Journal of Statistical Research *https://doi.org/10.3329/jsr.v55i2.58810*
2021, Vol. 55, No. 2, pp. 359-375 *ISSN 0256 - 422 X*

APPROXIMATE LIKELIHOOD INFERENCE IN GENERALIZED LINEAR MODELS WITH CENSORED COVARIATES

MAHDI TEIMOURI*
Department of Statistics, Faculty of Science and Engineering
Gonbad Kavous University, Gonbad Kavous, Iran
Email: teimouri@aut.ac.ir

SANJOY K. SINHA
School of Mathematics and Statistics
Carleton University, Ottawa, ON K1S 5B6 Canada
Email: sinha@math.carleton.ca

SUMMARY

In many surveys and clinical trials, we obtain measurements on covariates or biomarkers that are left-censored due to the limit of detection. In such cases, it is necessary to correct for the left-censoring when studying covariate effects in regression models. The expectation-maximization (EM) algorithm is widely used for the likelihood inference in generalized linear models with censored covariates. The EM method, however, requires intensive computation involving high-dimensional integration with respect to the covariates when the dimension of the censored covariates is large. To reduce such computational difficulties, we propose and explore a Monte Carlo EM method based on the Metropolis algorithm. The finite-sample properties of the proposed estimators are studied using Monte Carlo simulations. An application is also provided using actual data obtained from a health and nutrition examination survey.

Keywords and phrases: Expectation-maximization; Generalized linear model; Limit of detection; Maximum likelihood; Monte Carlo method

AMS Classification: MSC 2000: Primary 62F10; secondary 62F35

1 Introduction

In clinical trials and health examination surveys, we often encounter covariates or biomarkers that are left-censored. The left-censoring may occur due to low concentrations of biomarkers for which measuring devices are unable to detect or observe the true values. In other words, we encounter left-censored covariates when their actual values are below some predetermined level, often referred to as the limit of detection (LOD). In such cases, instead of actual measurements, the limit of detection is typically reported as observed values of covariates. For a valid statistical inference, it is often necessary to perform an analysis by correcting for the left-censored covariates. To address

* Corresponding author

the left-censoring in regression models, methods were developed and studied by authors in many different fields, such as biology (Bernhardt et al., 2015), ecology (Thompson and Nelson, 2003), environmental studies (Helsel, 2006), and medicine (Buckley and James, 1979; Sattar et al., 2015).

A straightforward approach to tackle the problem of the limit of detection is to replace the left-censored values by (1/2)LOD or $(1/\sqrt{2})$LOD (Nie et al., 2010) or by weighted average of the uncensored observations (Buckley and James, 1979). This simple and naive method has certain weaknesses, as it generally provides biased estimates (Nie et al., 2010; Lee et al., 2018). Also, substituting the left-censored values with (1/2)LOD or $(1/\sqrt{2})$LOD gives biased estimators with small standard errors (Thompson and Nelson, 2003). There are other techniques available to treat the limit of detection, which include the iterative least squares technique (Schmee and Hahn, 1979), EM algorithm (Aitkin, 1981), regression analysis with random censoring (Ireson and Rao, 1985), M-estimation with censored covariates (Ritov, 1990), multiple imputation with censored covariates (Wei and Tanner, 1991), and polynomial regression with missing covariates (Akritas, 1996). For further details on regression analyses with censored covariates, we refer readers to Lubin et al. (2004), Helsel (2006), Herring (2010), LaFleur et al. (2011), Barescut et al. (2011), May et al. (2011), Sattar et al. (2012), Sattar et al. (2015), Bernhardt et al. (2015), Holstein et al. (2015), and Lee et al. (2018).

To deal with left-censored cocariates, as an efficient tool, the EM method is widely used (May et al., 2011; Sattar et al., 2012, 2015; Bernhardt et al., 2015; Holstein et al., 2015; Lee et al., 2018). The EM method involves calculation of the conditional expectations of the log-likelihood, score function, and Fisher information with respect to the left-censored covarites given the observed data. When the dimension of the censored covariates is large, it may be impractical to calculate the conditional expectations numerically, as they require intensive computation involving high-dimensional integration with respect to the censored covariates. To reduce the computational burden, in this paper we propose a Monte Carlo EM method based on the Metropolis algorithm. The finite-sample properties of the Monte Carlo estimates are assessed based on a simulation study. As an application of the proposed method, we consider analyzing a real data set obtained from the National Health and Nutrition Examination Survey (NHANES), which is designed to assess the health and nutritional status of individuals through interviews and physical examinations.

The paper is organized as follows. Section 2 introduces the model and notation, and describes the proposed approximate EM method for the maximum likelihood estimation in generalized linear models under left-censored covariates. Section 3 presents an illustrative example to describe the computational issues of the proposed method. Section 4 investigates finite-sample properties of the proposed estimators based on a simulation study. Section 5 presents an application using actual health and nutrition examination survey data. Section 6 concludes the paper with some discussion.

2 Model and Notation

2.1 Generalized linear model

Suppose $\mathbf{y} = (y_1, \ldots, y_n)'$ represents a vector of n independent responses, where the ith response y_i follows a distribution in the exponential family

$$f_{y_i|z_i}(y_i|\mathbf{z}_i, \boldsymbol{\beta}, \phi) = \exp\left[\frac{y_i\eta_i - b(\eta_i)}{a(\phi)} + c(y_i, \phi)\right], \tag{2.1}$$

for some functions a, b, and c. The response variable y_i is assumed to be related to the vector of covariates $\mathbf{z}_i$ through the canonical parameter $\eta_i = \mathbf{z}_i'\boldsymbol{\beta}$, where $\mathbf{z}_i$ may contain 1 to incorporate an intercept term. For the exponential family (2.1), the log-likelihood function of the regression coefficients $\boldsymbol{\beta}$ and dispersion parameter ϕ may be obtained as

$$l(\boldsymbol{\beta}, \phi|\mathbf{y}, \mathbf{Z}) = \sum_{i=1}^{n}\left[\frac{y_i\eta_i - b(\eta_i)}{a(\phi)} + c(y_i, \phi)\right], \tag{2.2}$$

where $\mathbf{Z}$ is the design matrix with its ith row being equal to $\mathbf{z}_i'$.

For simplicity, we assume $\phi = 1$, as this is the case for both binary and Poisson regression models. The maximum likelihood estimator of $\boldsymbol{\beta}$ may be obtained by numerically maximizing the log-likelihood function (2.2) with respect to $\boldsymbol{\beta}$ or, equivalently, by solving the estimating equation

$$\sum_{i=1}^{n}\big[y_i - \mu_i(\boldsymbol{\beta}, \mathbf{z}_i)\big]\mathbf{z}_i = \mathbf{0} \tag{2.3}$$

with respect to $\boldsymbol{\beta}$, where $\mu_i(\boldsymbol{\beta}, \mathbf{z}_i)$ is the ith mean response, $\mu_i(\boldsymbol{\beta}, \mathbf{z}_i) = E(y_i|\mathbf{z}_i, \boldsymbol{\beta}) = \partial b(\eta_i)/(\partial\eta_i)$. Equation (2.3) may be solved numerically using a suitable method, such as the iteratively reweighted least squares (IRWLS) method.

2.2 Left-censored covariates

When covariates are left-censored, it is important to estimate the regression parameters by correcting for the censored covariates. Consider a vector of covariates $\mathbf{z}_i$ that may be partitioned as $\mathbf{z}_i = (\mathbf{x}_i, \mathbf{x}_i^*)$, where $\mathbf{x}_i$ represents a set of p continuous covariates that are subject to the limit of detection. Suppose $\mathbf{v}_i = (v_{i1}, \ldots, v_{ip})'$ denotes a vector of p binary variables indicating the censoring statuses of p covariates $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$, where v_{ij} is 1 if the jth covariate x_{ij} is observed (i.e., $x_{ij} \geq l_j$), and 0 if x_{ij} is left-censored (i.e., $x_{ij} < l_j$), with the detection limit l_j being considered known.

Note that the binary indicators $\mathbf{v}_i = (v_{i1}, \ldots, v_{ip})'$ of censored covariates (or biomarkers) may be correlated by nature, as a small value of one covariate may result in a small value of another and hence both covariates could be subject to the limit of detection or left-censoring. To find the joint distribution of $\mathbf{v}_i = (v_{i1}, \ldots, v_{ip})'$, we assume a multivariate Bahadur model (Bahadur, 1961), which is defined by the marginal probabilities $\theta_{ij} = P(v_{ij} = 1) = P(x_{ij} \geq l_j)$ of the binary

indicators and their pairwise and higher order associations. For example, in the case of a three-dimensional covariate vector $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3})'$, the multivariate Bahadur model for the indicator variables $\mathbf{v}_i = (v_{i1}, v_{i2}, v_{i3})'$ is given by

$$f_{v_i}(\mathbf{v}_i|\boldsymbol{\theta}) = \Big[\prod_{j=1}^{3} \theta_{ij}^{v_{ij}}(1-\theta_{ij})^{1-v_{ij}}\Big]\big(1+\gamma_{12}r_{i1}r_{i2}+\gamma_{13}r_{i1}r_{i3}+\gamma_{23}r_{i2}r_{i3}+\gamma_{123}r_{i1}r_{i2}r_{i3}\big),$$

where $r_{ij} = (v_{ij}-\theta_{ij})/\sqrt{\theta_{ij}(1-\theta_{ij})}$, $\gamma_{jk} = \text{corr}(v_{ij}, v_{ik}) = E(r_{ij}r_{ik})$, for $j, k = 1, 2, 3$, and $\gamma_{123} = E(r_{i1}r_{i2}r_{i3})$.

Suppose $\mathbf{x}_i^o$ represents the observed values and $\mathbf{x}_i^l$ the left-censored values of $\mathbf{x}_i$, so that a permutation of these values can be written as $\mathbf{x}_i = (\mathbf{x}_i^o, \mathbf{x}_i^l)$. Assume that $\mathbf{x}_i$ has the multivariate normal density $f_{x_i}(\mathbf{x}_i|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$. Given the observed data $\{(y_i, \mathbf{x}_i^o, \mathbf{x}^*, \mathbf{v}_i), i = 1, \ldots, n\}$, the likelihood of $\boldsymbol{\alpha} = (\boldsymbol{\beta}, \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x, \boldsymbol{\theta})$ may be obtained as

$$L(\boldsymbol{\alpha}) = \prod_{i=1}^{n} \int_{\mathbf{x}_i^l} f_{y_i|z_i}\big(y_i|\mathbf{z}_i, \boldsymbol{\beta}\big)\, f_{x_i}\big(\mathbf{x}_i|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x\big)\, f_{v_i}\big(\mathbf{v}_i|\boldsymbol{\theta}\big)\, d\mathbf{x}_i^l, \tag{2.4}$$

where the limits of integration for the jth left-censored covariate x_{ij}^l in $\mathbf{x}_i^l$ is $(-\infty, l_j)$. The estimators of $\boldsymbol{\alpha}$ may be obtained by maximizing the observed data likelihood function (2.4) with respect to $\boldsymbol{\alpha}$ using a numerical method. Equivalently, we can find the estimators by solving the likelihood score equations $S(\boldsymbol{\alpha}) = \mathbf{0}$ with respect to $\boldsymbol{\alpha}$, where the score function $S(\boldsymbol{\alpha})$ may be obtained as

$$\begin{aligned} S(\boldsymbol{\alpha}) &= \sum_{i=1}^{n} (\partial/\partial\boldsymbol{\alpha}) \log \int_{\mathbf{x}_i^l} f_{y_i|z_i}\big(y_i|\mathbf{z}_i, \boldsymbol{\beta}\big)\, f_{x_i}\big(\mathbf{x}_i|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x\big)\, f_{v_i}\big(\mathbf{v}_i|\boldsymbol{\theta}\big)\, d\mathbf{x}_i^l \\ &= \sum_{i=1}^{n} \int_{\mathbf{x}_i^l} B(\boldsymbol{\alpha}, \mathbf{z}_i)\, f_{x_i^l|\text{obs}_i}\big(\mathbf{x}_i^l|\text{obs}_i, \boldsymbol{\alpha}\big)\, d\mathbf{x}_i^l, \end{aligned} \tag{2.5}$$

with $B(\boldsymbol{\alpha}, \mathbf{z}_i)$ being the score function $B(\boldsymbol{\alpha}, \mathbf{z}_i) = (\partial/\partial\boldsymbol{\alpha}) l(\boldsymbol{\alpha}, \mathbf{z}_i)$ for the complete data log-likelihood

$$l(\boldsymbol{\alpha}, \mathbf{z}_i) = \log f_{y_i|z_i}\big(y_i|\mathbf{z}_i, \boldsymbol{\beta}\big) + \log f_{x_i}\big(\mathbf{x}_i|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x\big) + \log f_{v_i}\big(\mathbf{v}_i|\boldsymbol{\theta}\big). \tag{2.6}$$

The density function $f_{x_i^l|\text{obs}_i}\big(\mathbf{x}_i^l|\text{obs}_i, \boldsymbol{\alpha}\big)$ in Equation (2.5) denotes the conditional distribution of $\mathbf{x}_i^l$ given the observed data $\text{obs}_i = (y_i, \mathbf{x}_i^o, \mathbf{x}^*, \mathbf{v}_i)$ for the ith individual. This density function does not have a closed form, in general, and should be calculated numerically. When the dimension of $\mathbf{x}_i^l$ is large, it is difficult to obtain the exact maximum likelihood estimates, as the estimation involves high-dimensional integration with respect to the left-censored covariates $\mathbf{x}_i^l$. To reduce computational difficulties involving the high-dimensional integration, we propose an approximate maximum likelihood estimation using the Monte Carlo EM method as described in the next section.

2.3 Monte Carlo EM method

Recall the complete data log-likelihood (2.6). Given initial estimates $\boldsymbol{\alpha}^{(m)}$, the E-step of the EM method calculates

$$Q_i\big(\boldsymbol{\alpha}\big|\boldsymbol{\alpha}^{(m)}\big) = E\bigg[\Big\{\log f_{y_i|z_i}\big(y_i|\mathbf{z}_i,\boldsymbol{\beta}\big) + \log f_{x_i}\big(\mathbf{x}_i|\boldsymbol{\mu}_x,\boldsymbol{\Sigma}_x\big) + \log f_{v_i}\big(\mathbf{v}_i|\boldsymbol{\theta}\big)\Big\}\,\Big|\,\text{obs}_i,\boldsymbol{\alpha}^{(m)}\bigg], \quad (2.7)$$

where the expectation E is with respect to the conditional distribution of the left-censored covariates $\mathbf{x}_i^l$ given the observed data obs_i. In particular, if the value of a single covariate x_{i1} for the ith individual is below the detection limit l_1 and hence left-censored, then the E-step of the EM method calculates the integral

$$\begin{aligned} Q_i\big(\boldsymbol{\alpha}\big|\boldsymbol{\alpha}^{(m)}\big) = \int_{x_{i1}} &\Big[\log f_{y_i|z_i}\big(y_i|\mathbf{z}_i,\boldsymbol{\beta}\big) + \log f_{x_i}\big(\mathbf{x}_i|\boldsymbol{\mu}_x,\boldsymbol{\Sigma}_x\big) + \log f_{v_i}\big(\mathbf{v}_i|\boldsymbol{\theta}\big)\Big] \\ &\times\; f_{x_{i1}|\text{obs}_i}\big(x_{i1}|\text{obs}_i,\boldsymbol{\alpha}^{(m)}\big)\, I\big(-\infty < x_{i1} < l_1\big)\, dx_{i1}. \end{aligned}$$

To approximate the conditional expectations in (2.7), we draw random samples from the truncated distribution $f_{x_i^l|\text{obs}_i}\big(\mathbf{x}_i^l|\text{obs}_i,\boldsymbol{\alpha}^{(m)}\big)\, I\big(-\infty < \mathbf{x}_i^l < \mathbf{l}\big)$ using the Metropolis algorithm, where we use the notation $\big(-\infty < \mathbf{x}_i^l < \mathbf{l}\big)$ to indicate that each element of $\mathbf{x}_i^l$ assumes values within its corresponding censoring interval, i.e.,

$$\big(-\infty < \mathbf{x}_i^l < \mathbf{l}\big) \equiv \bigcap_{x_{ij}\in\mathbf{x}_i^l} \big(-\infty < x_{ij} < l_j\big).$$

The Metropolis step begins with a candidate distribution $h_{x_i^l}(\mathbf{x}_i^l)$ from which potential new draws are made. Then an acceptance function is specified in order to determine the probability of accepting a new value as opposed to retaining the previous one. Let $\mathbf{x}_i^l$ be the covariate vector of previous values drawn from the truncated conditional distribution $f_{x_i^l|\text{obs}_i}\big(\mathbf{x}_i^l|\text{obs}_i,\boldsymbol{\alpha}^{(m)}\big)\, I\big(-\infty < \mathbf{x}_i^l < \mathbf{l}\big)$. Consider a new value $\tilde{x}_{ij}^l$ for its jth element as $\tilde{\mathbf{x}}_i^l = \big(x_{i1}^l,\ldots,x_{i,j-1}^l,\tilde{x}_{ij}^l,x_{i,j+1}^l,\ldots,x_{ip_i}^l\big)$ drawn from the candidate distribution $h_{x_i^l}(\mathbf{x}_i^l)$. Then with probability

$$a_j\big(\mathbf{x}_i^l,\tilde{\mathbf{x}}_i^l\big) = \min\left\{1, \frac{f_{x_i^l|\text{obs}_i}\big(\tilde{\mathbf{x}}_i^l|\text{obs}_i,\boldsymbol{\alpha}^{(m)}\big)\, I\big(-\infty < \tilde{\mathbf{x}}_i^l < \mathbf{l}\big)\, h_{x_i^l}\big(\mathbf{x}_i^l\big)}{f_{x_i^l|\text{obs}_i}\big(\mathbf{x}_i^l|\text{obs}_i,\boldsymbol{\alpha}^{(m)}\big)\, I\big(-\infty < \mathbf{x}_i^l < \mathbf{l}\big)\, h_{x_i^l}\big(\tilde{\mathbf{x}}_i^l\big)}\right\}, \quad (2.8)$$

we accept the candidate value $\tilde{x}_{ij}^l$; otherwise, we reject and retain the previous value x_{ij}^l. We repeat the above process to update each element of the covariate vector $\mathbf{x}_i^l$. Note that if the candidate distribution is chosen as $h_{x_i^l}\big(\mathbf{x}_i^l\big) = f_{x_i^l|x_i^o}\big(\mathbf{x}_i^l|\mathbf{x}_i^o,\boldsymbol{\mu}_x,\boldsymbol{\Sigma}_x\big)\, I\big(-\infty < \mathbf{x}_i^l < \mathbf{l}\big)$, then the second term within the braces in (2.8) reduces to

$$\begin{aligned} &\frac{f_{x_i^l|\text{obs}_i}\big(\tilde{\mathbf{x}}_i^l|\text{obs}_i,\boldsymbol{\alpha}^{(m)}\big)\, f_{x_i^l|x_i^o}\big(\mathbf{x}_i^l|\mathbf{x}_i^o,\boldsymbol{\mu}_x,\boldsymbol{\Sigma}_x\big)}{f_{x_i^l|\text{obs}_i}\big(\mathbf{x}_i^l|\text{obs}_i,\boldsymbol{\alpha}^{(m)}\big)\, f_{x_i^l|x_i^o}\big(\tilde{\mathbf{x}}_i^l|\mathbf{x}_i^o,\boldsymbol{\mu}_x,\boldsymbol{\Sigma}_x\big)} \\ &= \frac{f_{y_i|z_i}\big(y_i|\tilde{\mathbf{x}}_i^l,\mathbf{x}_i^o,\mathbf{x}_i^*,\boldsymbol{\beta}\big)\, f_{x_i}\big(\tilde{\mathbf{x}}_i^l,\mathbf{x}_i^o|\boldsymbol{\mu}_x,\boldsymbol{\Sigma}_x\big)\, f_{x_i^l|x_i^o}\big(\mathbf{x}_i^l|\mathbf{x}_i^o,\boldsymbol{\mu}_x,\boldsymbol{\Sigma}_x\big)}{f_{y_i|z_i}\big(y_i|\mathbf{x}_i^l,\mathbf{x}_i^o,\mathbf{x}_i^*,\boldsymbol{\beta}\big)\, f_{x_i}\big(\mathbf{x}_i^l,\mathbf{x}_i^o|\boldsymbol{\mu}_x,\boldsymbol{\Sigma}_x\big)\, f_{x_i^l|x_i^o}\big(\tilde{\mathbf{x}}_i^l|\mathbf{x}_i^o,\boldsymbol{\mu}_x,\boldsymbol{\Sigma}_x\big)} \\ &= \frac{f_{y_i|z_i}\big(y_i|\tilde{\mathbf{x}}_i^l,\mathbf{x}_i^o,\mathbf{x}_i^*,\boldsymbol{\beta}\big)}{f_{y_i|z_i}\big(y_i|\mathbf{x}_i^l,\mathbf{x}_i^o,\mathbf{x}_i^*,\boldsymbol{\beta}\big)}. \end{aligned} \quad (2.9)$$

In this case, the acceptance function $a_j\big(\mathbf{x}_i^l, \tilde{\mathbf{x}}_i^l\big)$ involves only the specification of the conditional distribution of $y_i|\mathbf{z}_i$. The Metropolis step is then incorporated into the EM method to obtain Monte Carlo approximations to the conditional expectations. Suppose $\big(\mathbf{u}_i^{(1)}, \mathbf{u}_i^{(2)}, \ldots, \mathbf{u}_i^{(S_i)}\big)$ is a random sample drawn from the truncated distribution $f_{x_i^l|\text{obs}_i}\big(\mathbf{x}_i^l|\text{obs}_i, \boldsymbol{\alpha}^{(m)}\big)\, I\big(-\infty < \mathbf{x}_i^l < \mathbf{l}\big)$. Then the E-step in (2.7) may be approximated as

$$\begin{aligned} Q_i\big(\boldsymbol{\alpha}|\boldsymbol{\alpha}^{(m)}\big) &\approx \widehat{E}\bigg[\Big\{\log f_{y_i|z_i}\big(y_i|\mathbf{z}_i, \boldsymbol{\beta}\big) + \log f_{x_i}\big(\mathbf{x}_i|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x\big) + \log f_{v_i}\big(\mathbf{v}_i|\boldsymbol{\theta}\big)\Big\} \,\Big|\, \text{obs}_i, \boldsymbol{\alpha}^{(m)}\bigg] \\ &= \frac{1}{S_i}\sum_{s=1}^{S_i}\bigg[\log f_{y_i|z_i}\big(y_i|\mathbf{z}_i^{(s)}, \boldsymbol{\beta}\big) + \log f_{x_i}\big(\mathbf{x}_i^{(s)}|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x\big) + \log f_{v_i}\big(\mathbf{v}_i|\boldsymbol{\theta}\big)\bigg], \quad (2.10) \end{aligned}$$

where $\mathbf{z}_i^{(s)} = \big(\mathbf{x}_i^{(s)}, \mathbf{x}_i^*\big)$ with $\mathbf{x}_i^{(s)} = \big(\mathbf{u}_i^{(s)}, \mathbf{x}_i^o\big)$. The M-step of the EM method maximizes the objective function

$$Q\big(\boldsymbol{\alpha}|\boldsymbol{\alpha}^{(m)}\big) = \sum_{i=1}^{n} Q_i\big(\boldsymbol{\alpha}|\boldsymbol{\alpha}^{(m)}\big), \quad (2.11)$$

with respect to $\boldsymbol{\alpha}$. The complete EM algorithm for estimating the regression parameters $\boldsymbol{\beta}$ and nuisance parameters $\boldsymbol{\tau} = (\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x, \boldsymbol{\theta})$ may be described as follows:

1. Choose initial values $\boldsymbol{\alpha}^{(0)} = (\boldsymbol{\beta}^{(0)}, \boldsymbol{\tau}^{(0)})$. Set $m = 0$.

2. Calculate (with expectations E being replaced by $\widehat{E}$ based on Monte Carlo samples):

   (a) $\boldsymbol{\beta}^{(m+1)}$ that maximises $\sum_{i=1}^{n} \widehat{E}\Big[\log f_{y_i|z_i}(y_i|\mathbf{z}_i, \boldsymbol{\beta}) \,|\, \text{obs}_i, \boldsymbol{\alpha}^{(m)}\Big]$.

   (b) $\boldsymbol{\tau}^{(m+1)}$ that maximises $\sum_{i=1}^{n} \widehat{E}\Big[\big\{\log f_{x_i}(\mathbf{x}_i|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x) + \log f_{v_i}(\mathbf{v}_i|\boldsymbol{\theta})\big\} \,|\, \text{obs}_i, \boldsymbol{\alpha}^{(m)}\Big]$.

   (c) Set $m = m + 1$.

3. If convergence is achieved, declare $\boldsymbol{\alpha}^{(m+1)} = \big(\boldsymbol{\beta}^{(m+1)}, \boldsymbol{\tau}^{(m+1)}\big)$ to be the Monte Carlo EM (MCEM) estimates $\hat{\boldsymbol{\alpha}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\tau}})$.

2.4 Approximate variance of MCEM estimtors

The variance of the proposed MCEM estimators $\hat{\boldsymbol{\alpha}}$ may be approximated by the observed Fisher information. Following Louis (1982), the observed Fisher information matrix may be obtained as

$$\begin{aligned} \mathbf{I}(\boldsymbol{\alpha}) = &-\sum_{i=1}^{n} E\Big[\dot{B}(\boldsymbol{\alpha}, \mathbf{z}_i) \,\Big|\, \text{obs}_i, \boldsymbol{\alpha}\Big] - \sum_{i=1}^{n} E\Big[B(\boldsymbol{\alpha}, \mathbf{z}_i)B'(\boldsymbol{\alpha}, \mathbf{z}_i) \,\Big|\, \text{obs}_i, \boldsymbol{\alpha}\Big] \\ &+ \sum_{i=1}^{n} E\Big[B(\boldsymbol{\alpha}, \mathbf{z}_i) \,\Big|\, \text{obs}_i, \boldsymbol{\alpha}\Big] E\Big[B'(\boldsymbol{\alpha}, \mathbf{z}_i) \,\Big|\, \text{obs}_i, \boldsymbol{\alpha}\Big], \quad (2.12) \end{aligned}$$

where $B(\boldsymbol{\alpha}, \mathbf{z}_i)$ is the complete data score function, $B(\boldsymbol{\alpha}, \mathbf{z}_i) = (\partial/\partial\boldsymbol{\alpha}) l(\boldsymbol{\alpha}, \mathbf{z}_i)$, and $\dot{B}(\boldsymbol{\alpha}, \mathbf{z}_i) = \partial B(\boldsymbol{\alpha}, \mathbf{z}_i)/\partial\boldsymbol{\alpha}'$. A Monte Carlo approximation to the above Fisher information is obtained as

$$\begin{aligned}\mathbf{I}^*(\boldsymbol{\alpha}) = &- \sum_{i=1}^{n} \frac{1}{S_i} \sum_{s=1}^{S_i} \left[\dot{B}\big(\boldsymbol{\alpha}, \mathbf{z}_i^{(s)}\big) \right] - \sum_{i=1}^{n} \frac{1}{S_i} \sum_{s=1}^{S_i} \left[B\big(\boldsymbol{\alpha}, \mathbf{z}_i^{(s)}\big)\, B'\big(\boldsymbol{\alpha}, \mathbf{z}_i^{(s)}\big) \right] \\ &+ \sum_{i=1}^{n} \left[\frac{1}{S_i} \sum_{s=1}^{S_i} B\big(\boldsymbol{\alpha}, \mathbf{z}_i^{(s)}\big) \right] \left[\sum_{i=1}^{n} \frac{1}{S_i} \sum_{s=1}^{S_i} B'\big(\boldsymbol{\alpha}, \mathbf{z}_i^{(s)}\big) \right], \end{aligned} \tag{2.13}$$

where $\mathbf{z}_i^{(s)} = (\mathbf{x}_i^{(s)}, \mathbf{x}_i^*) = (\mathbf{u}_i^{(s)}, \mathbf{x}_i^o, \mathbf{x}_i^*)$ with $\mathbf{u}_i^{(s)}$ $(s = 1, \ldots, S_i)$ being drawn from the truncated conditional distribution $f_{x_i^l|\text{obs}_i}\big(\mathbf{x}_i^l|\text{obs}_i, \hat{\boldsymbol{\alpha}}\big)\, I(-\infty < \mathbf{x}_i^l < \mathbf{l})$. An approximate estimate of the asymptotic variance of $\hat{\boldsymbol{\alpha}}$ is then obtained as $V(\boldsymbol{\alpha}) = \mathbf{I}^*(\hat{\boldsymbol{\alpha}})^{-1}$, where $\mathbf{I}^*(\hat{\boldsymbol{\alpha}})$ is the Monte Carlo Fisher information $\mathbf{I}^*(\boldsymbol{\alpha})$ evaluated at $\hat{\boldsymbol{\alpha}}$.

3 Illustrative Example

Consider a simple binary logistic model with two left-censored covariates x_1 and x_2 in the form

$$\begin{aligned} y_i|(x_{i1}, x_{i2}) &\sim \text{Ind. Bernoulli}(p_i),\ i = 1, \ldots, n, \\ \text{logit}(p_i) &= \mathbf{z}_i'\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}, \\ \mathbf{x}_i &= (x_{i1}, x_{i2})' \sim \text{Ind. } N(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x), \end{aligned} \tag{3.1}$$

where $\mathbf{z}_i = (1, x_{i1}, x_{i2})'$, $\boldsymbol{\mu}_x = (\mu_{x_1}, \mu_{x_2})'$ and $\boldsymbol{\Sigma}_x$ is a 2×2 covariance matrix with the diagonal elements $(\sigma_{x_1}^2, \sigma_{x_2}^2)$ and off-diagonal element $\sigma_{x_1 x_2}$. Assume that the censoring indicators $\mathbf{r}_i = (r_{i1}, r_{i2})'$ follow a multivariate Bahadur model in the form

$$f_{v_i}(\mathbf{v}_i|\boldsymbol{\tau}) = \left[\prod_{j=1}^{2} \theta_{ij}^{v_{ij}} (1 - \theta_{ij})^{1 - v_{ij}} \right] \left[1 + \rho \frac{(v_{i1} - \theta_{i1})(v_{i2} - \theta_{i2})}{\sqrt{\theta_{i1}(1 - \theta_{i1})\theta_{i2}(1 - \theta_{i2})}} \right], \tag{3.2}$$

where $\theta_{ij} = P(v_{ij} = 1) = P(x_{ij} \geq l_j) = 1 - \Phi\big((l_j - \mu_{x_j})/\sigma_{x_j}\big)$ for $j = 1, 2$ with Φ being the standard normal distribution function, and $\rho = \text{corr}(v_{i1}, v_{i2})$. We can estimate the regression parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$ and nuisance parameters $\boldsymbol{\tau} = \big(\mu_{x_1}, \mu_{x_2}, \sigma_{x_1}^2, \sigma_{x_2}^2, \sigma_{x_1 x_2}, \rho\big)'$ using the MCEM algorithm described in Section 2.3. In particular, Step 2(a) of the EM algorithm leads to

the iterative equations for $\boldsymbol{\beta}$:

$$\begin{aligned}\boldsymbol{\beta}^{(m+1)} =& \widehat{E}\left[\sum_{i=1}^{n} w_i(\boldsymbol{\beta},\mathbf{z}_i)\mathbf{z}_i\mathbf{z}_i' \,\Big|\, \text{obs}_i, \boldsymbol{\beta}^{(m)}, \boldsymbol{\tau}^{(m)}\right]^{-1} \\ & \times\ \widehat{E}\left[\sum_{i=1}^{n} w_i(\boldsymbol{\beta},\mathbf{z}_i)d_i(\boldsymbol{\beta},\mathbf{z}_i)\mathbf{z}_i \,\Big|\, \text{obs}_i, \boldsymbol{\beta}^{(m)}, \boldsymbol{\tau}^{(m)}\right] \\ =& \left[\sum_{i=1}^{n}\frac{1}{S_i}\sum_{s=1}^{S_i} w_i\big(\boldsymbol{\beta}^{(m)},\mathbf{z}_i^{(s)}\big)\mathbf{z}_i^{(s)}\mathbf{z}_i'^{(s)}\right]^{-1} \\ & \times\ \left[\sum_{i=1}^{n}\frac{1}{S_i}\sum_{s=1}^{S_i} w_i\big(\boldsymbol{\beta}^{(m)},\mathbf{z}_i^{(s)}\big)d_i\big(\boldsymbol{\beta}^{(m)},\mathbf{z}_i^{(s)}\big)\mathbf{z}_i^{(s)}\right], \end{aligned} \tag{3.3}$$

for $m = 0, 1, 2, \ldots$, where $\mathbf{z}_i^{(s)} = (1, \mathbf{x}_i'^{(s)})'$, with the random observations $\mathbf{x}_i^{(s)}$ $(s = 1, \ldots, S_i)$ being drawn from the truncated conditional distribution $f_{x_i^l|\text{obs}_i}\big(\mathbf{x}_i^l|\text{obs}_i, \boldsymbol{\beta}^{(m)}, \boldsymbol{\tau}^{(m)}\big)\, I(-\infty < \mathbf{x}_i^l < \mathbf{l})$. Also, $w_i(\boldsymbol{\beta}, \mathbf{z}_i)$ is a weight function given by

$$w_i(\boldsymbol{\beta},\mathbf{z}_i) = \text{var}(y_i) = p_i(1-p_i) = \frac{\exp(\mathbf{z}_i'\boldsymbol{\beta})}{\big(1+\exp\big(\mathbf{z}_i'\boldsymbol{\beta}\big)\big)^2}, \tag{3.4}$$

and $d_i(\boldsymbol{\beta}, \mathbf{z}_i)$ is a "pseudo-observation" given by

$$d_i(\boldsymbol{\beta},\mathbf{z}_i) = \mathbf{z}_i'\boldsymbol{\beta} + \frac{(y_i - p_i)}{p_i(1-p_i)}. \tag{3.5}$$

In Step 2(b) of the MCEM algorithm, the nuisance parameters $\boldsymbol{\tau} = (\mu_{x_1}, \mu_{x_2}, \sigma_{x_1}^2, \sigma_{x_2}^2, \sigma_{x_1x_2}, \rho)'$ are estimated by numerically maximizing the approximate expectation

$$\begin{aligned}&\sum_{i=1}^{n}\widehat{E}\left[\Big\{\log f_{x_i}(\mathbf{x}_i|\boldsymbol{\mu}_x,\boldsymbol{\Sigma}_x) + \log f_{v_i}(\mathbf{v}_i)\Big\} \,\Big|\, \text{obs}_i, \boldsymbol{\beta}^{(m)}, \boldsymbol{\tau}^{(m)}\right] \\ &= \sum_{i=1}^{n}\frac{1}{S_i}\sum_{s=1}^{S_i}\log f_{x_i}\big(\mathbf{x}_i^{(s)}|\boldsymbol{\mu}_x,\boldsymbol{\Sigma}_x\big) + \sum_{i=1}^{n}\log f_{v_i}(\mathbf{v}_i|\boldsymbol{\tau}). \end{aligned} \tag{3.6}$$

For good approximations to the above conditional expectations, the number of Monte Carlo samples drawn by the Metropolis algorithm should be reasonably large. In the next section, we carry out a simulation study based on a set of $S_i = 1000$ Monte Carlo samples at each iteration.

Remark: Note that in the case of a continuous response y_i, we can use a linear model in the form $E(y_i|\mathbf{x}_i) = \mathbf{x}_i'\boldsymbol{\beta}$. In this setting, the iterative equation (3.3) for estimating the regression parameters $\boldsymbol{\beta}$ takes a simplified form where the pseudo observation $d_i(\boldsymbol{\beta}, \mathbf{z}_i)$ becomes the original response $d_i(\boldsymbol{\beta}, \mathbf{z}_i) = y_i$ and the weight function becomes $w_i(\boldsymbol{\beta}, \mathbf{z}_i) = 1$. As before, the nuisance parameters $\boldsymbol{\tau}$ are estimated by numerically maximizing Equation (3.6).

4 Simulation Study

To assess the performance of the proposed Monte Carlo EM (MCEM) method, we ran a series of simulations using the binary logistic regression model (3.1). The data were generated using two combinations of regression coefficients, $(\beta_0, \beta_1, \beta_2) = (-2, 0.5, 1)$ and $(-2, 1, 0.5)$, and a fixed set of nuisance parameters, $(\mu_{x_1}, \mu_{x_2}, \sigma^2_{x_1}, \sigma^2_{x_2}, \sigma_{x_1x_2}) = (0, 2, 1, 2, 0.5)$. The covariates x_1 and x_2 were subject to the limit of detection, where we considered two combinations, 30% and 50% LODs, of left-censored values for both covariates.

Also, the data were generated using two combinations of sample sizes, $n = 100$ and 200. Each simulation run was based on 500 replicates of data sets. The parameters were estimated using the proposed MCEM method described earlier. The number of Monte Carlo samples used in the MCEM method was fixed at $S_i = 1000$. Figure 1 exhibits the convergence of the MCEM estimates $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ of regression coefficients obtained from a representative sample of size $n = 200$. The plots indicate that when the number of Monte Carlo samples as used in the MCEM method is reasonably large, i.e., $S_i \geq 2000$, the stochastic estimates can approximate the exact maximum likelihood estimates with a good degree of accuracy.

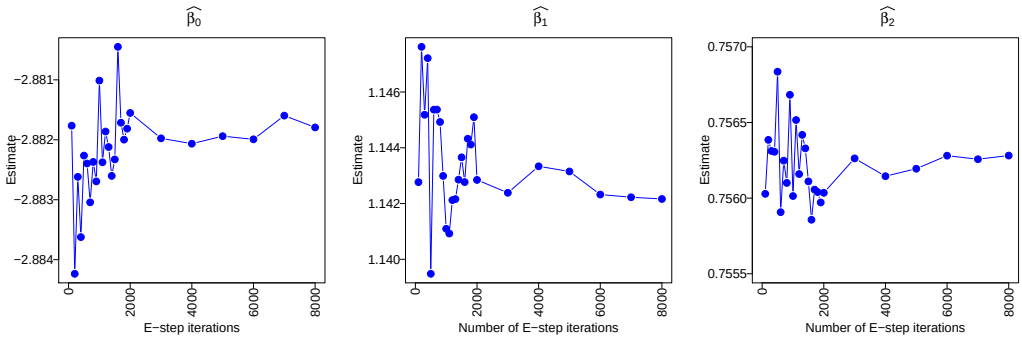

Figure 1: Convergence of estimates $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ obtained by the iterative Monte Carlo EM algorithm. Estimates are shown for a representative sample of size $n = 200$ drawn from the binary logistic model (3.1) with regression parameters $(\beta_0, \beta_1, \beta_2) = (-2, 1, 0.5)$ and nuisance parameters $(\mu_{x_1}, \mu_{x_2}, \sigma^2_{x_1}, \sigma^2_{x_2}, \sigma_{x_1x_2}) = (0, 2, 1, 2, 0.5)$. Covariates (x_1, x_2) are subject to detection limits with 50% left-censored values.

As we consider two covariates x_1 and x_2 in the logistic regression model (3.1), it is not so difficult to obtain the exact maximum likelihood estimates by numerically maximizing the observed data likelihood function. So here we assess how the proposed MCEM estimates compare to their exact counterparts. Boxplots of the exact and approximate (MCEM) estimates of $(\beta_0, \beta_1, \beta_2)$ constructed based on 500 replicates of data sets are shown in Figures 2 and 3 for two combinations of sample sizes, $n = 100$ and $n = 200$, respectively, for true parameters $(\beta_0, \beta_1, \beta_2) = (-2, 0.5, 1)$ and $(\mu_{x_1}, \mu_{x_2}, \sigma^2_{x_1}, \sigma^2_{x_2}, \sigma_{x_1x_2}) = (0, 2, 1, 2, 0.5)$. Figures 4 and 5 repeat the plots for the regression coefficients $(\beta_0, \beta_1, \beta_2) = (-2, 1, 0.5)$. It is clear from the boxplots that the Monte Carlo estimates are very similar to their exact counterparts, as expected. For both methods, the estimates are roughly unbiased and symmetric about their corresponding true parameter values.

Tables 1 and 2 supplement the boxplots in Figures 2–5 by showing the empirical biases and root

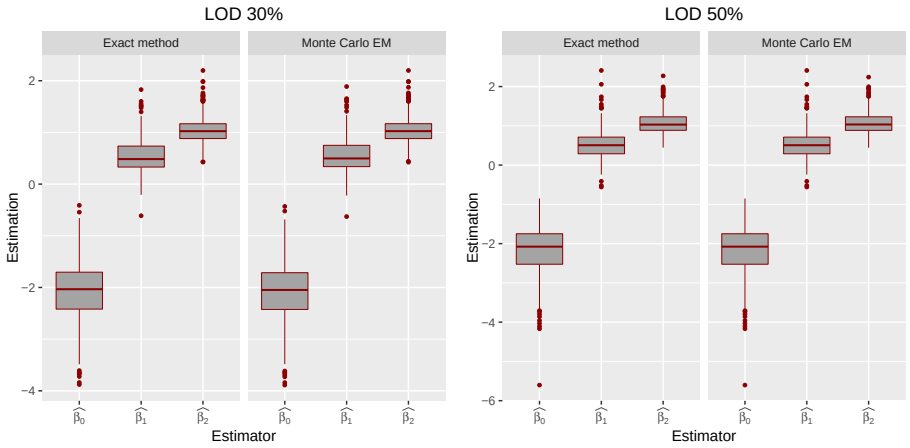

Figure 2: Boxplots of exact and approximate (MCEM) estimates of $(\beta_0, \beta_1, \beta_2)$ constructed based on 500 replicates of data sets each with sample size $n = 100$. Regression parameters fixed at $(\beta_0, \beta_1, \beta_2) = (-2, 0.5, 1)$; nuisance parameters at $(\mu_{x_1}, \mu_{x_2}, \sigma^2_{x_1}, \sigma^2_{x_2}, \sigma_{x_1x_2}) = (0, 2, 1, 2, 0.5)$.

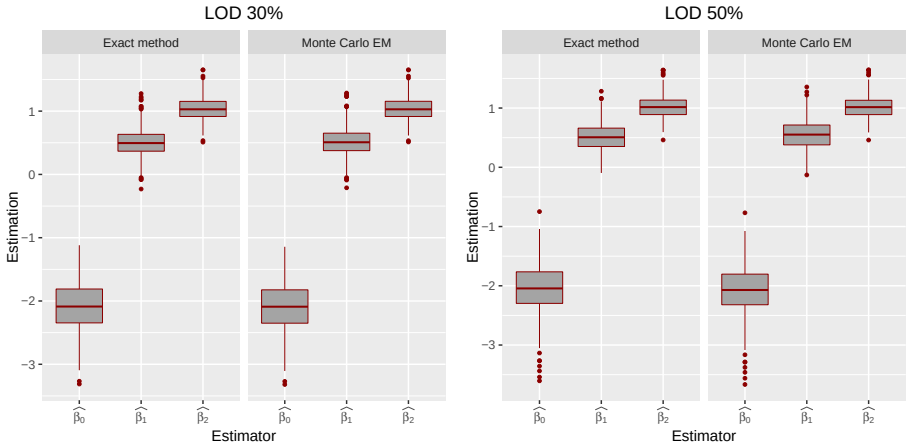

Figure 3: Boxplots of exact and approximate (MCEM) estimates of $(\beta_0, \beta_1, \beta_2)$ constructed based on 500 replicates of data sets each with sample size $n = 200$. Regression parameters fixed at $(\beta_0, \beta_1, \beta_2) = (-2, 0.5, 1)$; nuisance parameters at $(\mu_{x_1}, \mu_{x_2}, \sigma^2_{x_1}, \sigma^2_{x_2}, \sigma_{x_1x_2}) = (0, 2, 1, 2, 0.5)$.

of mean squared errors of the exact and stochastic MCEM estimates of the regression coefficients. Although both methods provide roughly unbiased estimates, the MCEM method generally provides more variability in the estimates. For example, when estimating β_1 with sample size $n = 200$, with 50% LOD, and $(\beta_0, \beta_1, \beta_2) = (-2, 0.5, 1)$, Table 1 shows that the MCEM method provides a root mean squared error of 0.3478, whereas the exact method provides a slightly smaller root mean squared error of 0.3188. Similarly, when estimating β_1 with sample size $n = 200$, with 50%

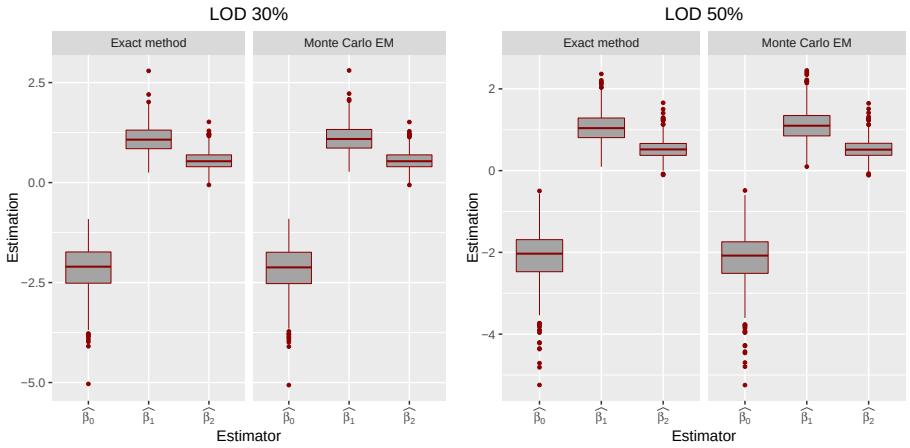

Figure 4: Boxplots of exact and approximate (MCEM) estimates of $(\beta_0, \beta_1, \beta_2)$ constructed based on 500 replicates of data sets each with sample size $n = 100$. Regression parameters fixed at $(\beta_0, \beta_1, \beta_2) = (-2, 1, 0.5)$; nuisance parameters at $(\mu_{x_1}, \mu_{x_2}, \sigma^2_{x_1}, \sigma^2_{x_2}, \sigma_{x_1x_2}) = (0, 2, 1, 2, 0.5)$.

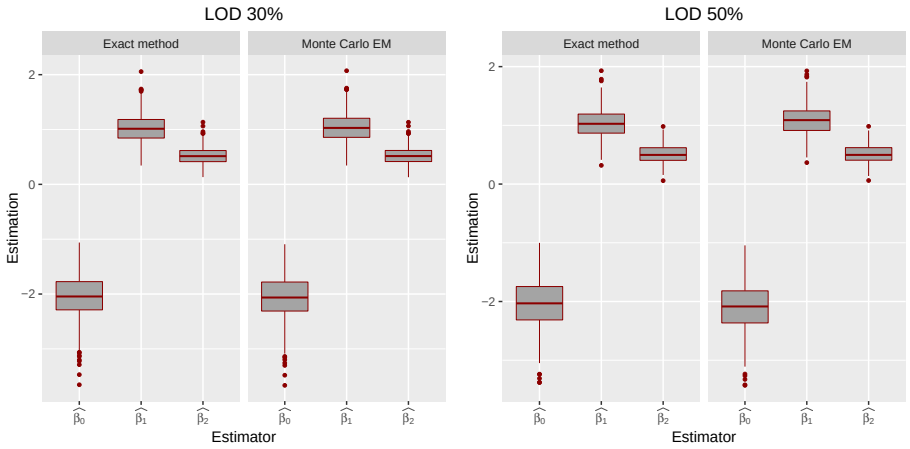

Figure 5: Boxplots of exact and approximate (MCEM) estimates of $(\beta_0, \beta_1, \beta_2)$ constructed based on 500 replicates of data sets each with sample size $n = 200$. Regression parameters fixed at $(\beta_0, \beta_1, \beta_2) = (-2, 1, 0.5)$; nuisance parameters at $(\mu_{x_1}, \mu_{x_2}, \sigma^2_{x_1}, \sigma^2_{x_2}, \sigma_{x_1x_2}) = (0, 2, 1, 2, 0.5)$.

LOD, and $(\beta_0, \beta_1, \beta_2) = (-2, 1, 0.5)$, Table 2 shows that the MCEM method provides a root mean squared error of 0.3451, whereas the exact method provides a similar root mean squared error of 0.3412. The difference between estimates obtained under the two methods appeared to be small when the proportion of left-censored values in covariates is small.

Table 1: Biases and root of mean squared errors (RMSEs) of $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ constructed based on exact and Monte Carlo EM methods using sample sizes $n = 100$ and $n = 200$. Each simulation run was based on 500 replicates of data sets. Regression parameters fixed at $(\beta_0, \beta_1, \beta_2) = (-2, 0.5, 1)$; nuisance parameters at $(\mu_{x_1}, \mu_{x_2}, \sigma^2_{x_1}, \sigma^2_{x_2}, \sigma_{x_1x_2}) = (0, 2, 1, 2, 0.5)$.

		$n = 100$				$n = 200$			
		MCEM		Exact		MCEM		Exact	
LOD%	Estimator	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
30	$\hat{\beta}_0$	–0.1976	0.8860	–0.1647	0.8827	–0.0866	0.7816	–0.0772	0.7800
	$\hat{\beta}_1$	0.0678	0.5507	0.0239	0.4931	0.0490	0.4795	0.0341	0.4616
	$\hat{\beta}_2$	0.0704	0.3923	0.0723	0.3917	0.0412	0.3621	0.0413	0.3619
50	$\hat{\beta}_0$	–0.0856	0.5918	–0.0531	0.5881	–0.0901	0.5714	–0.0802	0.5704
	$\hat{\beta}_1$	0.0515	0.3478	0.0136	0.3213	0.0247	0.3318	0.0087	0.3188
	$\hat{\beta}_2$	0.0256	0.2647	0.0248	0.2644	0.0379	0.2587	0.0376	0.2585

Table 2: Biases and root of mean squared errors (RMSEs) of $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ constructed based on exact and Monte Carlo EM methods using sample sizes $n = 100$ and $n = 200$. Each simulation run was based on 500 replicates of data sets. Regression parameters fixed at $(\beta_0, \beta_1, \beta_2) = (-2, 1, 0.5)$; nuisance parameters at $(\mu_{x_1}, \mu_{x_2}, \sigma^2_{x_1}, \sigma^2_{x_2}, \sigma_{x_1x_2}) = (0, 2, 1, 2, 0.5)$.

		$n = 100$				$n = 200$			
		MCEM		Exact		MCEM		Exact	
LOD%	Estimator	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
30	$\hat{\beta}_0$	–0.1789	0.9458	–0.1640	0.9393	–0.1665	0.8831	–0.1111	0.8805
	$\hat{\beta}_1$	0.1123	0.5726	0.0938	0.5334	0.1209	0.5013	0.0609	0.4925
	$\hat{\beta}_2$	0.0487	0.3481	0.0482	0.3472	0.0364	0.3398	0.0364	0.3393
50	$\hat{\beta}_0$	–0.0773	0.5935	–0.0625	0.5866	–0.1069	0.5792	–0.0510	0.5768
	$\hat{\beta}_1$	0.0388	0.3728	0.0206	0.3452	0.0938	0.3451	0.0344	0.3412
	$\hat{\beta}_2$	0.0231	0.2182	0.0228	0.2182	0.0083	0.2171	0.0086	0.2169

5 Application

In this section, we consider analyzing some actual health survey data from the National Health and Nutrition Examination Survey (NHANES) using the proposed MCEM method. The NHANES study is designed to assess the health and nutritional status of adults and children in the United States through interviews and physical examinations. It is a major program of the National Center for Health Statistics (NCHS) under the Centers for Disease Control and Prevention (CDC).

The NHANES study collects data from participating respondents on demographic, socioeconomic, dietary, and health-related variables through interviews. Also, physical examinations are performed to obtain measurements on medical, dental, and physiological characteristics, as well as laboratory tests. Here we investigate the cardiovascular (CV) fitness level of respondents aged 12–49 years. In this CV fitness study, the screening of the participants is done prior to a treadmill test, where individuals are excluded from the study depending upon their health conditions. We investigate the relationship between the cardiovascular fitness and other health conditions as well as risk factors, and thereby assess persons at risk due to poor physical fitness.

Our study is based on the NHANES data collected during the period 2003–2004. The data contained observations from 1230 individuals. The response variable of interest is the CV fitness with three levels categorized based on gender-age specific cut-points of maximal oxygen consumption (VO2max), which is determined by measuring the heart rates response to known levels of submaximal work. In our study, the response variable y represents the CV fitness, which is dichotomized into 0 for "low" CV fitness level and 1 for "moderate" to "high" CV fitness level. The goal is to identify factors that affect the cardiovascular fitness of the individuals under study.

For our analysis, we use a set of dichotomized covariates, which include *age* (0 for age ≤ 35 years, and 1, otherwise), physical activity readiness code *parc* (0 for participating regularly in recreation or work requiring little or no physical activity, and 1 for modest or heavy physical activity), and ratio of family income to poverty *pir* (0 for ratio < 5, and 1, otherwise). We also consider two continuous biomarkers, *cot* (serum cotinine level, ng/mL) and *crp* (C-reactive protein level, mg/dL). These biomarkers showed large variability in the measurements, as well as left-censored values due to the limit of detection. The detection limits for *cot* and *crp* were 0.011 and 0.01, respectively. To reduce the variability, we took the logarithm of the measurements on the two biomarkers, denoted by $lcot = \log(cot)$ and $lcrp = \log(crp)$. To adjust for the left-censoring, we considered analyzing the data using the proposed Monte Carlo EM (MCEM) method. The number of Monte Carlo samples used in the MCEM method was fixed at $S_i = 2000$.

We use a logistic regression model to describe the mean response $E(y)$ as a function of the covariates, given by

$$\text{logit}\{E(y)\} = \beta_0 + \beta_1 age + \beta_2 parc + \beta_3 pir + \beta_4 lcot + \beta_5 lcrp. \tag{5.1}$$

Both exact ML and MCEM estimates of the regression coefficients are presented in Table 3. The MCEM estimates appear to be generally close to their exact counterparts, as expected. From Table 3, it appears that the CV fitness is strongly associated with the two biomarkers, serum cotinine (*cot*) and C-reactive protein (*crp*), and all demographic variables considered. For example, older individuals (*age* > 35), individuals with increased physical activity (*parc* $= 1$), or with higher ratio of family income to poverty (*pir* ≥ 5) have a higher CV fitness level. Also, individuals with a higher log-serum cotinine level (*lcot*) tend to have a higher CV fitness level. On the other hand, individuals with a higher log-C-reactive protein level (*lcrp*) tend to have a lower CV fitness level. In particular,

given other variables fixed, for individuals participating regularly in recreation or work requiring modest or heavy physical activity (*parc* $= 1$), the odds of having a "moderate" to "high" CV fitness level increases by 1.39 $(= \exp(0.3283))$, as compared to individuals requiring little or no physical activity (*parc* $= 0$). Also, given other variables fixed, for every unit increase in log-C-reactive protein level (*lcrp*), the odds of having a "moderate" to "high" CV fitness level decreases by 0.89 $(= \exp(-0.1137))$, as obtained by the exact method.

Figure 6 displays the direction of the Monte Carlo EM algorithm, which shows that a convergence in estimation is achieved at the 30th iteration, where the initial values are shown at iteration zero. When comparing the estimation times by the two methods, the proposed MCEM method appeared to be faster than the exact maximum likelihood method for estimating the model parameters. Roughly, with two covariates as considered here, the MCEM method required 25% less time for the estimation, as compared to the exact likelihood method. For high-dimensional covariates, however, the gain in computation time from the MCEM method would be much higher.

It is important to note that under correctly specified models, the exact maximum likelihood estimators are, in fact, the most efficient. The purpose of this paper is to address the computational issues of the likelihood estimation involving multidimensional integration. Specifically, we discuss a Monte Carlo approach to approximating the maximum likelihood estimators by generating random draws from the conditional distribution of the 'missing' data given the observed data. As expected, the estimated standard errors of the proposed MCEM estimators are slightly higher than those obtained from the exact likelihood estimators. The accuracy of the standard errors of the MCEM estimators can be improved by increasing the number of bootstrap samples used by the Monte Carlo method.

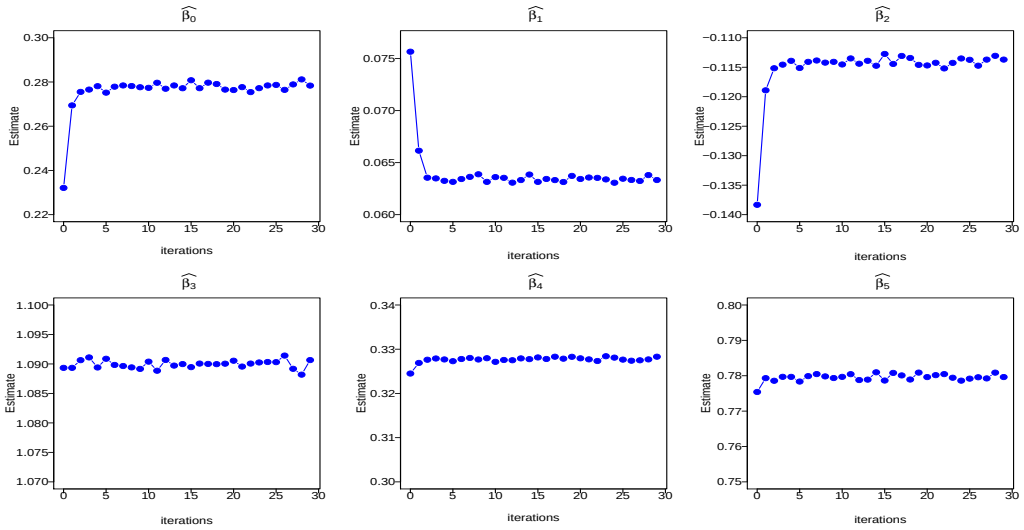

Figure 6: Convergence of estimates of regression coefficients for CV fitness data from NHANES study based on Monte Carlo EM method.

Table 3: Analysis of cardiovascular fitness data from NHANES study based on both exact and Monte Carlo EM (MCEM) methods.

	Exact method				MCEM method			
			95% Interval				95% Interval	
Coef	Estimate	SE	Low	Up	Estimate	SE	Low	Up
β_0	0.2783	0.0826	0.1164	0.4401	0.2321	0.1317	–0.0260	0.4902
age	1.0906	0.1334	0.8291	1.3521	1.0893	0.1671	0.7617	1.4168
parc	0.3283	0.0855	0.1607	0.4958	0.3245	0.1318	0.0661	0.5828
pir	0.7796	0.1260	0.5326	1.0265	0.7754	0.1893	0.4043	1.1464
lcot	0.0633	0.0126	0.0386	0.0879	0.0756	0.0203	0.0358	0.1153
lcrp	–0.1137	0.0535	–0.2185	–0.0088	–0.1383	0.0601	–0.2560	–0.0205

6 Discussion

The purpose of this paper was to provide a suitable method for analyzing generalized linear models when covariates are left-censored due to the limit of detection. We have developed and explored a Monte Carlo EM (MCEM) method to approximate the maximum likelihood estimates of the regression parameters, as well as other nuisance parameters. The proposed MCEM method provides approximate estimates that are generally close to their exact counterparts. It is also worth noting that the proposed stochastic Monte Carlo estimates reach the neighborhood of their exact counterparts very quickly, but they continue to show random variation. In fact, the number of Monte Carlo samples required to have the stochastic estimates converge with three- or four-decimal accuracy would be very large, as pointed out by McCulloch (1997).

In this paper, we have focused on developing the Monte Carlo approach in the context of generalized linear models, which include the commonly used binary and Poisson regression models. The proposed method may be extended to generalized linear mixed models for analyzing clustered or longitudinal data, where repeated outcomes from a given cluster or longitudinal outcomes from a given individual may be correlated by nature. Mixed models are commonly used to describe correlation structures among repeated outcomes. A full likelihood analysis of the generalized linear mixed model usually requires intensive calculations involving high-dimensional integrals. We intend to investigate the proposed method further for analyzing generalized linear mixed models for correlated outcomes with left-censored covariates in a future study.

Acknowledgements

The research of Sanjoy Sinha is partially supported by a grant from the Natural Sciences and Engineering Research Council of Canada. This research was conducted while Mahdi Teimouri was a postdoctoral fellow in the School of Mathematics and Statistics at Carleton University. The fund for Mahdi Teimouri was provided by the Fields Postdoctoral Fellowship. He would like to express

his sincere thank to Carleton University for providing excellent working conditions to complete the research.

References

Aitkin, M. (1981), "A note on the regression analysis of censored data," *Technometrics*, 23, 161–163.

Akritas, M. G. (1996), "On the use of nonparametric regression techniques for fitting parametric regression models," *Biometrics*, 52, 1342–1362.

Bahadur, R. R. (1961), "A representation of the joint distribution of responses to n dichotomous items," *In Studies in Item Analysis and Prediction, Stanford University Press*, 158–168.

Barescut, J., Lariviere, D., Stocki, T., Wood, M., Beresford, N., and Copplestone, D. (2011), "Limit of detection values in data analysis: Do they matter?" *Radioprotection*, 46, S85–S90.

Bernhardt, P. W., Wang, H. J., and Zhang, D. (2015), "Statistical methods for generalized linear models with covariates subject to detection limits," *Statistics in Biosciences*, 7, 68–89.

Buckley, J. and James, I. (1979), "Linear regression with censored data," *Biometrika*, 66, 429–436.

Helsel, D. R. (2006), "Fabricating data: how substituting values for nondetects can ruin results, and what can be done about it," *Chemosphere*, 65, 2434–2439.

Herring, A. H. (2010), "Nonparametric Bayes shrinkage for assessing exposures to mixtures subject to limits of detection," *Epidemiology (Cambridge, Mass.)*, 21, S71.

Holstein, C. A., Griffin, M., Hong, J., and Sampson, P. D. (2015), "Statistical method for determining and comparing limits of detection of bioassays," *Analytical chemistry*, 87, 9795–9801.

Ireson, M. and Rao, P. (1985), "Interval estimation of slope with right-censored data," *Biometrika*, 72, 601–608.

LaFleur, B., Lee, W., Billhiemer, D., Lockhart, C., Liu, J., and Merchant, N. (2011), "Statistical methods for assays with limits of detection: Serum bile acid as a differentiator between patients with normal colons, adenomas, and colorectal cancer," *Journal of Carcinogenesis*, 10, 12.

Lee, W.-C., Sinha, S. K., Arbuckle, T. E., and Fisher, M. (2018), "Estimation in generalized linear models under censored covariates with an application to MIREC data," *Statistics in Medicine*, 37, 4539–4556.

Louis, T. A. (1982), "Finding the observed information matrix when using the EM algorithm," *Journal of the Royal Statistical Society: Series B*, 44, 226–233.

Lubin, J. H., Colt, J. S., Camann, D., Davis, S., Cerhan, J. R., Severson, R. K., Bernstein, L., and Hartge, P. (2004), "Epidemiologic evaluation of measurement data in the presence of detection limits," *Environmental Health Perspectives*, 112, 1691–1696.

May, R. C., Ibrahim, J. G., and Chu, H. (2011), "Maximum likelihood estimation in generalized linear models with multiple covariates subject to detection limits," *Statistics in Medicine*, 30, 2551–2561.

McCulloch, C. E. (1997), "Maximum Likelihood Algorithms for Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 92, 162–170.

Nie, L., Chu, H., Liu, C., Cole, S. R., Vexler, A., and Schisterman, E. F. (2010), "Linear regression with an independent variable subject to a detection limit," *Epidemiology (Cambridge, Mass.)*, 21, S17.

Ritov, Y. (1990), "Estimation in a linear regression model with censored data," *The Annals of Statistics*, 18, 303–328.

Sattar, A., Sinha, S. K., and Morris, N. J. (2012), "A parametric survival model when a covariate is subject to left-censoring," *Journal of Biometrics & Biostatistics*, S3, 1–6.

Sattar, A., Sinha, S. K., Wang, X.-F., and Li, Y. (2015), "Frailty models for pneumonia to death with a left-censored covariate," *Statistics in Medicine*, 34, 2266–2280.

Schmee, J. and Hahn, G. J. (1979), "A simple method for regression analysis with censored data," *Technometrics*, 21, 417–432.

Thompson, M. L. and Nelson, K. P. (2003), "Linear regression with Type I interval-and left-censored response data," *Environmental and Ecological Statistics*, 10, 221–230.

Wei, G. C. and Tanner, M. A. (1991), "Applications of multiple imputation to the analysis of censored regression data," *Biometrics*, 47, 1297–1309.

Received: March 3, 2021

Accepted: November 28, 2021