Published in final edited form as:

Stat Biosci. 2015 May ; 7(1): 68–89. doi:10.1007/s12561-013-9099-4.

Statistical Methods for Generalized Linear Models with Covariates Subject to Detection Limits

Paul W. Bernhardt,

Department of Mathematics and Statistics, Villanova University, Villanova, USA

Huixia J. Wang, and

Department of Statistics, North Carolina State University, Raleigh, USA

Daowen Zhang

Department of Statistics, North Carolina State University, Raleigh, USA

Paul W. Bernhardt: paul.bernhardt@villanova.edu; Huixia J. Wang: hwang3@ncsu.edu; Daowen Zhang: dzhang2@stat.ncsu.edu

Abstract

Censored observations are a common occurrence in biomedical data sets. Although a large amount of research has been devoted to estimation and inference for data with censored responses, very little research has focused on proper statistical procedures when predictors are censored. In this paper, we consider statistical methods for dealing with multiple predictors subject to detection limits within the context of generalized linear models. We investigate and adapt several conventional methods and develop a new multiple imputation approach for analyzing data sets with predictors censored due to detection limits. We establish the consistency and asymptotic normality of the proposed multiple imputation estimator and suggest a computationally simple and consistent variance estimator. We also demonstrate that the conditional mean imputation method often leads to inconsistent estimates in generalized linear models, while several other methods are either computationally intensive or lead to parameter estimates that are biased or more variable compared to the proposed multiple imputation estimator. In an extensive simulation study, we assess the bias and variability of different approaches within the context of a logistic regression model and compare variance estimation methods for the proposed multiple imputation estimator. Lastly, we apply several methods to analyze the data set from a recently-conducted GenIMS study.

Keywords

Censored predictor; Complete case; Conditional mean imputation; Detection limit; Improper multiple imputation

Correspondence to: Huixia J. Wang, hwang3@ncsu.edu.

Electronic supplementary material: The online version of this article (doi:10.1007/s12561-013-9099-4) contains supplementary material, which is available to authorized users.

Supplementary Material: The reader is referred to the on-line Supplementary Material for technical appendices and to http://www4.stat.ncsu.edu/∼wang/software.html for simulation code, application code, and data similar to that analyzed in the application.

1 Introduction

Biomedical data sets frequently contain variables that are subject to censoring. Although censoring is commonly associated with time-to-event data, censored data also arise due to detection limits (DLs). Two well-known examples of left-censoring due to DLs include viral RNA measurements in individuals with the human immunodeficiency virus [12, 20, 27, 35] and antibody concentration in response to vaccines [24].

While methods have been developed for analyzing missing covariate data [8, 9, 13, 14, 17, 25], these methods do not apply in general to censored data. Additionally, research for censored data due to DLs has mainly focused on censored responses ([20, 28, 35, 39, 42], among others). Our main objective in this paper is to develop statistical inference methods in the context of generalized linear models (GLMs) with covariates censored due to DLs.

Traditionally, three common approaches have been used to handle censored covariate data. First, a complete case analysis simply discards data from individuals with censored covariates. We show in Sect. 2.2 that the complete case method leads to consistent estimates in GLMs when covariates are censored due to DLs, but efficiency is lost due to the discarding of data. A second common approach replaces censored observations with a fixed value such as the DL, DL/2, or DL/√2 [11]. Using these naive substitution methods leads to biased parameter estimates and incorrect inference [2, 7, 19, 22, 30, 31]. Another conventional approach is conditional mean imputation, in which censored values are replaced with their conditional expectations given the other observed variables [1, 3, 6, 22, 26]. Although conditional mean imputation is more sophisticated than the substitution approach, we prove that this method generally leads to biased estimates; see Sect. 2.3 for a more detailed discussion.

Several authors have considered maximum likelihood methods for dealing with covariates censored due to DLs; see for instance [21, 22], and [2, 3, 29], and [6]. Recently, May et al. [23] developed a Monte Carlo EM algorithm for maximizing the likelihood of a GLM with multiple covariates subject to DLs. Although maximum likelihood methods lead to consistent and efficient parameter estimates, computation can be very intensive since representations of the censored-data likelihood involve integrations and generally do not have closed forms. To avoid this computational issue and weaken the required distributional assumptions, Tsimikas et al. [37] proposed an estimating equation approach, though the method is only described for a single censored covariate.

As an alternative to maximum likelihood methods, a few authors have looked at multiple imputation approaches. Lyles et al. [21] used multiple imputation to estimate the parameters of a bivariate normal distribution when one of the variables is censored, and Lynn [22] considered several ad hoc multiple imputation methods for a logistic regression model with a single censored covariate. Taking a different approach, Lubin et al. [19] developed a multiple imputation algorithm for models with a single censored covariate which uses bootstrapping to account for uncertainty in initial parameter estimates. Lee et al. [16] proposed a multiple imputation algorithm that allows for multiple censored covariates, though they assumed a multivariate normal model for generating imputations. Wang and

Feng [38] developed a semiparametric multiple imputation procedure for linear regression models, but the method requires the conditional quantiles of the censored covariates given the response and other un-censored covariates to be linear and the method can only perform partial imputation under heavy censoring.

The methods mentioned above are computationally intensive, designed for a limited number of covariates, or suffer from inefficiency or estimation bias. Our work advances the field in three major ways. First, we describe and compare several methods for handling multiple covariates subject to DLs in the context of GLMs. Most of these methods were not previously proposed for this problem or are extensions to the case of multiple covariates. Second, we study the theoretical validity of the conventional complete case and mean imputation methods. To our knowledge, this is the first time such properties are studied theoretically in context of GLMs with multiple covariates subject to DLs. Third, and most substantially, we develop a new and computationally efficient "improper" multiple imputation algorithm which overcomes many of the deficiencies present in other methods. We establish the asymptotic properties of the improper multiple imputation estimator and propose a convenient variance estimation method. Even though improper multiple imputation has been studied for missing data, our development for censored covariates is nontrivial. The special feature of censoring due to DLs makes it technically challenging to incorporate the censoring mechanism in a practical way to achieve consistency and efficiency improvement. Through numerical studies, we demonstrate that the proposed improper multiple imputation estimator performs comparably to maximum likelihood, fully Bayesian, and "proper" multiple imputation methods while requiring much less computational effort.

The remainder of the paper is organized as follows. In Sect. 2, we investigate several conventional approaches for dealing with censored covariates within the context of GLMs. In Sect. 3, we present our proposed improper multiple imputation method, as well as a consistent variance estimator for the parameter estimates. In Sect. 4, we carry out an extensive simulation study to compare the performance of the proposed methods with various existing approaches. In Sect. 5, we apply the proposed methods to the data set from the GenIMS study. Finally, in Sect. 6, we discuss the limitations of the improper multiple imputation method and some avenues for further research. All the technical details are provided in the Supplementary Material.

2 Model Setup and Conventional Analysis Methods

2.1 Generalized Linear Model

Suppose we observe independent samples $(y_i, \boldsymbol{w}_i)$, $i = 1,\ldots,n$, where y_i is a univariate response and $\boldsymbol{w}_i$ is a p-dimensional vector of covariates. We assume that the distribution of y_i given $\boldsymbol{w}_i$ belongs to the exponential dispersion family,

$$f(y_i|\boldsymbol{w}_i,\xi_i,\phi)=\exp[\{y_i\xi_i-b(\xi_i)\}/a_i(\phi)+c(y_i,\phi)], \quad (1)$$

where ξ_i is the natural parameter, ϕ is a possible dispersion parameter, and $a_i(\cdot)$, $b(\cdot)$, and $c(\cdot, \cdot)$ are known functions determining the specific distribution. For simplicity, we assume that ϕ is known, though most of the results in this paper could easily be extended to situations where ϕ is unknown.

The exponential dispersion family has the property that $\mu_i = E(y_i \mid \boldsymbol{w}_i) = b'(\xi_i)$ and $h_i(\mu_i) = \text{var}(y_i \mid \boldsymbol{w}_i) = b''(\xi_i)a_i(\phi)$. A GLM then relates μ_i to the covariates by assuming

$$g(\mu_i)=\sum_{j=1}^{p} w_{ij}\beta_j, \quad (2)$$

where $g(\cdot)$ is referred to as the link function, and w_{ij} and β_j are the jth elements in the covariate vector $\boldsymbol{w}_i$ and the parameter vector $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_p)^T$, respectively.

When there is no censoring, $\boldsymbol{\beta}$ can be estimated by solving the estimating equations

$$n^{-1}\sum_{i=1}^{n}\frac{w_{ij}(y_i-\mu_i)}{h_i(\mu_i)g'(\mu_i)}=0, \ \ j=1,\ldots,p, \quad (3)$$

where $g'(\mu_i) = \partial g(\mu_i)/\partial \mu_i$. For a GLM with $a_i(\phi) = a(\phi)$, $i = 1,\ldots,n$, and the canonical link function $g(\mu_i) = \xi_i$, (3) simplifies to $n^{-1}\sum_{i=1}^{n} w_{ij}(y_i - \mu_i)=0$, $j=1,\ldots,p$.

For the remainder of this paper, we assume that some covariates in $\boldsymbol{w}_i$ are subject to censoring due to lower DLs. Thus, for the ith individual, we let $\boldsymbol{w}_i=(\boldsymbol{z}_i^T, \boldsymbol{x}_i^T)^T$, where $\boldsymbol{z}_i = (z_{i1},\ldots,z_{i(p-q)})^T$ is the $(p - q)$-dimensional vector of covariates fully observed for each individual, and $\boldsymbol{x}_i = (x_{i1},\ldots,x_{iq})^T$ is the q-dimensional vector of covariates subject to censoring below $\boldsymbol{d} = (d_1,\ldots,d_q)^T$, the vector of DLs. For x_{ij}, $j = 1,\ldots,q$, we only observe $x_{ij}^*=\max(x_{ij}, d_j)$ and $\delta_{ij} = I(x_{ij} > d_j)$, so that the complete set of observed data for the ith individual is $(y_i, \boldsymbol{z}_i, \boldsymbol{x}_i^*, \boldsymbol{\delta}_i)$, where $\boldsymbol{x}_i^*=(x_{i1}^*,\ldots,x_{iq}^*)^T$ and $\boldsymbol{\delta}_i = (\delta_{i1},\ldots, \delta_{iq})^T$. Lastly, for the ith subject, we define $\boldsymbol{x}_i^c$ as the subset of $\boldsymbol{x}_i$ for which $\delta_{ij} = 0$ and $\boldsymbol{x}_i^o$ as the subset of $\boldsymbol{x}_i$ for which $\delta_{ij} = 1$. Similarly, we let $\boldsymbol{d}_i^c$ be the subset of $\boldsymbol{d}$ corresponding to $\boldsymbol{x}_i^c$. For simplicity of notation, we assume that the DLs do not change with subjects. However, all the methods discussed in this paper can be applied to data with subject-specific DLs, which occur for instance in experiments where measurements are taken using different instruments.

Although we focus on lower DLs in this paper, the ideas we discuss can be generalized to covariates censored due to upper DLs. For simplicity of notation, we do not differentiate between random variables and their observed values, necessitating the reader to distinguish based on the context. Finally, in this paper we assume that the distribution $f(\boldsymbol{x}_i \mid \boldsymbol{z}_i; \boldsymbol{\gamma})$ is a known q-variate distribution indexed by the parameter vector $\boldsymbol{\gamma}$, from which we can obtain $f(\boldsymbol{x}_i^c|\boldsymbol{z}_i, \boldsymbol{x}_i^0;\boldsymbol{\gamma})$ for all possible subsets of censored covariates, $\boldsymbol{x}_i^c$. For example, if $f(\boldsymbol{x}_i \mid \boldsymbol{z}_i; \boldsymbol{\gamma})$ is a q-variate normal distribution, then $f(\boldsymbol{x}_i^c|\boldsymbol{z}_i, \boldsymbol{x}_i^o;\boldsymbol{\gamma})$ is also a normal distribution of dimension $\leq q$. Although this distributional assumption may seem restrictive, it provides us

with a useful way to extrapolate information into the censored data region. For positive random variables, we suggest using a Box–Cox transformation so that the multivariate normal distribution can be employed. If such an assumption is unreasonable, we recommend modeling $f(\boldsymbol{x}_i \mid \boldsymbol{z}_i)$ using a series of conditional univariate distributions $f(x_{i1} \mid x_{i2}, x_{i3},\ldots, x_{iq}, \boldsymbol{z}_i) f(x_{i2} \mid x_{i3},\ldots, x_{iq}, \boldsymbol{z}_i) \ldots f(x_{iq} \mid \boldsymbol{z}_i)$. Each of these univariate distributions can then be modeled by a regression model with a flexible error term. We henceforth denote $\boldsymbol{\theta}$ as the vector of all parameters in a particular model, though we emphasize that the regression parameter vector $\boldsymbol{\beta}$, the subset of $\boldsymbol{\theta}$ defined through (1) and (2), is of primary interest for statistical inference. In the next four sections, we describe several conventional estimation methods and discuss their limitations.

2.2 Complete Case Estimation

In the presence of censored predictors, one simple method for estimating the parameters in a GLM is the complete case approach, where the statistical analysis is restricted to those individuals for whom all covariates are completely observed. That is, the complete case estimator, $\hat{\boldsymbol{\beta}}_{cc}$, is the solution to

$$n^{*-1}\sum_{i=1}^{n}\delta_i^*\frac{w_{ij}(y_i-\mu_i)}{h_i(\mu_i)g'(\mu_i)}=0,\ j=1,\ldots,p,$$

where $\delta_i^*=\prod_{j=1}^{q}\delta_{ij}$ and $n^*=\sum_{i=1}^{n}\delta_i^*$ is the effective sample size.

Proposition 1—Under the GLM defined by (1) and (2) and the regularity conditions (C1)–(C3) outlined in the on-line Supplementary Material, $\hat{\boldsymbol{\beta}}_{cc} \xrightarrow{p} \boldsymbol{\beta}$ as $n^* \to \infty$, where $\boldsymbol{\beta}$ is the true parameter vector.

Proposition 1 is a direct extension from missing-data literature, as the censoring in our context does not depend on the response. In contrast to cases with censored responses, the complete case estimator is still consistent when the covariates are subject to fixed censoring. However, as it does not take into account any data obtained for individuals with censored covariates, the complete case estimator is inefficient. If a large percentage of individuals have censored covariates, the loss in efficiency can be substantial.

2.3 Simple Imputation Methods

An alternative approach for dealing with censored covariates is to use imputation methods, where censored data are filled-in with reasonable values. A variety of imputation methods have been proposed for missing-data problems (refer to [18] for a comprehensive review), but only a few have been explored in the context of censored predictors. A simple method for dealing with censored predictors is to impute, or substitute, the censored values by DL, DL/2, or DL/√2. However, Helsel [7] and others have concluded that using functions of the DLs is inappropriate, resulting in biased parameter and standard error estimators.

An alternative imputation method is conditional mean imputation, where censored covariates are imputed with either (i) $E(\boldsymbol{x}_i^c|\boldsymbol{z}_i, \boldsymbol{x}_i^o, \boldsymbol{x}_j^c<\boldsymbol{d}_i^c;\boldsymbol{\theta})$ or (ii) $E(\boldsymbol{x}_i^c|y_i\boldsymbol{z}_i, \boldsymbol{x}_i^o, \boldsymbol{x}_j^c<\boldsymbol{d}_i^c;\boldsymbol{\theta})$, and the unknown $\boldsymbol{\theta}$ can be replaced by some consistent initial estimator. The imputed data can then be used to obtain estimates of $\boldsymbol{\beta}$ via solving (3). We refer to the resulting estimators based on these two types of imputations as $\hat{\boldsymbol{\beta}}_{M1}$ and $\hat{\boldsymbol{\beta}}_{M2}$, respectively.

Calculating the conditional expectations may require numerical integration. However, unlike maximum likelihood methods which we describe in Sect. 2.4, conditional mean imputation only requires one integration for each individual with censored covariates as opposed to one integration at each step in a maximization algorithm. Thus, this method is relatively less computationally intensive. On the other hand, unlike maximum likelihood, where the observed Fisher information matrix can be used to obtain standard errors, variance estimation for $\hat{\boldsymbol{\beta}}_{M1}$ and $\hat{\boldsymbol{\beta}}_{M2}$ is difficult. Even in the case of linear regression with missing at random covariates, Little [18] explained that the standard error estimates based on information matrices will be biased without complicated variance corrections.

To our knowledge, the use of conditional mean imputation has not been explored thoroughly in the context of GLMs with censored covariates. We demonstrate in Theorem 1 that conditional mean imputation generally gives biased estimates. We focus on the case where only one covariate, x, is subject to censoring below DL d. We consider this special case for the ease of notation and because it illustrates that conditional mean imputation usually gives biased estimates even in the simplest censoring scenario.

Theorem 1—Consider a GLM defined by (1) and (2) with covariate vector $\boldsymbol{w} = (\boldsymbol{z}^T, x)^T$. Under the regularity conditions (C1)–(C3) outlined in the on-line Supplementary Material,

1. $\hat{\boldsymbol{\beta}}_{M1} \xrightarrow{p} \boldsymbol{\beta}$ as $n \to \infty$ if and only if

$$\int_{\mathcal{Z}} \frac{\int_{-\infty}^{d} \boldsymbol{w} f(x|\boldsymbol{z})\,dx}{h(\tilde{\mu})g'(\tilde{\mu})}\{E_x(\mu|\boldsymbol{z}, x<d) - \tilde{\mu}\}f(\boldsymbol{z})d\boldsymbol{z}=\boldsymbol{0},$$

   where $\tilde{\mu} = \mu(\boldsymbol{z}, \tilde{x})$, $\tilde{x} = E(x \mid \boldsymbol{z}, x < d; \boldsymbol{\theta})$, and $\mathcal{Z}$ is the support of $\boldsymbol{z}$;

2. $\hat{\boldsymbol{\beta}}_{M2} \xrightarrow{p} \boldsymbol{\beta}$ as $n \to \infty$ if and only if

$$\int_{\mathcal{Z}}\int_{-\infty}^{d} \boldsymbol{w}\left\{\int_y \frac{(y-\tilde{\mu})}{h(\tilde{\mu})g'(\tilde{\mu})} f(y|\boldsymbol{z}, x)\,dy\right\} f(\boldsymbol{z}, x)\,dx\,d\boldsymbol{z}=\boldsymbol{0},$$

   where $\tilde{x} = E(x \mid y, \boldsymbol{z}, x < d; \boldsymbol{\theta})$.

Remark 1: The consistency of $\hat{\boldsymbol{\beta}}_{M1}$ will be fulfilled if either $\int_{-\infty}^{d} \boldsymbol{w} f(x|\boldsymbol{z})\,dx=\boldsymbol{0}$ or $E_x(\mu \mid \boldsymbol{z}, x < d) = \tilde{\mu}$. Now, $\int_{-\infty}^{d} \boldsymbol{z} f(x|\boldsymbol{z})\,dx=\boldsymbol{z}\int_{-\infty}^{d} f(x|\boldsymbol{z})\,dx=\boldsymbol{z}P(x<d|\boldsymbol{z}) \neq 0$ Unless $\boldsymbol{z} = 0$ since we implicitly assume in this paper that $P(x < d \mid \boldsymbol{z}) > 0$ for at least some values of $\boldsymbol{z}$. Thus, to show consistency regardless of the distribution for $\boldsymbol{z}$, we must have that $E_x(\mu \mid \boldsymbol{z}, x < d) = \tilde{\mu}$.

For a GLM defined by the identity link, $\mu = \boldsymbol{w}^T\boldsymbol{\beta}$, the estimator $\hat{\boldsymbol{\beta}}_{M1}$ is consistent for $\boldsymbol{\beta}$ since $E_x(\mu \mid \boldsymbol{z}, x < d) = E_x\{(\boldsymbol{z}^T, x)\boldsymbol{\beta} \mid \boldsymbol{z}, x < d\} = (\boldsymbol{z}^T, \tilde{x})\boldsymbol{\beta} = \tilde{\mu}$. When the identity link is not used, $\hat{\boldsymbol{\beta}}_{M1}$ will generally not be consistent unless the $\boldsymbol{\beta}$ parameter associated with x is zero. For example, for Poisson regression with a single covariate that is subject to censoring and the canonical link, $\mu = \exp(x\beta)$, we have $E(\mu \mid x < d) = E\{\exp(x\beta) \mid x < d\} \geq \exp\{E(x \mid x < d)\beta\} = \tilde{\mu}$ by the Conditional Jensen's Inequality, with equality only when $\beta = 0$. While it may be possible to find a distribution for $\boldsymbol{z}$ such that $\hat{\beta}_{M1} \xrightarrow{p} \beta$ even when $E_x(\mu \mid \boldsymbol{z}, x < d) \neq \tilde{\mu}$, this will generally not be the case.

Remark 2: For the canonical link, we can use similar arguments as those in Remark 1 to argue that $\hat{\boldsymbol{\beta}}_{M2}$ is consistent for $\boldsymbol{\beta}$ generally only if $E_x(\mu \mid \boldsymbol{z}, x < d) = \tilde{\mu}$, where $\tilde{\mu}$ is now based on $\tilde{x} = E(x \mid y, \boldsymbol{z}, x < d)$. Now, even with the identity link, $\mu = \boldsymbol{w}^T\boldsymbol{\beta}$, $E_x(\mu \mid \boldsymbol{z}, x < d) = E_x\{(\boldsymbol{z}^T, x)\boldsymbol{\beta} \mid \boldsymbol{z}, x < d\} \neq (\boldsymbol{z}^T, \tilde{x})\boldsymbol{\beta} = \tilde{\mu}$ unless x is conditionally independent of y so that $\tilde{x} = E(x \mid y, \boldsymbol{z}, x < d) = E(x \mid \boldsymbol{z}, x < d)$. Thus, when considering GLMs based on a canonical link, $\hat{\boldsymbol{\beta}}_{M2}$ will generally only be consistent if the response y is conditionally independent of the covariate x given z, so that the β parameter associated with x will be zero. Again, we note that while it may be possible to find a distribution for $\boldsymbol{z}$ such that $\hat{\beta}_{M2} \xrightarrow{p} \beta$ even when $E_x(\mu \mid \boldsymbol{z}, x < d) \neq \tilde{\mu}$, this will generally not be the case.

2.4 Maximum Likelihood Estimation

Another method to estimate $\boldsymbol{\beta}$ is through maximizing the likelihood contributed by all the observed data. We represent this likelihood as

$$\prod_{i=1}^{n} \int_{-\infty}^{\boldsymbol{d}_i^c} f(y_i, \boldsymbol{x}_i^o, \boldsymbol{x}_i^c \mid \boldsymbol{z}_i; \theta) d\boldsymbol{x}_i^c, \quad (4)$$

where $\int_{-\infty}^{\boldsymbol{d}_i^c}$ is an integral whose dimension corresponds to the length of $\boldsymbol{x}_i^c$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)^T$ is the vector of parameters indexing $f(y_i, \boldsymbol{x}_i^o, \boldsymbol{x}_i^c \mid \boldsymbol{z}_i; \boldsymbol{\theta}) = f(y_i \mid \boldsymbol{z}_i, \boldsymbol{x}_i; \boldsymbol{\beta}) \times f(\boldsymbol{x}_i^o, \boldsymbol{x}_i^c \mid \boldsymbol{z}_i; \boldsymbol{\gamma})$, with $f(y_i \mid \boldsymbol{z}_i, \boldsymbol{x}_i; \boldsymbol{\beta})$ distributed according to (1) and (2) and $f(\boldsymbol{x}_i^o, \boldsymbol{x}_i^c \mid \boldsymbol{z}_i; \boldsymbol{\gamma})$ as the conditional density of $(\boldsymbol{x}_i^o, \boldsymbol{x}_i^c)$ given $\boldsymbol{z}_i$. Note that in (4), $\boldsymbol{x}_i^c$ is a subject-specific set of covariates whose length corresponds to the number of censored covariates for the ith individual.

For illustrative purposes, suppose that $\boldsymbol{z} = (z_1, z_2)^T$ and $\boldsymbol{x} = (x_1, x_2, x_3)^T$. The contribution to (4) for an individual i for whom we observe the DLs d_1 and d_3 rather than x_{i1} and x_{i3} is

$$\int_{-\infty}^{d_3} \int_{-\infty}^{d_1} f(y_i, x_1, x_{i2}, x_3 \mid z_{i1}, z_{i2}; \boldsymbol{\theta}) \, dx_1 \, dx_3.$$

The likelihood in (4) generally does not have a closed form, and for this reason we often must resort to approximations or numerical integration when maximizing this likelihood. In general, approximations are difficult to derive while numerical integration is notoriously difficult in higher dimensions. Recently, May et al. [23] proposed a Monte Carlo EM

algorithm for maximizing (4) which was shown to work well but is also computationally intensive because it uses Monte Carlo methods to approximate numerical integration.

2.5 Bayesian Methods

Bayesian methods have been considered as a practical method for handling missing covariate data. Until a recent work by Wu et al. [41], Bayesian methods had not been explored in the specific context of GLMs with covariates subject to DLs. In the following, we summarize a basic Bayesian approach, which is an extension of the missing-data methods described in detail by Ibrahim et al. [13, 14].

In a fully Bayesian analysis, inference for $\boldsymbol{\beta}$ is carried out using the observed data posterior, given by

$$\pi(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{w}^o) \propto \left\{\prod_{i=1}^{n}\int_{-\infty}^{\boldsymbol{d}_i^c} f(y_i,\boldsymbol{x}_i^o,\boldsymbol{x}_i^c|\boldsymbol{z}_i;\boldsymbol{\theta})d\boldsymbol{x}_i^c\right\}\pi(\boldsymbol{\theta}), \quad (5)$$

where $\pi(\boldsymbol{\theta})$ is a prior for $\boldsymbol{\theta}=(\boldsymbol{\beta}^T,\boldsymbol{\gamma}^T)^T$, $\boldsymbol{w}^o=\{\boldsymbol{z}_i,\boldsymbol{x}_i^*,\boldsymbol{\rho}_i\}_{i=1}^n$, and all other notation is identical to that in Sect. 2.4.

After specifying a prior, $\pi(\boldsymbol{\theta})$, we can obtain samples from (5) using Gibbs sampling. That is, we iteratively sample from the full conditional distributions $f(\boldsymbol{x}_i^c|\boldsymbol{\beta},\boldsymbol{\gamma},y_i,\boldsymbol{z}_i,\boldsymbol{x}_i^o,\boldsymbol{x}_i^c<\boldsymbol{d}_i^c)$, $i=1,\ldots,n$, $\pi(\boldsymbol{\beta}|\,\boldsymbol{\gamma},\boldsymbol{y},\boldsymbol{z},\boldsymbol{x}^o,\boldsymbol{x}^c)$, and $\pi(\boldsymbol{\gamma}|\,\boldsymbol{\beta},\boldsymbol{y},\boldsymbol{z},\boldsymbol{x}^o,\boldsymbol{x}^c)$, where $\boldsymbol{z}=\{\boldsymbol{z}_i\}_{i=1}^n$, $\boldsymbol{x}^o=\{\boldsymbol{x}_i^o\}_{i=1}^n$, $\boldsymbol{x}^c=\{\boldsymbol{x}_i^c\}_{i=1}^n$, and initial values for $\boldsymbol{\beta}$ at the first iteration can be chosen as complete case estimates while initials values for $\boldsymbol{\gamma}$ can be chosen as the maximum likelihood estimates or any other reasonable values. Note that since we do not actually observe $\boldsymbol{x}^c$, we are treating each element in this set as a parameter vector. After obtaining N samples for $\boldsymbol{\beta}$, $\boldsymbol{\beta}^{(1)},\ldots,\boldsymbol{\beta}^{(N)}$, we can estimate $\boldsymbol{\beta}$ and its standard error by the posterior mode and standard deviation. Standard Markov chain Monte Carlo (MCMC) sampling methods can be used to obtain draws from the full conditional distributions above when closed forms are not available.

As an alternative to fully Bayesian methods, multiple imputation has become increasingly popular over the past several decades. The multiple imputation estimator was originally proposed by Rubin [33] for parameter estimation in the presence of missing data. The idea is based on obtaining imputations for missing data by making m random draws for each missing value from the conditional distribution of the missing variable(s) given the other variables in the model. These m sets of draws can then be used to construct m complete data sets, based on which conventional methods designed for complete data can be used to calculate a final parameter estimate,

$$\hat{\boldsymbol{\theta}}=m^{-1}\sum_{k=1}^{m}\hat{\boldsymbol{\theta}}^{(k)},$$

where $\hat{\boldsymbol{\theta}}^{(k)}$ is the estimate based on the kth imputed data set.

To obtain *m* imputations in a Bayesian framework, we must assume a prior for $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta})$, and then make random draws from the posterior predictive distribution of the censored data,

$$f(\boldsymbol{x}_i^c|y_i,\boldsymbol{z}_i,\boldsymbol{x}_i^o,\boldsymbol{x}_i^c<\boldsymbol{d}_i^c) \propto \int f(\boldsymbol{x}_i^c|y_i,\boldsymbol{z}_i,\boldsymbol{x}_i^o,\boldsymbol{x}_i^c<\boldsymbol{d}_i^c;\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{w}^o)d\boldsymbol{\theta},$$

where $\pi(\boldsymbol{\theta}|\boldsymbol{y},\boldsymbol{w}^o)$ is the posterior distribution given by (5). This Bayesian multiple imputation is also referred to as "proper" imputation in the literature [33, 34] because it leads to valid large sample inferences in a Bayesian framework.

One attraction of the proper (Bayesian) multiple imputation lies in the ability to construct a simple variance estimator [33],

$$\hat{\text{var}}(\hat{\boldsymbol{\theta}})=V_W+(1+m^{-1})V_B, \quad (6)$$

where $V_W=m^{-1}\sum_{k=1}^{m}\hat{V}^{(k)}$, with $\hat{V}^{(k)}$ representing the covariance matrix for the *k*th imputed data set, is the "within" imputation variance, and

$V_B=(m-1)^{-1}\sum_{k=1}^{m}(\hat{\boldsymbol{\theta}}^{(k)}-\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}^{(k)}-\hat{\boldsymbol{\theta}})^T$ is the between imputation variance.

All of the methods described thus far have clear disadvantages. Complete case analyses lead to inefficient estimates while substitution and conditional mean imputation methods generally lead to inconsistent estimates. Although Bayesian methods are flexible, like maximum likelihood methods they often require a great deal of computational effort, and they also necessitate the user to define priors and monitor the convergence of sampling chains. In the following section, we present a frequentist-based, "improper" multiple imputation estimator that does not suffer from many of these limitations.

3 Improper Multiple Imputation Method

Alternatively to the proper (Bayesian) multiple imputation, we can conduct multiple imputation in a frequentist framework by obtaining *m* draws from the distribution $f(\boldsymbol{x}_i^c|y_i,\boldsymbol{z}_i,\boldsymbol{x}_i^o,\boldsymbol{x}_i^c<\boldsymbol{d}_i^c;\hat{\boldsymbol{\theta}}^I)$, where $\hat{\boldsymbol{\theta}}^I$ is a consistent initial estimator for $\boldsymbol{\theta}$. This type of multiple imputation is generally referred to as "improper" multiple imputation [33], for which (6) is no longer consistent because the variation in $\hat{\boldsymbol{\theta}}^I$ is not taken into account. However, we believe that the improper multiple imputation approach is advantageous over proper multiple imputation for two primary reasons: (i) obtaining draws from the posterior predictive distribution can be difficult and computationally intensive, and (ii) improper multiple imputation produces more efficient estimates for a finite number of imputations (Wang and Robins, 1998). A similar discussion of the computational advantage of improper multiple imputation was provided in [34] for a different setup.

In the remainder of this section, we develop an improper multiple imputation method for GLMs with covariates subject to DLs and propose a consistent variance estimator using the theory of [40] and [32]. We show that the method is practical and straightforward and that

obtaining standard errors is only minimally more difficult than for proper multiple imputation.

Our proposed multiple imputation algorithm is as follows:

Step 1 Using all the data ($\boldsymbol{z}_i$, $\boldsymbol{x}_i^*$, δ_i), $i = 1,\ldots,n$, and the DLs $\boldsymbol{d}$, obtain the maximum likelihood estimate, $\hat{\boldsymbol{\gamma}}^I$, as an initial estimate for $\boldsymbol{\gamma}$, the parameters indexing $f(\boldsymbol{x}_i \mid \boldsymbol{z}_i)$.

Step 2 Using the complete cases, obtain the maximum likelihood estimate, $\hat{\boldsymbol{\beta}}^I$, as an initial estimate for $\boldsymbol{\beta}$, the parameters indexing $f(y_i \mid \boldsymbol{z}_i, \boldsymbol{x}_i)$.

Step 3 Form the 1 st imputed data set by generating $\tilde{\boldsymbol{x}}_i^c$ for each individual i with censoring, $i = 1,\ldots,n$, from $f(\boldsymbol{x}_i^c | y_l, \boldsymbol{z}_i, \boldsymbol{x}_i^o, \boldsymbol{x}_i^c < \boldsymbol{d}_i^c; \hat{\boldsymbol{\theta}}^I)$ to replace the censored covariates $\boldsymbol{x}_i^c$, where $\hat{\boldsymbol{\theta}}^I = (\hat{\boldsymbol{\beta}}^{I^T}, \hat{\boldsymbol{\gamma}}^{I^T})^T$ Denote $\tilde{\boldsymbol{x}_i}$ as the resulting covariate vector for the ith subject that includes both $\boldsymbol{x}_i^o$ and $\tilde{\boldsymbol{x}}_i^c$.

Step 4 Using the complete data set, $(y_i, \boldsymbol{z}_i, \tilde{\boldsymbol{x}_i}), i = 1,\ldots,n$, obtain an estimate $\hat{\boldsymbol{\beta}}^{(1)}$ for $\boldsymbol{\beta}$ by solving the estimating equations given by (3), and an estimate $\hat{\boldsymbol{\gamma}}^{(1)}$ for $\boldsymbol{\gamma}$ by maximum likelihood estimation. Let $\hat{\boldsymbol{\theta}}^{(1)} = (\hat{\boldsymbol{\beta}}^{(1)T}, \hat{\boldsymbol{\gamma}}^{(1)T})^T$.

Step 5 Repeat Steps 3 and 4 m times, resulting in $\hat{\boldsymbol{\theta}}, \ldots, \hat{\boldsymbol{\theta}}^{(m)}$.

Step 6 Finally, the improper multiple imputation estimator is defined as $\hat{\boldsymbol{\theta}}_{\mathrm{IMI}} = m^{-1} \sum_{k=1}^{m} \hat{\boldsymbol{\theta}}^{(k)}$.

In Step 3 of the above algorithm, the method for obtaining draws from the truncated distribution of $f(\boldsymbol{x}_i^c | y_i, \boldsymbol{z}_i, \boldsymbol{x}_i^o; \hat{\boldsymbol{\theta}}^I)$ may differ for various GLMs and distributions $f(\boldsymbol{x}_i \mid \boldsymbol{z}_i; \boldsymbol{\gamma})$. In normal linear models where $f(\boldsymbol{x}_i \mid \boldsymbol{z}_i)$ is also normal, we can sample directly from a truncated jointly normal distribution. Most generally, by noting that Bayes theorem gives

$$f(\boldsymbol{x}_i^c | y_i, \boldsymbol{z}_i, \boldsymbol{x}_i^o) = \frac{f(y_i | \boldsymbol{z}_i, \boldsymbol{x}_i^o, \boldsymbol{x}_i^c) f(\boldsymbol{x}_i^c | \boldsymbol{z}_i, \boldsymbol{x}_i^o)}{f(y_i | \boldsymbol{z}_i, \boldsymbol{x}_i^o)},$$

the Metropolis–Hastings algorithm or other MCMC methods may be used. For the simulation in Sect. 4 and application in Sect. 5, we consider a logistic model for a binary y, $y_i, y_i | \boldsymbol{x}_i, \boldsymbol{z}_i \overset{\text{ind.}}{\sim} \text{Bernoulli}([1+\exp\{-(\boldsymbol{z}_i^T, \boldsymbol{x}_i^T)\boldsymbol{\beta}\}]^{-1})$, and a multivariate normal distribution for $\boldsymbol{x} | \boldsymbol{z}, \boldsymbol{x}_i | \boldsymbol{z}_i \overset{\text{ind.}}{\sim} N(\Lambda \boldsymbol{z}_i, \Omega)$, where Λ is a $q \times (p - q)$ matrix of mean parameters. Then, after obtaining the initial estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma} = (\Lambda, \Omega)$, $\hat{\boldsymbol{\beta}}^I$ and $\hat{\boldsymbol{\gamma}}^I = (\hat{\Lambda}^I, \hat{\Omega}^I)$, we can use the following acceptance-rejection algorithm to obtain an imputation draw for the ith subject:

1. Generate a candidate imputation, $\tilde{\boldsymbol{x}}_i^c$, for the vector $\boldsymbol{x}_i^c$ from the truncated conditional normal distribution $f(\boldsymbol{x}_i^c | \boldsymbol{z}_i, \boldsymbol{x}_i^o, \boldsymbol{x}_i^c < \boldsymbol{d}_i^c; \hat{\boldsymbol{\gamma}}^I)$;
2. Generate U from a Uniform(0,1) distribution;

3. Let $\tilde{\boldsymbol{x}}_i^c$ be the imputed vector for $\boldsymbol{x}_i^c$ If $U < f(y_i|\boldsymbol{z}_i, \boldsymbol{x}_i^o, \tilde{x}_i^c)$, and repeat Step 1 otherwise.

This acceptance-rejection algorithm works for logistic regression because

$$\frac{f(y_i|\boldsymbol{z}_i, \boldsymbol{x}_i^o, \boldsymbol{x}_i^c) f(\boldsymbol{x}_i^c|\boldsymbol{z}_i, \boldsymbol{x}_i^o)}{f(y_i|\boldsymbol{z}_i, \boldsymbol{x}_i^o)} \leq \frac{f(\boldsymbol{x}_i^c|\boldsymbol{z}_i, \boldsymbol{x}_i^o)}{f(y_i|\boldsymbol{z}_i, \boldsymbol{x}_i^o)}$$

since $f(y_i|\boldsymbol{z}_i, \boldsymbol{x}_i^o, \boldsymbol{x}_i^c) \leq 1$ and $f(y_i|\boldsymbol{z}_i, \boldsymbol{x}_i^o)$ does not depend on $\boldsymbol{x}_i^c$.

For improper multiple imputation, Rubin's variance formula (6) is not consistent for the variance of $\hat{\boldsymbol{\theta}}_{\text{IMI}}$ because it does not take into account variation in the initial estimates $\hat{\boldsymbol{\beta}}^I$ and $\hat{\boldsymbol{\gamma}}^I$. Extending theory in [40] and [36], we obtain the following asymptotic properties of $\hat{\boldsymbol{\theta}}_{\text{IMI}}$.

Theorem 2

Consider a GLM defined by (1) and (2). Then, under the regularity conditions (C1)–(C3) outlined in the on-line Supplementary Material,

$$n^{1/2}(\hat{\boldsymbol{\theta}}_{\text{IMI}} - \boldsymbol{\theta} \xrightarrow{d} N(0, \sum),$$

where

$$\begin{aligned}\sum = &\{I^F(\boldsymbol{\theta})\}^{-1} + (1+m^{-1})\{I^F(\boldsymbol{\theta})\}^{-1} D \{I^F(\boldsymbol{\theta})\}^{-1} \\ &+ \{I^F(\boldsymbol{\theta})\}^{-1} D \text{var}(\hat{\boldsymbol{\theta}}^I) D \{I^F(\boldsymbol{\theta})\}^{-1},\end{aligned} \tag{7}$$

$D = \{I^F(\boldsymbol{\theta}) - I(\boldsymbol{\theta})\}$, $I^F(\boldsymbol{\theta}) = E\{S^F(y, \boldsymbol{w}, \boldsymbol{\theta}) S^{F^T}(y, \boldsymbol{w}, \boldsymbol{\theta})\}$ is the full-data information matrix, and $I(\boldsymbol{\theta}) = E\{S(y, \boldsymbol{z}, \boldsymbol{x}^*, \delta, \boldsymbol{\theta}) S^T(y, \boldsymbol{z}, \boldsymbol{x}^*, \delta, \boldsymbol{\theta})\}$ is the observed-data information matrix, evaluated at the true parameter vector, $\boldsymbol{\theta}$.

To clarify, the full-data information matrix and score vector for $\boldsymbol{\theta}$, I^F and S^F are based on the complete data with no censoring, while the observed-data information matrix and score vector for $\boldsymbol{\theta}$, I and S are based on the censored data. Although the theory here is driven based on a joint model for $(y_i, \boldsymbol{z}_i, \boldsymbol{x}_i)$, we note that in this paper we have only suggested modeling $f(y_i \mid \boldsymbol{z}_i, \boldsymbol{x}_i)$ and $f(\boldsymbol{x}_i \mid \boldsymbol{z}_i)$ in practice. Also, while the technical details for Theorem 2, given in the on-line Supplementary Material, are similar to those provided by Wang and Robins [40], the specific type of missingness due to DLs that we encounter is not missing at random and thus does not fit into the general framework considered by Wang and Robins [40]. The same basic theory follows, however, because likelihood methods can be applied to the censoring problem.

To estimate (7), Wang and Robins [40] proposed estimating $I^F(\boldsymbol{\theta}_0)$ and $I(\boldsymbol{\theta}_0)$ using the consistent estimators

$$\hat{I}^F(\boldsymbol{\theta}_0) = -m^{-1}\sum_{k=1}^{m}\left[n^{-1}\sum_{i=1}^{n}\frac{\partial S^F(y_i, \boldsymbol{z}_i, \tilde{\boldsymbol{x}}_{ik}, \hat{\boldsymbol{\theta}}^{(k)})}{\partial \boldsymbol{\theta}^T}\right], \quad (8)$$

$$\hat{I}(\boldsymbol{\theta}_0) = n^{-1}\sum_{i=1}^{n}\{m(m-1)\}^{-1} \times \sum_{k,k'=1,\ldots,m,k\neq k'}\sum S^F(y_i, \boldsymbol{z}_i, \tilde{\boldsymbol{x}}_{ik}, \hat{\boldsymbol{\theta}}^{(k)}) S^{F^T}(y_i, \boldsymbol{z}_i, \tilde{\boldsymbol{x}}_{ik'}, \hat{\boldsymbol{\theta}}^{(k')}),$$

Where $S^F = \frac{\partial}{\partial\theta}\{\ell(\boldsymbol{\theta}; y_i, \boldsymbol{z}_i, \tilde{\boldsymbol{x}}_{ik})\}$ is the score vector from the full data with log-likelihood $\ell(\boldsymbol{\theta}; y_i, \boldsymbol{z}_i, \boldsymbol{x}_{ik}, \tilde{\boldsymbol{x}}_{ik})$, $\tilde{\boldsymbol{x}}_{ik}$ denotes the kth vector of imputed values for the ith individual, and $\hat{\boldsymbol{\theta}}^{(k)}$ denotes the estimated parameter vector based on the kth data set. The var($\hat{\boldsymbol{\theta}}$) can be estimated by the observed information matrix for the initial parameter estimates $\hat{\boldsymbol{\theta}}$.

It is often laborious to estimate the asymptotic variance (7) using the consistent estimators given by (8). For this reason, we propose an alternative easy-to-calculate approximate variance formula,

$$\hat{\sum} = V_W + (1+m^{-1})V_B + V_B V_W^{-1}\mathrm{var}(\hat{\boldsymbol{\theta}}^I)V_W^{-1}V_B, \quad (9)$$

where V_W and V_B are the within- and between-imputation variances, respectively. This estimator can be derived by noting that Rubin's variance formula (6) is consistent for the first two terms of (7) as $n \to \infty$ and $m \to \infty$ [40]. Although (9) is not a consistent estimator for finite m, we have found that it works well even for a high percentage of censoring (>50 %) and a moderate number of imputed data sets, say $m \in [15, 30]$.

Finally, as a caution to the reader, we want to emphasize that multiple imputation procedures, in general, only lead to valid inference when imputations are drawn conditional on both the covariates and the response. In the accompanying on-line Supplementary Material, we provide Theorem 3 to formally give this result in the context of censored covariates. Additionally, we provide some insight on the direction of bias that will be observed if the response is ignored during imputation.

4 Simulation Study

We considered a situation with two covariates, x_1 and x_2, both subject to censoring, and a single binary response variable, y. We let x_1 and x_2 be bivariate normal with zero mean, unit variance, and correlation $\rho = 0.5$. We assumed that $y \mid x_1, x_2$ follows a logistic regression model, $y \mid x_1, x_2 \sim$ Bernoulli($[1 + \exp\{-\beta_0 - x_1\beta_1 - x_2\beta_2\}]^{-1}$). We let $\boldsymbol{\beta} = (3, 1.5, -3)^T$ and considered three levels of censoring percentage, 20,40, and 60 %, where the censoring percentage is defined as the proportion of cases with at least one covariate below the DL $d = d_1 = d_2$. For each scenario, the simulation was repeated 2000 times with 200 subjects. We provide results for other simulation setups in the on-line Supplementary Material.

4.1 Estimation of β

We obtained estimates for β using the following methods: (i) omniscient method (Omni) in which all observations are known as if no censoring occurred; (ii) complete case analysis (CC); (iii) substitution method (SUB) using DL/√2 to replace censored values; (iv) conditional mean imputation (Mean1) conditioning only on the uncensored covariates; (v) conditional mean imputation (Mean2) conditioning on both the response and the uncensored covariates; (vi) maximum likelihood method (ML) using a Monte Carlo EM algorithm; (vii) fully Bayesian method (FB); (viii) proper (Bayesian) multiple imputation (PMI); and (ix) improper (frequentist) multiple imputation (IMI).

The omniscient method serves as a gold standard while the IMI method is the proposed improper multiple imputation method described in Sect. 3. For the ML method, we estimated β using the Monte Carlo EM algorithm described in [23] with 250 samples for each censored observation in the E-step. For the FB and PMI methods we used Jeffrey's prior for β and a multivariate normal-Wishart conjugate prior for the γ parameters indexing the distribution of the covariates. We used the Metropolis-Hastings algorithm to obtain 1000 draws from the full conditional for β. For the ML, FB, PMI, and IMI methods, we implemented the acceptance-rejection algorithm described in Sect. 3 to draw imputations for censored observations. For both the imputation methods PMI and IMI, we obtained $m = 15$ imputed data sets. For each of the methods except the FB method we used the bias correction methods described by Firth [5] to eliminate the first-order bias in the β estimates for logistic regression. For the FB method, we estimated β by the posterior mode.

To estimate the standard errors for the SUB, Mean1, and Mean2 estimates, we used the observed Fisher's information matrix based on the filled-in data set. For the maximum likelihood method, we employed bootstrapping as described in [23], while for the fully Bayesian method we used the posterior standard deviation. For the PMI and IMI methods, we applied the formulas (6) and (9), respectively.

Table 1 summarizes the average bias, simulation standard deviation, average estimated standard error, and empirical 95 % coverage probability of the parameter estimates based on the above estimation methods for 20, 40, and 60 % censoring. The last column of Table 1 shows the average elapsed time (in seconds) needed to compute the estimates for a single simulated data set. For fair comparison, we used R to program each of the methods.

The CC, ML, PMI, and IMI methods yielded estimates for β with minimal or no bias. The CC and IMI estimates are asymptotically unbiased as proven by Proposition 1 and Theorem 2, while the unbiasedness of the ML estimates is expected since they are derived by directly maximizing the likelihood of the available data. The FB method has a slight bias while the SUB, Mean1, and Mean2 methods are more clearly biased, as was indicated by previous literature [7] and Theorem 1.

We note that among the approximately unbiased methods, the ML, FB, PMI, and IMI methods are clearly more efficient than the CC analysis. For example, with 60 % censoring, the relative efficiency gains for these estimators compared to the CC estimator range from 54 to 59 % for $\hat{\beta_0}$, 55 to 63 % for $\hat{\beta_1}$, and 36 to 43 % for $\hat{\beta_2}$, where the relative efficiency

gain is calculated by $\{\text{var}(\hat{\beta}_{CC}) - \text{var}(\hat{\beta})\}/\text{var}(\hat{\beta}_{CC})$, and $\hat{\beta}_{CC}$ and $\hat{\beta}$ are the estimators for the CC and unbiased methods, respectively. While the ML method is the most efficient of these methods, the FB, PMI, and IMI methods perform very comparably.

For the majority of the methods, the average of the standard error estimates is a relatively accurate estimate of the simulation standard deviation, with the exception of the Mean2 estimator. However, while coverage probabilities are good for the CC, ML, FB, PMI, and IMI methods, they are poor for the SUB, Mean1, and Mean2 methods, especially with higher censoring.

Each of the methods was implemented in R on a computer with an Intel Core i7 870 @ 2.93-GHz processor and 8 GB of RAM. We found that the IMI is much faster and easier to program than both the PMI and ML methods. Additionally, since the PMI method relies on Bayesian techniques, convergence of all the parameters must be monitored. The PMI method is shown to be faster than the FB method since we only need m independent draws from the posterior predictive distribution for the data rather than the entire posterior distribution of the parameters. We also note that we included standard error estimation in the calculation for the elapsed computational time for all the methods except the ML method. Thus the ML method actually takes longer than indicated in Table 1.

In summary, the CC, ML, FB, PMI, and IMI methods all produce approximately unbiased parameter estimates while the SUB, Mean1, and Mean2 methods lead to biased estimates. Among the unbiased methods, the CC analysis is inferior as it leads to estimates with higher variance. The ML, FB, PMI, and IMI estimators behave similarly, though the ML estimator slightly outperforms the others in terms of efficiency while the proposed IMI method is computationally superior.

4.2 Variance Estimation of $\hat{\beta}_{IMI}$

We considered four methods for estimating the variance of the proposed improper multiple imputation estimator, $\hat{\beta}_{IMI}$: (i) a consistent estimator using (8), (ii) the approximate estimator given by (9), (iii) Rubin's variance formula (6), and (iv) a boot-strap procedure. We label the corresponding standard error estimates as SE_C, SE_A, SE_R, and SE_B, respectively. To implement the bootstrap procedure, we randomly re-sampled the individuals in the simulated data set 50 times, calculated $\hat{\beta}_{IMI}$ for each bootstrapped data set, and found the standard deviation of these 50 estimates. Table 2 gives the simulation standard deviations (*SDs*) of $\hat{\beta}_{IMI}$ and the standard error estimates SE_C, SE_A, SE_R, and SE_B averaged across the 2000 simulated data sets.

As we would expect by the theory outlined in Sect. 3, the SE_R estimates are smaller than the simulation *SDs*, confirming that Rubin's formula (6) underestimates the variation in the improper multiple imputation estimator. The SE_C and SE_A estimates are very close to the simulation *SDs* for all three parameters, validating the corrected asymptotic variance given by (7) and both estimation methods (8) and (9). The SE_B estimates tend to be high, possibly due to heavy censoring in some bootstrap samples.

4.3 Incorrect Distributional Assumptions on the Covariates

In practice, it may often be the case that we inaccurately model the conditional distribution of the covariates subject to DLs, $f(\boldsymbol{x}_i \mid \boldsymbol{z}_i; \gamma)$. To assess the sensitivity of various methods against the misspecification of the parametric distribution, we conducted an additional simulation where the assumed distribution of the censored covariates is incorrect. Specifically, we let the covariates x_1 and x_2 be bivariate gamma, as defined by a Clayton copula [4] with each variable having marginal shape and scale parameters 2 and 0.5. We generated samples from this distribution using the Copula package in R [10]. We naively modeled the joint distribution of log (x_1) and log (x_2) by a bivariate normal even though the Q-Q plots shown in Fig. 1 do not seem to support this assumption. The remainder of the simulation setup was not changed.

Table 3 summarizes the average bias, simulation standard deviation, average estimated standard error, and empirical coverage probability of the parameter estimates for the Omni, CC, ML, FB, PMI, and IMI methods for 20, 40, and 60 % censoring. Because the normal assumption is incorrect, all of the methods except the Omni and CC methods have clear biases on the parameter estimates. However, standard error estimation is still reasonable and coverage probabilities are close to 95 %, even at high censoring percentages. While this simulation demonstrates that the estimates are not overly sensitive to incorrect distributional assumptions for $f(\boldsymbol{x}_i \mid \boldsymbol{z}_{i:}; \gamma)$, we refer to the last paragraph of Sect. 2.1 for modeling suggestions when normality is unreasonable.

5 Application to the GenIMS Data

We illustrate several of the analysis methods discussed in Sects. 2 and 3 by applying them to the Genetic and Inflammatory Markers of Sepsis (GenIMS) data set. One of the main purposes of the GenIMS study was to identify relationships between cytokine levels in the body and the development of severe sepsis, defined in this study as community acquired pneumonia (CAP) plus organ dysfunction. A second purpose of this study, upon which we focus in this paper, was to study the relationship between cytokine levels and 90-day survival (event of surviving at least 90 days) for patients with CAP [15].

Cytokines are cell-signaling protein molecules that are sent out by cells in the immune system. Three cytokines were measured in this study: tumor necrosis factor (TNF), interleukin-6 (IL-6), and interleukin-10 (IL-10). The TNF and IL-6 cytokines serve as biomarkers of pro-inflammatory responses to CAP while IL-10 serves as a biomarker of anti-inflammatory responses to CAP. It has been thought that pro- and anti-inflammatory responses in the body help explain the development of severe sepsis and resulting deaths, and that understanding these relationships could be important for developing medical treatments.

The data for the GenIMS study were obtained by first enrolling individuals with CAP immediately after admission to a hospital and then collecting biological measurements and demographic information for each individual. In our analysis, we considered the 1418 patients who actually acquired CAP, necessitated a hospital stay, and had TNF, IL-6, and IL-10 measurements taken on the first day of hospitalization.

We used a logistic regression model to explain the relationship between 90-day survival and six covariates: the levels of the three cytokines, TNF, IL-6, and IL-10, on the first day of hospitalization, sex (1 representing males, 0 representing females), race (1 representing Caucasians, 0 representing all other races), and age. The three biomarkers, TNF, IL-6, and IL-10, were all subject to censoring below the detection thresholds of 4,2 or 5, and 5 pg/ml, respectively, with censoring proportions of 35.54, 13.40, and 46.83 %, respectively. A total of 64 % of the individuals had at least one of these measurements censored.

We assumed that given the sex, race, and age covariates, the natural logarithms of TNF, IL-6, and IL-10 are multivariate normal. Although our analysis was based on the logarithms of TNF, IL-6, and IL-10, we henceforth continue to refer to the log-cytokines as TNF, IL-6, and IL-10. Specifically, we assumed $(\mathrm{TNF}_i, \mathrm{IL}-6_i, \mathrm{IL}-10_i)^T \underset{\sim}{\mathrm{ind.}} N_3(\Lambda z_i, \Omega)$, where Λ is a 3 × 4 matrix of mean parameters, $\boldsymbol{z}_i = (1, \mathrm{sex}_i, \mathrm{race}_i, \mathrm{age}_i)^T$, and Ω is a 3 × 3 variance-covariance matrix. To informally verify this normality assumption, we plotted normal Q-Q plots for the log of the three cytokines, shown in Fig. 2. These Q-Q plots are only partially complete because they show only the sample quantiles above the p_0th quantile, where p_0 is the sample censoring proportion, as the lower quantiles cannot be estimated nonparametrically. The Q-Q plots indicate that the marginal distributions of the log-transformed TNF, IL-6, and IL-10 covariates are approximately normal.

With the multivariate normal distributional assumption for the censored covariates, we conducted a logistic regression analysis for 90-day survival using the proposed improper multiple imputation method (IMI), maximum likelihood method (ML), fully Bayesian method (FB), proper multiple imputation method (PMI), and complete case analysis (CC; based on 511 individuals). Table 4 summarizes the parameter estimates and estimated standard errors for all these methods. We used (9) to estimate the standard errors for the IMI estimates and Rubin's formula (6) to estimate the standard errors for the PMI estimates. For the maximum likelihood method, we obtained standard errors by bootstrapping while for the fully Bayesian method we estimated the standard errors by the standard deviations of the posterior distributions of the parameters.

All of the analysis methods produced similar parameter estimates, but the standard errors for the CC method are significantly higher. As an example, the parameter estimates based on the IMI method indicate that TNF, IL-6, and IL-10 are only minimally important for predicting the 90-day survival of patients, with *p*-values of 0.347, 0.263, and 0.251, respectively. However, the sex and age covariates are statistically significant in the model at the 0.05 level, with older males having lower probabilities of 90-day survival. For the CC analysis, age is the only statistically significant covariate. The computation times for the PMI, FB, and ML methods were approximately 12, 15, and 40 times longer than the proposed IMI method.

Although our analysis suggested that none of the three cytokines, TNF, IL-6, and IL-10, are statistically significant in the logistic regression, we note that there are moderate correlations between these variables, ranging from 0.22 to 0.40 in absolute value. We repeated the analysis including only one of the cytokines in the model at a time. The associated *p*-values for TNF, IL-6, and IL-10 using the proposed IMI method are 0.049, 0.014, and 0.030,

respectively. With a CC analysis, the *p*-values are 0.007, 0.028, and 0.364, respectively. Thus, based on the IMI analysis, we found that the levels of the three cytokines are important for predicting 90-day survival, but including all three in the model leads to some prediction redundancy. For the CC analysis, we were still unable to conclude that IL-10 is significant.

6 Discussion

We proposed an improper multiple imputation method for handling covariates censored due to DLs in GLMs. We have proven that the proposed estimator is consistent and asymptotically normal. We demonstrated that Rubin's variance formula (6) is inadequate for improper imputation and a corrected variance estimator is needed to account for variation in the initial parameter estimates. We suggested two variance estimation methods for the improper multiple imputation estimator and showed empirically that these estimators perform well.

We also extended several missing-data methods to the context of GLMs with covariates censored due to DLs and compared these and other competing methods to the proposed improper multiple imputation estimator. We showed both theoretically and empirically that conditional mean imputation methods may give biased parameter estimates. We also showed that substitution methods are inadequate while complete case analyses are inefficient. Maximum likelihood, fully Bayesian, proper multiple imputation, and improper multiple imputation methods all perform similarly with regard to estimation, but the improper multiple imputation estimator is computationally more efficient.

In this paper, we generally assumed a multivariate normal model for $f(\boldsymbol{x}_i \mid \boldsymbol{z}_i)$, possibly after transformation, which eases the computational burden. The implementation of the improper multiple imputation method is more challenging for other distributional models, especially when multiple covariates are subject to censoring. However, this comment applies to each of the methods. Future research could focus on using more flexible multivariate distributions or semiparametric techniques.

In cases where censoring is on the response y, the complete case estimator is generally both invalid and inefficient. In contrast, when censoring is on the covariates $\boldsymbol{x}$, we showed that the complete case estimator is still valid. Differently from the imputation and maximum likelihood methods, the validity of the complete case estimator in our setup does not rely on any parametric assumption on the distribution of $\boldsymbol{x} \mid \boldsymbol{z}$. We note that in practice it is difficult to assess the parametric distributional assumptions for observations below the DLs. Therefore, if the censoring level on $\boldsymbol{x}$ is not high, the complete case analysis would be a good option since it is robust, valid, and easy to apply. However, in cases with heavy censoring on $\boldsymbol{x}$, more complicated imputation methods should be considered to improve the efficiency.

Lastly, we note that if in addition to the censored covariates there is also data that is missing at random, the fully Bayesian, proper multiple imputation and maximum likelihood methods can all easily be extended using standard methods. However, the complete case method and

the improper multiple imputation method may no longer be appropriate if missingness depends on the response.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Wang's research was supported in part by NSF award DMS-1007420 and NSF CAREER award DMS-1149355 and Zhang's research was supported in part by the HIH grant R01 CA85848-12 and the NIH/NIAID grant R37 AI031789-20.

The authors are grateful to the editor, an associate editor, and two anonymous referees for their valuable comments. The authors also would like to thank Dr. Lan Kong of Penn State College of Medicine and the CRISMA (Clinical Research, Investigation, and Systems Modeling of Acute Illness) Center at the University of Pittsburgh for providing the GenIMS data set.

References

1. Arunajadai SG, Rauh VA. Handling covariates subject to limits of detection in regression. Environ Ecol Stat. 2012; 157:369–391.
2. Austin PC, Brunner LJ. Type I error inflation in the presence of a ceiling effect. Am Stat. 2003; 57:97–104.
3. Austin PC, Hoch JS. Estimating linear regression models in the presence of a censored independent variable. Stat Med. 2004; 23:411–429. [PubMed: 14748036]
4. Clayton DG. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. Biometrika. 1978; 65:677–692.
5. Firth D. Bias reduction of maximum likelihood estimates. Biometrika. 1993; 23:27–38.
6. Giovanini, J. PhD thesis. Oregon State University: 2008. Generalized linear mixed models with censored covariates.
7. Helsel, DR. Statistics for censored environmental data using minitab and R. Wiley; New York: 2012.
8. Herring AH, Ibrahim JG. Likelihood-based methods for missing covariates in the Cox proportional hazards model. J Am Stat Assoc. 2001; 96:292–302.
9. Herring AH, Ibrahim JG, Lipsitz SR. Non-ignorable missing covariate data in survival analysis: a case-study of an international breast cancer study group trial. J R Stat Soc, Ser C, Appl Stat. 2004; 53:293–310.
10. Hofert, M.; Kojadinovic, I.; Maechler, M.; Yan, J. Copula: Multivariate dependence with copulas. R package version 0.999-5. 2012. http://CRAN.R-project.org/package=copula
11. Hornung RW, Reed LD. Estimation of average concentration in the presence of nondetectable values. Appl Occup Environ Hyg. 1990; 5:46–51.
12. Hughes JP. Mixed effects models with censored data with application to HIV RNA levels. Biometrics. 1999; 55:625–629. [PubMed: 11318225]
13. Ibrahim JG, Chen MH, Lipsitz SR. Bayesian methods for generalized linear models with covariates missing at random. Can J Stat. 2002; 30:55–78.
14. Ibrahim JG, Chen MH, Lipsitz SR, Herring AH. Missing-data methods for generalized linear models: a comparative review. J Am Stat Assoc. 2005; 100:332–346.
15. Kellum JA, Kong L, Fink MP, Weissfeld LA, Yealy DM, Pinsky MR, Fine J, Krichevsky A, Delude R, Angus D. Understanding the inflammatory cytokine response in pneumonia and sepsis. Arch Intern Med. 2007; 167:1655–1663. [PubMed: 17698689]
16. Lee M, Kong L, Weissfeld L. Multiple imputation for left-censored biomarker data based on Gibbs sampling method. Stat Med. 2012; 31:1838–1848. [PubMed: 22359320]

17. Lipsitz SR, Ibrahim JG, Chen MH, Peterson H. Non-ignorable missing covariates in generalized linear models. Stat Med. 1999; 18:2435–2448. [PubMed: 10474151]
18. Little RJA. Regression with missing x's: a review. J Am Stat Assoc. 1992; 87:1227–1237.
19. Lubin JH, Colt JS, Camann D, Davis S, Cerhan JR, Severson RK, Bernstein L, Hartge P. Epidemiologic evaluation of measurement data in the presence of detection limits. Environ Health Perspect. 2004; 112:1691–1696. [PubMed: 15579415]
20. Lyles RH, Lyles CM, Taylor DJ. Random regression models for human immunodeficiency virus ribonucleic acid data subject to left censoring and informative drop-outs. J R Stat Soc, Ser C, Appl Stat. 2000; 49:485–497.
21. Lyles RH, Fan D, Chauchoowon R. Correlation coefficient estimation involving a left censored laboratory assay variable. Stat Med. 2001; 20:2921–2933. [PubMed: 11568949]
22. Lynn HS. Maximum likelihood inference for left-censored HIV RNA data. Stat Med. 2001; 20:33–45. [PubMed: 11135346]
23. May RC, Ibrahim JG, Chu H. Maximum likelihood estimation in generalized linear models with multiple covariates subject to detection limits. Stat Med. 2011; 30:2551–2561. [PubMed: 21710558]
24. Moulton LH, Halsey NA. A mixture model with detection limits for regression analysis of antibody response to vaccine. Biometrics. 1995; 51:1570–1578. [PubMed: 8589241]
25. Nan B, Kalbfleisch JD, Yu M. Asymptotic theory for the semiparametric accelerated failure time model with missing data. Ann Stat. 2009; 37:2351–2376.
26. Nie L, Chu H, Liu C, Cole SR, Vexler A, Schisterman EF. Linear regression with an independent variable subject to a detection limit. Epidemiology. 2010; 21S:S17–S24. [PubMed: 21422965]
27. Paxton WB, Coombs RW, McElrath MJ, Keefer MC, Hughes J, Sinagil F, Chernoff D, Demeter L, Williams B, Corey L. Longitudinal analysis of quantitative virologic measures in human immunodeficiency virus-infected subjects with ≥ 400 cd4 lymphocytes: implications for applying measurements to individual patients. J Infect Dis. 1997; 175:247–254. [PubMed: 9203644]
28. Pettitt AN. Censored observations, repeated measures and mixed effects models—An approach using the em algorithm and normal errors. Biometrika. 1986; 73:635–643.
29. Piepho HP, Thoni H, Müller HM. Estimating the product-moment correlation in samples with censoring on both variables. Biom J. 2002; 44:657–670.
30. Rigobon R, Stoker TM. Estimation with censored regressors: basic issues. Int Econ Rev. 2007; 48:1441–1467.
31. Rigobon R, Stoker TM. Bias from censored regressors. J Bus Econ Stat. 2009; 27:340–353.
32. Robins JM, Wang N. Inference for imputation estimators. Biometrika. 2000; 87:113–124.
33. Rubin, DB. Multiple imputation for nonresponse In: Surveys. Wiley; New York: 1987.
34. Schaubel DE, Cai J. Multiple imputation methods for recurrent event data with missing event category. Can J Stat. 2006; 34:677–692.
35. Thiebaut R, Jacqmin-Gadda H, Babiker A, Commenges D. The CASCADE Collaboration. Joint modelling of bivariate longitudinal data with informative dropout and left-censoring, with application to the evolution of cd4+ cell count and HIV RNA viral load in response to treatment of HIV infection. Stat Med. 2005; 24:65–82. [PubMed: 15523706]
36. Tsiatis, AA. Semiparametric theory and missing data. Springer; Berlin: 2006.
37. Tsimikas JV, Bantis LE, Georgiou SD. Inference in generalized linear regression models with a censored covariate. Comput Stat Data Anal. 2012; 56:1854–1868.
38. Wang H, Feng X. Multiple imputation for m-regression with censored covariates. J Am Stat Assoc. 2012; 107:194–204.
39. Wang H, Fygenson M. Inference for censored quantile regression models in longitudinal studies. Ann Stat. 2009; 37:756–781.
40. Wang N, Robins JM. Large-sample theory for parametric multiple imputation procedures. Biometrika. 1998; 84:935–948.
41. Wu H, Chen Q, Ware LB, Koyama T. A Bayesian approach for generalized linear models with explanatory biomarker measurement variables subject to detection limit—An application to acute lung injury. Appl Stat. 2012; 39:1733–1747.

42. Wu L. A joint model for nonlinear mixed-effects models with censoring and covariates measured with error, with applications to aids studies. J Am Stat Assoc. 2002; 97:955–964.

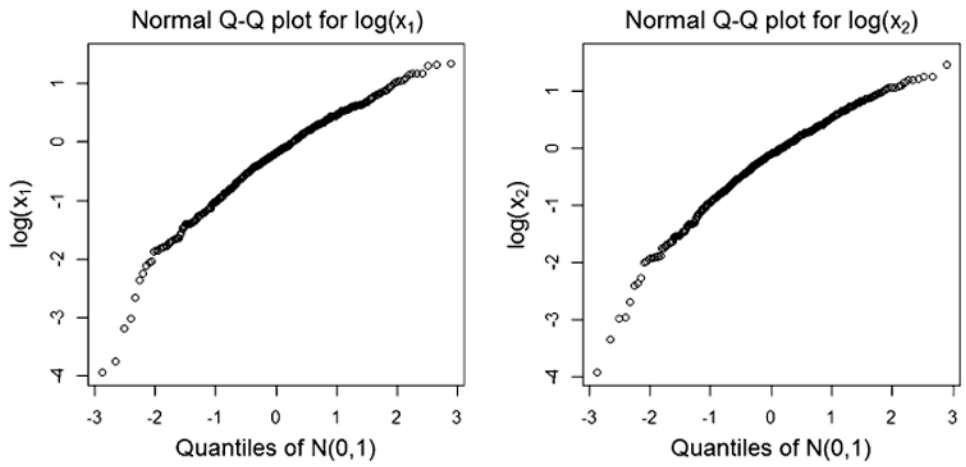

Fig. 1. Normal Q-Q plots for the latent log (x_1) and log (x_2) using a single simulated data set

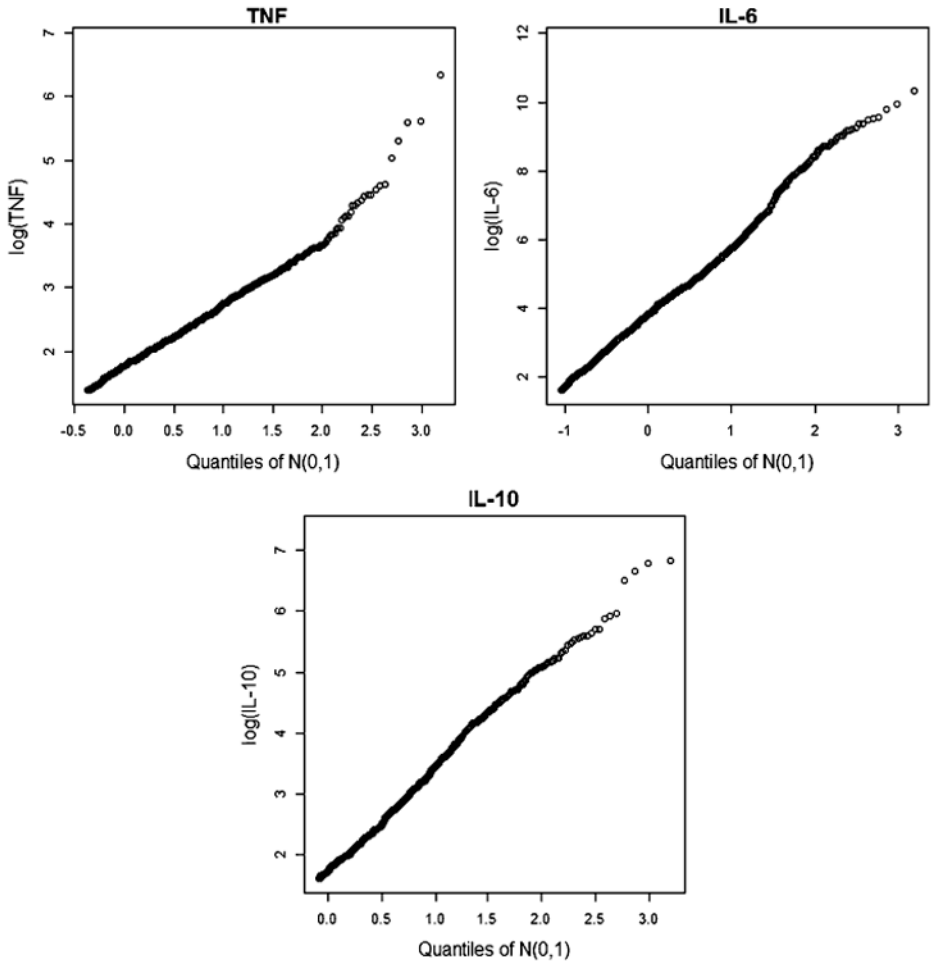

Fig. 2.

Normal Q-Q plots for censored cytokines: log (TNF), log (IL-6), and log (IL-10). The Y-axis corresponds to the sample quantiles above the p_0th quantile of the cytokine biomarkers, where p_0 is the sample censoring proportion. The X-axis corresponds to the quantiles of an N (0,1)

Table 1

Simulation results for a logistic regression on 2000 data sets with 200 observations and two covariates of correlation $\rho = 0.5$

Cens. %	Method	$\beta_0 = 3$				$\beta_1 = 1.5$				$\beta_2 = -3$				Time (s)
		Bias	SD	SE	CP	Bias	SD	SE	CP	Bias	SD	SE	CP	
20	Omni	0.00	0.43	0.44	0.95	0.01	0.36	0.37	0.96	0.00	0.51	0.52	0.96	< 0.1
	CC	0.01	0.48	0.48	0.95	−0.01	0.44	0.44	0.95	0.00	0.56	0.56	0.95	< 0.1
	SUB	−0.04	0.43	0.43	0.94	0.01	0.37	0.37	0.96	0.04	0.51	0.51	0.94	< 0.1
	Mean1	0.01	0.45	0.45	0.95	−0.08	0.34	0.35	0.94	0.03	0.52	0.52	0.95	0.8
	Mean2	0.02	0.45	0.45	0.95	0.06	0.40	0.38	0.95	−0.03	0.54	0.52	0.95	2.3
	ML	−0.01	0.43	0.43	0.95	−0.02	0.37	0.37	0.95	0.03	0.52	0.52	0.95	41.0
	FB	0.03	0.45	0.46	0.96	0.03	0.39	0.39	0.96	−0.03	0.53	0.54	0.96	56.1
	PMI	0.01	0.44	0.45	0.95	0.01	0.38	0.38	0.96	0.00	0.52	0.53	0.95	26.2
	IMI	−0.01	0.44	0.45	0.95	−0.01	0.38	0.38	0.96	0.02	0.52	0.53	0.95	1.0
40	Omni	0.00	0.43	0.44	0.95	0.01	0.36	0.37	0.96	0.00	0.51	0.52	0.96	< 0.1
	CC	0.02	0.58	0.58	0.96	−0.01	0.54	0.53	0.95	−0.01	0.65	0.64	0.96	< 0.1
	SUB	−0.14	0.41	0.41	0.90	0.05	0.39	0.40	0.97	0.10	0.48	0.49	0.93	< 0.1
	Mean1	−0.07	0.46	0.44	0.92	−0.15	0.34	0.35	0.91	0.14	0.53	0.51	0.90	1.2
	Mean2	0.02	0.50	0.45	0.93	0.16	0.47	0.40	0.93	−0.06	0.60	0.53	0.93	3.4
	ML	−0.01	0.46	0.47	0.95	−0.01	0.39	0.39	0.95	0.02	0.54	0.54	0.95	56.6
	FB	0.04	0.47	0.47	0.95	0.03	0.41	0.41	0.95	−0.04	0.56	0.53	0.94	71.0
	PMI	0.01	0.47	0.46	0.95	0.02	0.41	0.40	0.96	−0.01	0.56	0.55	0.95	45.8
	IMI	0.00	0.47	0.46	0.95	0.01	0.41	0.40	0.96	0.00	0.55	0.55	0.95	1.1
60	Omni	0.00	0.43	0.44	0.95	0.01	0.36	0.37	0.96	0.00	0.51	0.52	0.96	< 0.1
	CC	0.03	0.77	0.78	0.96	-0.02	0.69	0.67	0.95	0.00	0.77	0.77	0.95	< 0.1
	SUB	−0.14	0.37	0.38	0.84	0.14	0.45	0.45	0.96	0.11	0.47	0.48	0.92	< 0.1
	Mean1	−0.31	0.44	0.39	0.77	−0.23	0.33	0.34	0.88	0.41	0.49	0.45	0.74	1.6
	Mean2	0.13	0.54	0.43	0.85	0.24	0.55	0.43	0.90	−0.04	0.63	0.51	0.89	4.2
	ML	−0.01	0.49	0.49	0.95	−0.02	0.43	0.44	0.96	0.03	0.58	0.59	0.96	103.2
	FB	0.03	0.52	0.51	0.95	0.02	0.46	0.44	0.94	−0.03	0.61	0.60	0.94	79.9
	PMI	0.02	0.53	0.51	0.94	0.03	0.46	0.44	0.95	−0.02	0.62	0.60	0.95	62.1
	IMI	−0.02	0.50	0.50	0.94	0.02	0.47	0.47	0.96	0.02	0.60	0.61	0.95	1.3

Cens. %: percent of individuals with at least one censored covariate; SD: simulation standard deviation of $\hat{\beta}$; SE: average estimated standard error; CP: empirical 95 % coverage probability; Time (s): average elapsed computational time in seconds for a single simulated data set; Omni: omniscient method; CC: complete case analysis; SUB: substitution by DL/√2; Mean1: conditional mean imputation method conditioning on the covariates only; Mean2: conditional mean imputation method conditioning on the response and covariates; ML: maximum likelihood method; FB: fully Bayesian method; PMI: proper (Bayesian) multiple imputation method; IMI: improper (frequentist) multiple imputation method

Table 2

A comparison of the standard error estimates for the proposed improper multiple imputation estimator based on simulation results for 2000 data sets with 200 observations and two covariates of correlation $\rho = 0.5$. Standard errors of the estimates in the table range between 0.001 and 0.003

	20 % Censoring			40 % Censoring			60 % Censoring		
	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
SD	0.440	0.377	0.520	0.469	0.407	0.554	0.500	0.467	0.600
SE_C	0.449	0.382	0.526	0.464	0.402	0.550	0.496	0.464	0.598
SE_A	0.449	0.382	0.525	0.464	0.403	0.550	0.502	0.468	0.606
SE_R	0.447	0.381	0.522	0.458	0.392	0.542	0.472	0.418	0.565
SE_B	0.469	0.401	0.538	0.512	0.433	0.605	0.571	0.506	0.675

SD: simulation standard deviation of $\hat{\boldsymbol{\beta}}_{\mathrm{IMI}}$; SE_C: estimated using (8); SE_A: estimated using (9); SE_R: estimated using (6); SE_B: estimated via a bootstrap procedure

Table 3

Simulation results for a logistic regression on 2000 data sets with 200 individuals and an incorrect distributional assumption on the covariates

Cens. %	Method	$\beta_0 = 3$				$\beta_1 = 1.5$				$\beta_2 = -3$			
		Bias	SD	SE	CP	Bias	SD	SE	CP	Bias	SD	SE	CP
20	Omni	0.00	0.45	0.44	0.95	0.00	0.65	0.64	0.95	0.00	0.73	0.73	0.95
	CC	−0.01	0.48	0.46	0.94	0.02	0.72	0.71	0.96	0.00	0.79	0.79	0.95
	ML	0.01	0.46	0.45	0.94	−0.02	0.67	0.66	0.95	−0.02	0.75	0.74	0.95
	FB	0.04	0.47	0.46	0.96	−0.02	0.72	0.67	0.95	−0.04	0.77	0.75	0.95
	PMI	0.01	0.46	0.45	0.94	0.00	0.68	0.64	0.95	−0.02	0.75	0.73	0.95
	IMI	0.01	0.46	0.45	0.94	−0.02	0.68	0.64	0.95	−0.02	0.75	0.73	0.95
40	Omni	0.00	0.45	0.44	0.95	0.00	0.65	0.64	0.95	0.00	0.73	0.73	0.95
	CC	0.00	0.59	0.59	0.95	0.02	0.92	0.91	0.95	−0.01	0.96	0.89	0.95
	ML	0.05	0.47	0.44	0.94	−0.05	0.66	0.63	0.94	−0.06	0.74	0.70	0.94
	FB	0.08	0.50	0.47	0.94	−0.06	0.72	0.69	0.94	−0.08	0.83	0.79	0.95
	PMI	0.04	0.47	0.45	0.95	−0.04	0.69	0.66	0.95	−0.06	0.77	0.74	0.95
	IMI	0.04	0.47	0.45	0.95	−0.05	0.70	0.68	0.94	−0.06	0.76	0.74	0.95
60	Omni	0.00	0.45	0.44	0.95	0.00	0.65	0.64	0.95	0.00	0.73	0.73	0.95
	CC	−0.01	0.82	0.78	0.94	0.02	1.24	1.23	0.96	−0.01	1.18	1.14	0.96
	ML	0.07	0.47	0.45	0.94	−0.13	0.66	0.64	0.93	−0.13	0.76	0.73	0.93
	FB	0.12	0.52	0.50	0.94	−0.10	0.76	0.75	0.93	−0.15	0.86	0.83	0.93
	PMI	0.08	0.50	0.46	0.94	−0.08	0.75	0.71	0.93	−0.11	0.82	0.78	0.94
	IMI	0.07	0.48	0.47	0.94	−0.08	0.78	0.75	0.93	−0.10	0.79	0.81	0.95

Cens. %: percent of individuals with at least one censored covariate; SD: simulation standard deviation of $\hat{\beta}$; SE: average estimated standard error; CP: empirical 95 % coverage probability; Omni: omniscient method; CC: complete case analysis; ML: maximum likelihood method; FB: fully Bayesian method; PMI: proper multiple imputation method; IMI: improper multiple imputation method

Table 4

Estimates and standard errors for coefficients of log-cytokine and demographic covariates in the logistic regression model for 90-day survival, based on the GenIMS data set

	CC		ML		FB		PMI		IMI	
	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE
Int.	8.48	1.11	6.71	0.54	6.89	0.56	6.72	0.56	6.71	0.56
TNF	−0.25	0.16	−0.08	0.07	−0.10	0.08	−0.09	0.09	−0.09	0.09
IL-6	−0.08	0.07	−0.05	0.04	−0.04	0.04	−0.05	0.05	−0.04	0.04
IL-10	−0.05	0.11	−0.08	0.05	−0.06	0.06	−0.07	0.06	−0.08	0.06
Sex	−0.18	0.25	−0.42	0.15	−0.41	0.16	−0.42	0.16	−0.42	0.16
Race	−0.29	0.50	−0.31	0.27	−0.34	0.28	−0.32	0.28	−0.32	0.28
Age	−0.07	0.01	−0.05	0.01	−0.06	0.01	−0.05	0.01	−0.05	0.01

CC: complete case analysis; ML: maximum likelihood method; FB: fully Bayesian method; PMI: proper multiple imputation method; IMI: improper multiple imputation method