
J_\theta(Y)_{26} = -\frac{1}{k} \sum_{i=1}^{n} \sum_{k=1}^{k} (X'_i(r_i - \psi_i))((x_2 - \mu_2)/\sigma_2^2), \\
J_\theta(Y)_{27} = -\frac{1}{k} \sum_{i=1}^{n} \sum_{k=1}^{k} (X'_i(r_i - \psi_i))(1/\sigma_2^2 + (x_2 - \mu_2)/\sigma_2^2), \\
J_\theta(Y)_{33} = -\frac{\partial^2 \log L(\theta)}{\partial \theta \partial \theta'} - \frac{1}{k} \sum_{i=1}^{n} \sum_{k=1}^{k} (y_i(r_i - \psi_i))(y_i(r_i - \psi_i))', \\
J_\theta(Y)_{34} = -\frac{1}{k} \sum_{i=1}^{n} \sum_{k=1}^{k} (y_i(r_i - \psi_i))(X'_i(s_i - \phi_i))', \\
J_\theta(Y)_{35} = -\frac{1}{k} \sum_{i=1}^{n} \sum_{k=1}^{k} (y_i(r_i - \psi_i))(y_i (s_i - \phi_i))', \\
J_\theta(Y)_{36} = -\frac{1}{k} \sum_{i=1}^{n} \sum_{k=1}^{k} (y_i(r_i - \psi_i))((x_2 - \mu_2)/\sigma_2^2), \\
J_\theta(Y)_{37} = -\frac{1}{k} \sum_{i=1}^{n} \sum_{k=1}^{k} (y_i(r_i - \psi_i))(1/\sigma_2^2 + (x_2 - \mu_2)/\sigma_2^2),
\[ J_{\theta}(Y)_{44} = -\frac{\partial^2 \log L(\theta)}{\partial \phi \partial \phi'} - \frac{1}{k} \sum_{i=1}^{n} \sum_{k=1}^{k} (X'_i(s_i - \phi_i))(X'_i(s_i - \phi_i))', \]
\[ J_{\theta}(Y)_{45} = -\frac{1}{k} \sum_{i=1}^{n} \sum_{k=1}^{k} (X'_i(s_i - \phi_i))(y_i(s_i - \phi_i))', \]
\[ J_{\theta}(Y)_{46} = -\frac{1}{k} \sum_{i=1}^{n} \sum_{k=1}^{k} (X'_i(s_i - \phi_i))(x_2 - \mu_2)/\sigma_2^2, \]
\[ J_{\theta}(Y)_{47} = -\frac{1}{k} \sum_{i=1}^{n} \sum_{k=1}^{k} (X'_i(s_i - \phi_i))(1/\sigma_2^2 + (x_2 - \mu_2)/\sigma_2^2), \]
\[ J_{\theta}(Y)_{55} = -\frac{\partial^2 \log L(\theta)}{\partial \tau \partial \tau'} - \frac{1}{k} \sum_{i=1}^{n} \sum_{k=1}^{k} (y_i(s_i - \phi_i))(y_i(s_i - \phi_i))', \]
\[ J_{\theta}(Y)_{56} = -\frac{1}{k} \sum_{i=1}^{n} \sum_{k=1}^{k} (y_i(s_i - \phi_i))(x_2 - \mu_2)/\sigma_2^2, \]
\[ J_{\theta}(Y)_{57} = -\frac{1}{k} \sum_{i=1}^{n} \sum_{k=1}^{k} (y_i(s_i - \phi_i))(1/\sigma_2^2 + (x_2 - \mu_2)/\sigma_2^2), \]
\[ J_{\theta}(Y)_{66} = -\frac{\partial^2 \log L(\theta)}{\partial \mu_2 \partial \mu_2} - \frac{1}{k} \sum_{i=1}^{n} \sum_{k=1}^{k} (x_2 - \mu_2)/\sigma_2^2, \]
\[ J_{\theta}(Y)_{67} = -\frac{1}{k} \sum_{i=1}^{n} \sum_{k=1}^{k} (x_2 - \mu_2)/\sigma_2^2(1/\sigma_2^2 + (x_2 - \mu_2)/\sigma_2^2), \]
\[ J_{\theta}(Y)_{77} = -\frac{1}{k} \sum_{i=1}^{n} \sum_{k=1}^{k} (1/\sigma_2^2 + (x_2 - \mu_2)/\sigma_2^2)^2. \]

References