


RESEARCH ARTICLE

Estimation in generalized linear models under censored covariates with an application to MIREC data

Wan-Chen Lee¹  | Sanjoy K. Sinha² | Tye E. Arbuckle¹ | Mandy Fisher¹

¹Environmental Health Science and Research Bureau, Health Canada, Ottawa, Canada

²School of Mathematics and Statistics, Carleton University, Ottawa, Canada

Correspondence

Wan-Chen Lee, 101 Tunney's Pasture Driveway, Tunney's Pasture, AL 0201A, Ottawa, ON K1A 0K9, Canada.
Email: ophelia0825@gmail.com

Funding information

Health Canada's Chemicals Management Plan; Canadian Institute of Health Research, Grant/Award Number: MOP-81285; Ontario Ministry of the Environment; Natural Sciences and Engineering Research Council of Canada

In many biological experiments, certain values of a biomarker are often non-detectable due to low concentrations of an analyte or the limitations of a chemical analysis device, resulting in left-censored values. There is an increasing demand for the analysis of data subject to detection limits in clinical and environmental studies. In this paper, we develop a novel statistical method for the maximum likelihood estimation in generalized linear models with covariates subject to detection limits. Simulations are carried out to study the relative performance of the proposed estimators, as compared to other existing estimators. The proposed method is also applied to a real dataset from the Maternal-Infant Research on Environmental Chemicals cohort study, where we investigate how different chemical mixtures affect the health outcomes of infants and pregnant women.

KEYWORDS

generalized linear model, limit of detection, logistic regression, maximum likelihood estimate

1 | INTRODUCTION

In many clinical studies, it is common to have biomarkers that are measured with detection limits. “Nondetects” (samples that, for various reasons, have undetectable concentrations of the analyte) are often due to low-level concentrations of biomarkers with values known only up to the laboratory's detection limits. The problem of nondetects in bioassays often translates into the problem of left-censored covariates in the regression analysis. The simplest approach to deal with the left censoring is the complete-case analysis in which we remove all observations falling below the limit of detection (LOD) and perform a standard analysis based on the “truncated data.” Such analysis is not generally recommended due to the loss of useful information in the data. Another approach is the substitution method in which the nondetect values of a covariate are replaced by the LOD, LOD/2, or LOD/ $\sqrt{2}$. Such methods are sometimes employed as they are easy to implement and also simple to understand. However, Cole et al¹ demonstrated in a simulation study that, as the proportion of nondetects increases, replacing the left-censored values by LOD, LOD/2, or LOD/ $\sqrt{2}$ results in increasingly biased estimators of model parameters and also produces increasingly poor coverage probabilities of confidence intervals. Thompson and Nelson² found that replacing the left-censored values by half the detection limit led to biased parameter estimators and also artificially small standard errors of the estimators. These studies clearly provide evidence against any use of ad hoc substitution methods when analyzing data with detection limits.

Helsel³ reviewed a number of existing methods for dealing with censored observations commonly encountered in clinical and environmental studies, which include the nonparametric Kaplan-Meier method for determining

.....
The copyright line for this article was changed on 1 October 2018 after original online publication.

© 2018 Her Majesty the Queen in Right of Canada Statistics in Medicine © 2018 John Wiley & Sons Ltd.

Reproduced with the permission of Health Canada, Government of Canada.

descriptive statistics and regression on order statistics for imputing the nondetects. Analysis of data subject to detection limits has been extensively studied in the literature in recent years (eg, other works⁴⁻¹³). Helsel⁴ reviewed existing statistical methods for dealing with nondetects in environmental data. Herring⁵ proposed a nonparametric Bayesian approach for model selection and for handling truncation of exposures to complex chemical mixtures and health outcomes at limits of detection in the framework of a complex hierarchical model. May et al⁶ proposed a Monte Carlo version of the expectation-maximization algorithm to handle a large number of left-censored predictors in generalized linear models, which required intensive computation. Satter et al⁷ discussed a likelihood method for estimation and inference with a parametric proportional hazards model, where specific values of some biomarkers are left censored due to detection limits. Satter et al⁸ proposed a flexible semiparametric approach for estimation in frailty models with left-censored covariates. Bernhardt et al⁹ developed a multiple imputation approach for analyzing data with multiple predictors subject to detection limits in the context of generalized linear models.

In the setting of linear regression models, substitution methods have also been used for cases when a single covariate is subject to a detection limit. Richardson and Ciampi¹⁴ considered replacing left-censored values of a covariate by the conditional expected value of the censored covariate given all observed values of the covariates. This method requires specification of the underlying covariate distribution, which, in practice, may not be known with certainty. Schisterman et al¹⁵ considered an unknown covariate distribution and proposed substituting the average of all observed values of the left-censored covariate in the regression model, which was shown to provide unbiased estimates of the model parameters. To deal with left-censored covariates, the maximum likelihood method is often used, which also requires specification of the covariate distribution. Nie et al¹⁶ compared these methods with the aforementioned substitution methods, when a single covariate is subject to an LOD, where the maximum likelihood method appeared to perform the best when the covariate distribution was known. The study concludes that the maximum likelihood method achieves unbiased and efficient estimates of regression parameters under a known covariate distribution of a left-censored covariate.

This research was motivated by an epidemiologic study, referred to as the Maternal-Infant Research on Environmental Chemicals (MIREC) study,¹⁷ which is a large cohort study of pregnant women and their newborns in Canada. The study was established to obtain biomonitoring data on pregnant women and infants to examine potential adverse effects of prenatal exposure to environmental chemicals on pregnancy and infant health. Pregnancy and infant health outcomes are usually dichotomous and a common feature of the exposure data on the chemical mixtures is that a large proportion of the data are truncated at the detection limits. There were 81 available chemicals in the MIREC dataset. Since not every chemical is highly linearly correlated to other chemicals, the multiple imputation approach is not applicable in this case. To investigate the effects of exposures to chemical mixtures on health outcomes, we developed a likelihood approach in the framework of generalized linear models by addressing the issue of nondetects. The fundamental idea of our proposed method is to take into account all possible values below LOD. Specifically, when studying the effects of covariates with nondetects, we incorporated a weight function into the observed data likelihood function, where the weights are obtained from a multivariate distribution of the indicators of nondetects. The analysis of the MIREC data is discussed further in the Application section.

The paper is organized as follows. Section 2 introduces the model and notation and discusses the proposed maximum likelihood approach for estimation in generalized linear models with covariates subject to detection limits. Section 3 presents an illustrative example to describe the computational issues of the proposed maximum likelihood method using a simple logistic regression model. Section 4 studies the performance of the proposed estimators based on a series of Monte Carlo simulations, where the empirical biases and mean squared errors (MSEs) of the regression estimators are presented for scenarios where the response and left-censored covariates are assumed to follow different known distributions including Bernoulli and Gaussian distributions for discrete and continuous outcomes, and Gaussian and gamma distributions for left-censored covariates. Empirical results for the nuisance parameters are shown in the Appendix. Section 5 presents an application of the proposed method using the MIREC data introduced earlier. Section 6 concludes the paper with some discussions.

2 | MODEL, NOTATION, AND METHOD

2.1 | Complete data

Suppose the elements of the observed response vector $\mathbf{y} = (y_1, \dots, y_n)'$ are independent and follow a distribution in the exponential family¹⁸

$$f_{y_i|x_i}(y_i|\mathbf{x}_i, \boldsymbol{\beta}, \phi) = \exp \left\{ \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\}, \quad (1)$$

for some functions a , b , and c , where the canonical parameter $\theta_i = \mathbf{x}_i' \boldsymbol{\beta}$, with \mathbf{x}_i' being the i th row of the design matrix \mathbf{X} for the fixed effects, which may contain 1 to incorporate an intercept term. The log-likelihood function for (1) is obtained as

$$l(\boldsymbol{\beta}, \phi | \mathbf{y}, \mathbf{X}) = \sum_{i=1}^n \left\{ \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\}. \quad (2)$$

In many practical situations including binary and Poisson regression models, the dispersion parameter ϕ is fixed at unity. Therefore, we choose $\phi = 1$, for simplicity. In some situations, however, it may be necessary to estimate ϕ as a dispersion parameter of the marginal distribution of the response vector \mathbf{y} . From (2), the maximum likelihood estimating equation for $\boldsymbol{\beta}$ is given by

$$\sum_{i=1}^n \{y_i - \mu_i(\boldsymbol{\beta}, \mathbf{x}_i)\} \mathbf{x}_i = \mathbf{0}, \quad (3)$$

where $\mu_i(\boldsymbol{\beta}, \mathbf{x}_i)$ is the i th mean response, $\mu_i(\boldsymbol{\beta}, \mathbf{x}_i) = E(y_i | \mathbf{x}_i, \boldsymbol{\beta}) = b'(\theta_i)$. Equation (3) can be solved numerically using an iterative method, such as the iteratively reweighted least squares method, given by

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \{\mathbf{X}' \mathbf{W}(\hat{\boldsymbol{\beta}}^{(k)}) \mathbf{X}\}^{-1} \mathbf{X}' \mathbf{W}(\hat{\boldsymbol{\beta}}^{(k)}) \mathbf{z}(\hat{\boldsymbol{\beta}}^{(k)}), \quad (4)$$

for $k = 0, 1, 2, \dots$, where $\mathbf{W}(\boldsymbol{\beta})$ is an $n \times n$ diagonal matrix with its i th diagonal element $w_i = \text{var}(y_i)$ and $\mathbf{z}(\boldsymbol{\beta}) = (z_1, \dots, z_n)'$ is a vector of “pseudo-observations” with its i th element $z_i = \theta_i + (y_i - \mu_i)/\text{var}(y_i)$. An approximate variance of the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ may be obtained as

$$\text{var}(\hat{\boldsymbol{\beta}}) = \{\mathbf{X}' \mathbf{W}(\hat{\boldsymbol{\beta}}) \mathbf{X}\}^{-1}.$$

The estimator $\hat{\boldsymbol{\beta}}$ has an asymptotic normal distribution with mean $\boldsymbol{\beta}$ and variance $\text{var}(\hat{\boldsymbol{\beta}})$. In the next section, we consider the fact that some covariates are measured with the LOD resulting in left-censored values. We address this issue of left-censoring when finding the maximum likelihood estimators of the model parameters.

2.2 | Estimation with left-censored covariates

Let $\{(y_i, \mathbf{x}_i); i = 1, \dots, n\}$ denote the data that would occur in the absence of censored values. To denote the censoring status, consider a vector of indicator variables $\mathbf{v}_i = (v_{i1}, \dots, v_{ip})'$ whose j th element v_{ij} is 1 if the corresponding value of the covariate x_{ij} is observed (ie, if $x_{ij} \geq c_j$) and v_{ij} is 0 if x_{ij} is left censored (ie, if $x_{ij} \leq c_j$), where c_j 's are known constants. We assume that the marginal distribution of the j th binary indicator, ie, v_{ij} , is Bernoulli with “success” probability $\pi_{ij} = P(v_{ij} = 1) = P(x_{ij} \geq c_j)$. To define the joint distribution of the p binary indicators (v_{i1}, \dots, v_{ip}) , we consider a Bahadur type multivariate binary distribution.¹⁹ For example, when $p = 3$, the Bahadur multivariate density of $\mathbf{v}_i = (v_{i1}, v_{i2}, v_{i3})'$ has the form

$$f_{\mathbf{v}_i}(\mathbf{v}_i | \boldsymbol{\tau}) = \left\{ \prod_{j=1}^3 \pi_{ij}^{v_{ij}} (1 - \pi_{ij})^{(1-v_{ij})} \right\} \{1 + \rho_{12} z_{i1} z_{i2} + \rho_{13} z_{i1} z_{i3} + \rho_{23} z_{i2} z_{i3} + \rho_{123} z_{i1} z_{i2} z_{i3}\}, \quad (5)$$

where

$$z_{ij} = \frac{v_{ij} - \pi_{ij}}{\sqrt{\pi_{ij}(1 - \pi_{ij})}},$$

$$\rho_{jk} = \text{corr}(v_{ij}, v_{ik}) = \frac{E\{(v_{ij} - \pi_{ij})(v_{ik} - \pi_{ik})\}}{\sqrt{\pi_{ij}(1 - \pi_{ij})\pi_{ik}(1 - \pi_{ik})}},$$

$$\rho_{123} = \frac{E\{(v_{i1} - \pi_{i1})(v_{i2} - \pi_{i2})(v_{i3} - \pi_{i3})\}}{\sqrt{\pi_{i1}(1 - \pi_{i1})\pi_{i2}(1 - \pi_{i2})\pi_{i3}(1 - \pi_{i3})}},$$

for $j, k = 1, 2, 3$. Let $\mathbf{x}_{\text{obs},i}$ denote the observed values and $\mathbf{x}_{\text{lod},i}$ the left-censored values of \mathbf{x}_i . Assuming arbitrary, non-monotone patterns of censoring for \mathbf{x}_i , some permutation of the indices of \mathbf{x}_i may be written as $\mathbf{x}_i = (\mathbf{x}_{\text{obs},i}, \mathbf{x}_{\text{lod},i})$, where $\mathbf{x}_{\text{lod},i}$ is a $p_i \times 1$ vector of left-censored values of \mathbf{x}_i . The vector of covariates \mathbf{x}_i is assumed to follow a density function $f_{\mathbf{x}_i}(\mathbf{x}_i | \boldsymbol{\alpha})$, depending on parameters $\boldsymbol{\alpha}$. Our focus is on the regression parameters $\boldsymbol{\beta}$, with $\boldsymbol{\alpha}$ and $\boldsymbol{\tau}$ being viewed as nuisance parameters.

For the i th observation, the actual observed data consist of values of the variables $(y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i)$. The distribution of the observed data is obtained by integrating $\mathbf{x}_{\text{lod},i}$ out of the joint density of $(y_i, \mathbf{x}_i, \mathbf{v}_i)$, that is,

$$f_{y_i, \mathbf{x}_i, \mathbf{v}_i}(y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i | \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\tau}) = \int_{-\infty}^{\mathbf{c}} f_{y_i | \mathbf{x}_i}(y_i | \mathbf{x}_{\text{obs},i}, \mathbf{x}_{\text{lod},i}, \boldsymbol{\beta}) f_{\mathbf{x}_i}(\mathbf{x}_{\text{obs},i}, \mathbf{x}_{\text{lod},i} | \boldsymbol{\alpha}) f_{\mathbf{v}_i}(\mathbf{v}_i | \boldsymbol{\tau}) d\mathbf{x}_{\text{lod},i}, \quad (6)$$

where \mathbf{c} is a vector of p_i elements representing the upper limits of integration, obtained from the corresponding LOD values of the left-censored covariates $\mathbf{x}_{\text{lod},i}$. The full likelihood of $\boldsymbol{\gamma} = (\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\tau})$ is any function of $\boldsymbol{\gamma}$ proportional to the products of (6) for all n observations

$$L_{\text{full}}(\boldsymbol{\gamma} | \mathbf{y}, \mathbf{X}_{\text{obs}}, \mathbf{v}) \propto \prod_{i=1}^n f_{y_i, \mathbf{x}_i, \mathbf{v}_i}(y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i | \boldsymbol{\gamma}), \quad (7)$$

where $\mathbf{X}_{\text{obs}} = \{\mathbf{x}_{\text{obs},i}; i = 1, \dots, n\}$ and $\mathbf{v} = \{\mathbf{v}_i; i = 1, \dots, n\}$. This likelihood cannot usually be evaluated in a closed form because the density $f_{y_i, \mathbf{x}_i, \mathbf{v}_i}(y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i | \boldsymbol{\gamma})$ of the observed data for the i th unit has an integral with dimension equal to the dimension of $\mathbf{x}_{\text{lod},i}$. The maximum likelihood estimators of $\boldsymbol{\gamma}$ may be obtained by numerically maximizing the full likelihood function (7). We develop an iterative algorithm to calculate the maximum likelihood estimators. For this, we can write the likelihood score function for $\boldsymbol{\gamma}$ in the form

$$\begin{aligned} U(\boldsymbol{\gamma}) &= \sum_{i=1}^n \frac{\partial}{\partial \boldsymbol{\gamma}} \log \int_{-\infty}^{\mathbf{c}} f_{y_i | \mathbf{x}_i}(y_i | \mathbf{x}_{\text{obs},i}, \mathbf{x}_{\text{lod},i}, \boldsymbol{\beta}) f_{\mathbf{x}_i}(\mathbf{x}_{\text{obs},i}, \mathbf{x}_{\text{lod},i} | \boldsymbol{\alpha}) f_{\mathbf{v}_i}(\mathbf{v}_i | \boldsymbol{\tau}) d\mathbf{x}_{\text{lod},i} \\ &= \sum_{i=1}^n \int_{-\infty}^{\mathbf{c}} B(\boldsymbol{\gamma}, \mathbf{x}_{\text{lod},i}) f_{\mathbf{x}_i | y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i}(\mathbf{x}_{\text{lod},i} | y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i; \boldsymbol{\gamma}) d\mathbf{x}_{\text{lod},i}, \end{aligned} \quad (8)$$

where $B(\boldsymbol{\gamma}, \mathbf{x}_{\text{lod},i})$ represents the “complete data” score vector, given by

$$B(\boldsymbol{\gamma}, \mathbf{x}_{\text{lod},i}) = \frac{\partial}{\partial \boldsymbol{\gamma}} \left\{ \log f_{y_i | \mathbf{x}_i}(y_i | \mathbf{x}_{\text{obs},i}, \mathbf{x}_{\text{lod},i}, \boldsymbol{\beta}) + \log f_{\mathbf{x}_i}(\mathbf{x}_{\text{obs},i}, \mathbf{x}_{\text{lod},i} | \boldsymbol{\alpha}) + \log f_{\mathbf{v}_i}(\mathbf{v}_i | \boldsymbol{\tau}) \right\},$$

and $f_{\mathbf{x}_i | y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i}(\mathbf{x}_{\text{lod},i} | y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i; \boldsymbol{\gamma})$ is the conditional density of the vector of covariates $\mathbf{x}_{\text{lod},i}$, given the observed data $(y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i)$. This density function does not have a closed form, and numerical methods may be needed to calculate the log-likelihood function, score function, and Fisher information. Given some initial estimates $\boldsymbol{\gamma}^{(0)}$, we can find the maximum likelihood estimator of $\boldsymbol{\gamma}$ by solving the Newton-Raphson iterative equations

$$\boldsymbol{\gamma}^{(k+1)} = \boldsymbol{\gamma}^{(k)} - \left\{ U^{(1)}(\boldsymbol{\gamma}^{(k)}) \right\}^{-1} U(\boldsymbol{\gamma}^{(k)}), \quad (9)$$

for $k = 0, 1, 2, \dots$, where $U(\boldsymbol{\gamma}^{(k)})$ is the likelihood score function $U(\boldsymbol{\gamma})$ evaluated at $\boldsymbol{\gamma}^{(k)}$ and $U^{(1)}(\boldsymbol{\gamma}^{(k)})$ is the first derivative of the score function $U(\boldsymbol{\gamma})$ with respect to $\boldsymbol{\gamma}$, evaluated at $\boldsymbol{\gamma}^{(k)}$. Here, the derivative of the score function provides the observed Fisher information matrix defined by $I(\boldsymbol{\gamma}) = -U^{(1)}(\boldsymbol{\gamma})$. After some algebra, this Fisher information may be expressed in the form

$$\begin{aligned} -I(\boldsymbol{\gamma}) &= \sum_{i=1}^n \int_{-\infty}^{\mathbf{c}} \frac{\partial B(\boldsymbol{\gamma}, \mathbf{x}_{\text{lod},i})}{\partial \boldsymbol{\gamma}} f_{\mathbf{x}_i | y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i}(\mathbf{x}_{\text{lod},i} | y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i; \boldsymbol{\gamma}) d\mathbf{x}_{\text{lod},i} \\ &\quad + \sum_{i=1}^n \int_{-\infty}^{\mathbf{c}} B(\boldsymbol{\gamma}, \mathbf{x}_{\text{lod},i}) B'(\boldsymbol{\gamma}, \mathbf{x}_{\text{lod},i}) f_{\mathbf{x}_i | y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i}(\mathbf{x}_{\text{lod},i} | y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i; \boldsymbol{\gamma}) d\mathbf{x}_{\text{lod},i} \\ &\quad - \sum_{i=1}^n \int_{-\infty}^{\mathbf{c}} B(\boldsymbol{\gamma}, \mathbf{x}_{\text{lod},i}) f_{\mathbf{x}_i | y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i}(\mathbf{x}_{\text{lod},i} | y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i; \boldsymbol{\gamma}) d\mathbf{x}_{\text{lod},i} \\ &\quad \times \int_{-\infty}^{\mathbf{c}} B'(\boldsymbol{\gamma}, \mathbf{x}_{\text{lod},i}) f_{\mathbf{x}_i | y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i}(\mathbf{x}_{\text{lod},i} | y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i; \boldsymbol{\gamma}) d\mathbf{x}_{\text{lod},i}. \end{aligned}$$

Here, the integrations involving the score and Fisher information may be performed numerically using existing software. We use the R function “integrate” for the numerical integration. Based on some initial estimate $\boldsymbol{\gamma}^{(0)}$ of $\boldsymbol{\gamma}$, we continue iterations (9) until a convergence is met. The initial estimates may be chosen as the ordinary maximum likelihood estimates for “complete case” data with no LOD observations. At convergence, we obtain the maximum likelihood estimators of the parameters $\boldsymbol{\gamma}$, denoted by $\hat{\boldsymbol{\gamma}}$. The large-sample variance-covariance matrix of the maximum likelihood estimator $\hat{\boldsymbol{\gamma}}$ may be obtained from the observed Fisher information given by $\text{Var}(\hat{\boldsymbol{\gamma}}) = I^{-1}(\boldsymbol{\gamma})$, which may be approximated by evaluating the variance function at the likelihood estimator $\hat{\boldsymbol{\gamma}}$ as $\widehat{\text{Var}}(\hat{\boldsymbol{\gamma}}) = I^{-1}(\hat{\boldsymbol{\gamma}})$.

2.3 | Asymptotics

Under the assumption that the response and covariate distributions are correctly specified, it is typically the case that, as the sample size n increases, the maximum likelihood estimator $\hat{\boldsymbol{\gamma}}$ is consistent and follows an asymptotic normal distribution with mean vector $\boldsymbol{\gamma}$ and covariance matrix $I^{-1}(\boldsymbol{\gamma})$

$$\hat{\boldsymbol{\gamma}} \sim N(\boldsymbol{\gamma}, I^{-1}(\boldsymbol{\gamma})),$$

where $I(\boldsymbol{\gamma})$ is the Fisher information as defined earlier. We study finite-sample properties of the maximum likelihood estimators using Monte Carlo simulations in Section 4. The empirical results justify the use of normal theory inference procedure for the maximum likelihood estimators in the setting of generalized linear models with left-censored covariates.

3 | ILLUSTRATIVE EXAMPLE

This section provides some computational details for fitting a simple binary regression model with covariates that are subject to the LOD. Consider a binary model with two covariates x_1 and x_2

$y_i | x_{i1}, x_{i2} \sim$ independent Bernoulli (μ_i), $i = 1, \dots, n$,

$$\theta_i = \log\left(\frac{\mu_i}{1 - \mu_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}. \quad (10)$$

In this setup, $E(y_i) = \mu_i$ and $\text{var}(y_i) = \mu_i(1 - \mu_i)$. Suppose the covariates x_{i1} and x_{i2} are left censored due to the LOD. The covariates are assumed to follow a bivariate normal distribution with means $E(x_{i1}) = \mu_{x_1}$ and $E(x_{i2}) = \mu_{x_2}$, and with corresponding variances $\text{var}(x_{i1}) = \sigma_{x_1}^2$, $\text{var}(x_{i2}) = \sigma_{x_2}^2$, and $\text{cov}(x_{i1}, x_{i2}) = \sigma_{x_1 x_2}$. Let v_{ij} ($j = 1, 2$) denote an indicator variable, with $v_{ij} = 1$ when x_{ij} is observed (ie, $x_{ij} \geq c_j$) and $v_{ij} = 0$ when x_{ij} is left censored (ie, $x_{ij} \leq c_j$), with $E(v_{ij}) = \pi_{ij} = P(x_{ij} \geq c_j) = 1 - \Phi((c_j - \mu_{x_j})/\sigma_{x_j})$, where Φ is the cumulative distribution function of the standard normal distribution. We further assume that the vector of indicator variables $\mathbf{v}_i = (v_{i1}, v_{i2})'$ follows a bivariate Bahadur model

$$f_{\mathbf{v}_i}(\mathbf{v}_i | \boldsymbol{\alpha}, \rho) = \left\{ \prod_{j=1}^2 \pi_{ij}^{v_{ij}} (1 - \pi_{ij})^{(1-v_{ij})} \right\} \left(1 + \rho \frac{(v_{i1} - \pi_{i1})(v_{i2} - \pi_{i2})}{\sqrt{\pi_{i1}\pi_{i2}(1 - \pi_{i1})(1 - \pi_{i2})}} \right), \quad (11)$$

where ρ is the correlation between v_{i1} and v_{i2} , and $\boldsymbol{\alpha} = (\mu_{x_1}, \mu_{x_2}, \sigma_{x_1}, \sigma_{x_2}, \sigma_{x_1 x_2})'$ is the vector of nuisance parameters of the distribution of (x_{i1}, x_{i2}) , which are estimated simultaneously along with the regression coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$ for the binary regression model (10).

From (8), the likelihood score equations for $\boldsymbol{\beta}$ take the form

$$\sum_{i=1}^n \int_{-\infty}^c \{y_i - \mu_i(\boldsymbol{\beta}, \mathbf{x}_i)\} \mathbf{x}_i f_{x_{i1}|y_i, x_{o,i}, v_i}(\mathbf{x}_{\text{lod},i} | y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i; \boldsymbol{\gamma}) d\mathbf{x}_{\text{lod},i} = \mathbf{0}, \quad (12)$$

where $\boldsymbol{\gamma} = (\boldsymbol{\beta}, \boldsymbol{\alpha}, \rho)$. These equations can be solved numerically using the Newton-Raphson iterative equations

$$\begin{aligned} \boldsymbol{\beta}^{(k+1)} &= \boldsymbol{\beta}^{(k)} + \left\{ \sum_{i=1}^n \int_{-\infty}^c w_i(\hat{\boldsymbol{\beta}}^{(k)}) \mathbf{x}_i \mathbf{x}_i' f_{x_{i1}|y_i, x_{o,i}, v_i}(\mathbf{x}_{\text{lod},i} | y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i; \boldsymbol{\gamma}^{(k)}) d\mathbf{x}_{\text{lod},i} \right\}^{-1} \\ &\quad \times \sum_{i=1}^n \int_{-\infty}^c w_i(\hat{\boldsymbol{\beta}}^{(k)}) \mathbf{x}_i \{y_i - \mu_i(\boldsymbol{\beta}^{(k)}, \mathbf{x}_i)\} f_{x_{i1}|y_i, x_{o,i}, v_i}(\mathbf{x}_{\text{lod},i} | y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i; \boldsymbol{\gamma}^{(k)}) d\mathbf{x}_{\text{lod},i}, \end{aligned} \quad (13)$$

for $k = 0, 1, 2, \dots$, where the weights are given by $w_i(\boldsymbol{\beta}) = \mu_i(\boldsymbol{\beta}, \mathbf{x}_i)(1 - \mu_i(\boldsymbol{\beta}, \mathbf{x}_i))$ and the conditional density $f_{x_{i1}|y_i, x_{o,i}, v_i}(\mathbf{x}_{\text{lod},i} | y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i; \boldsymbol{\gamma})$ is obtained as

$$\begin{aligned} f_{x_{i1}|y_i, x_{o,i}, v_i}(\mathbf{x}_{\text{lod},i} | y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i; \boldsymbol{\gamma}) &= \frac{f_{y_i|x_i}(y_i | \mathbf{x}_{\text{obs},i}, \mathbf{x}_{\text{lod},i}; \boldsymbol{\beta}) f_{x_i}(\mathbf{x}_{\text{obs},i}, \mathbf{x}_{\text{lod},i} | \boldsymbol{\alpha}) f_{\mathbf{v}_i}(\mathbf{v}_i | \boldsymbol{\alpha}, \rho)}{\int_{-\infty}^c f_{y_i|x_i}(y_i | \mathbf{x}_{\text{obs},i}, \mathbf{x}_{\text{lod},i}; \boldsymbol{\beta}) f_{x_i}(\mathbf{x}_{\text{obs},i}, \mathbf{x}_{\text{lod},i} | \boldsymbol{\alpha}) f_{\mathbf{v}_i}(\mathbf{v}_i | \boldsymbol{\alpha}, \rho) d\mathbf{x}_{\text{lod},i}} \\ &= \frac{f_{y_i|x_i}(y_i | \mathbf{x}_{\text{obs},i}, \mathbf{x}_{\text{lod},i}; \boldsymbol{\beta}) f_{x_i}(\mathbf{x}_{\text{obs},i}, \mathbf{x}_{\text{lod},i} | \boldsymbol{\alpha})}{\int_{-\infty}^c f_{y_i|x_i}(y_i | \mathbf{x}_{\text{obs},i}, \mathbf{x}_{\text{lod},i}; \boldsymbol{\beta}) f_{x_i}(\mathbf{x}_{\text{obs},i}, \mathbf{x}_{\text{lod},i} | \boldsymbol{\alpha}) d\mathbf{x}_{\text{lod},i}}. \end{aligned}$$

The initial values of the regression parameters, ie, $\boldsymbol{\beta}^{(0)}$, may be chosen as the ordinary maximum likelihood estimates of the regression parameters obtained by substituting the left-censored values of the covariates x_1 and x_2 with corresponding $(1/2)\text{LOD}$ values $0.5c_1$ and $0.5c_2$, respectively, and by treating them as actual complete data.

To estimate the vector of nuisance parameters $\alpha^* = (\alpha, \rho)$, following Equation (8), we solve the likelihood score equations

$$\sum_{i=1}^n \int_{-\infty}^c \frac{\partial}{\partial \alpha^*} \{ \log f_{x_i}(\mathbf{x}_i | \alpha) + \log f_{v_i}(\mathbf{v}_i | \alpha, \rho) \} f_{x_i | y_i, x_{0i}, v_i}(\mathbf{x}_{\text{lod},i} | y_i, \mathbf{x}_{\text{obs},i}, \mathbf{v}_i; \gamma) d\mathbf{x}_{\text{lod},i} = \mathbf{0}, \quad (14)$$

with respect to α^* using an iterative equation similar to (13).

Note that, as we consider only two covariates in the binary model (10), it is not very tedious to find the exact maximum likelihood estimates by evaluating the integrals within the iterative equations using a numerical integration technique. Therefore, for our numerical analysis discussed in the next two sections, we consider finding the exact estimates of the model parameters. For high-dimensional integration involving multiple covariates, however, the iterative method would require intensive computation. Some resampling algorithms, such as the Metropolis-Hastings algorithm (see, eg, the work of McCulloch²⁰), may be used to approximate the high-dimensional integrals and hence to obtain approximate values of the corresponding maximum likelihood estimators.

4 | SIMULATION STUDY

To investigate the performance of the proposed method, we conducted a series of Monte Carlo simulations based on the following two binary regression models.

- i. $y_i | x_{i1} \sim \text{ind. Bernoulli}(\mu_i)$; $\log\{\mu_i/(1 - \mu_i)\} = \beta_0 + \beta_1 x_{i1}$;
- ii. $y_i | x_{i1}, x_{i2} \sim \text{ind. Bernoulli}(\mu_i)$; $\log\{\mu_i/(1 - \mu_i)\} = \beta_1 + \beta_2 x_{i1} + \beta_2 x_{i2}$,

for $i = 1, \dots, n$, where the regression parameters were fixed at $\beta_0 = -2$, $\beta_1 = 0.5$, and $\beta_2 = 1$. The values of the covariates (x_{i1}, x_{i2}) were generated from a bivariate normal distribution with means $E(x_{i1}) = \mu_{x_1} = 0$ and $E(x_{i2}) = \mu_{x_2} = 2$, and covariance terms $\text{var}(x_{i1}) = \sigma_{x_1}^2 = 1$, $\text{var}(x_{i2}) = \sigma_{x_2}^2 = 2$, and $\text{cov}(x_{i1}, x_{i2}) = \sigma_{x_1 x_2} = 0.25$. The response y_i was considered fully observed, whereas the covariates were assumed left censored due to the LOD. We chose the LOD values so that the overall proportion of left-censored covariates was either 0.3 or 0.5 for each of the aforementioned two models. For example, when the covariate $x_1 \sim N(0, 1)$, we can choose the LOD values $\Phi^{-1}(0.3) = -0.5244$ and $\Phi^{-1}(0.5) = 0$ so as to get 30% and 50% left-censored values of the covariate, respectively. We ran the simulations for each combination of the sample sizes $n = 100, 200, 500$. Each simulation run was based on 1000 replicates of datasets. The statistical software R version 3.1.1 was used for the numerical study. An R program to compute parameter estimates for generalized linear models with nondetects is available from the authors upon request.

We compared the following three methods.

1. *Naive method* (N): The ordinary maximum likelihood estimates of the model parameters are found by replacing the left-censored values of the covariates with the (1/2)LOD value and by treating them as actual complete data.
2. *Nonweighted method* (NW): The maximum likelihood estimates are found by ignoring the “weight” $f_{v_i}(\mathbf{v}_i | \tau)$ in Equation (6), that is, by maximizing the likelihood

$$L_0 = \prod_{i=1}^n \int_{-\infty}^c f_{y_i | x_i}(y_i | \mathbf{x}_{\text{obs},i}, \mathbf{x}_{\text{lod},i}, \beta) f_{x_i}(\mathbf{x}_{\text{obs},i}, \mathbf{x}_{\text{lod},i} | \alpha) d\mathbf{x}_{\text{lod},i}.$$

3. *Weighted method* (W): The proposed estimates are found by maximizing the full likelihood function (7).

Figure 1 exhibits the empirical biases of the estimators of the regression parameters β_0 and β_1 , and nuisance parameters μ_{x_1} and σ_{x_1} for various proportions (0.1 – 0.6) of the left-censored (LOD) covariate x_1 , where we assume a binary response y_i with success probability μ_i , which is related to the covariate x_{i1} by the logit link function $\log(\mu_i/(1 - \mu_i)) = \beta_0 + \beta_1 x_{i1}$. We assume that x_{i1} follows an independent $N(\mu_{x_1}, \sigma_{x_1}^2)$ distribution. The values of the model parameters were fixed at $\beta_0 = -2$, $\beta_1 = 0.5$, $\mu_{x_1} = 0$, and $\sigma_{x_1} = 1$. It is clear from the figure that the naive method (N) produces systematic biases of the estimators for all model parameters and sample sizes considered. The other two methods (NW and W) produce slight biases of the regression estimators for a smaller sample size ($n = 100$). However, these biases tend to decrease when the sample size increases. As expected, both NW and W methods appear to be roughly unbiased for large samples.

Table 1 supplements the results in Figure 1 by presenting the empirical biases and MSEs of the estimators of the regression coefficients β_0 and β_1 for the binary regression model with a left-censored (LOD) covariate x_1 by considering two proportions of LOD, ie, 0.3 and 0.5. Table A1 in the Appendix presents the empirical results for estimators of the nuisance

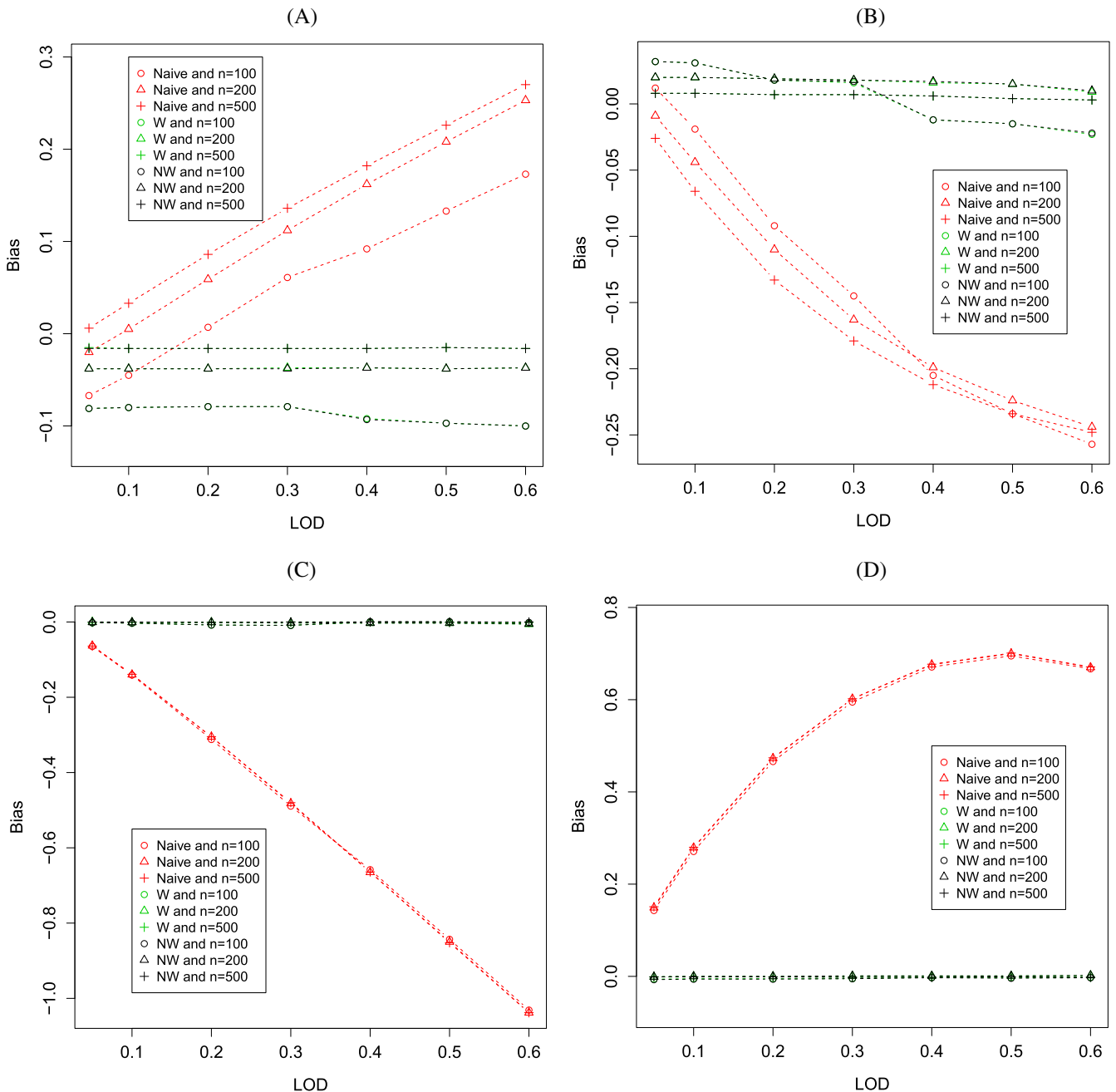


FIGURE 1 Empirical biases of estimators of regression parameters β_0 and β_1 , and nuisance parameters μ_{x_1} and σ_{x_1} for various proportions of left-censored (LOD) covariate x_1 . Binary response y_i is used with success probability μ_i , where $\log(\mu_i/(1 - \mu_i)) = \beta_0 + \beta_1 x_{i1}$ and $x_{i1} \sim \text{ind. Normal}(\mu_{x_1}, \sigma_{x_1}^2)$. Parameters were fixed at $\beta_0 = -2$, $\beta_1 = 0.5$, $\mu_{x_1} = 0$, and $\sigma_{x_1} = 1$. A, $E(\hat{\beta}_0) - \beta_0$; B, $E(\hat{\beta}_1) - \beta_1$; C, $E(\hat{\mu}_{x_1}) - \mu_{x_1}$; D, $E(\hat{\sigma}_{x_1}) - \sigma_{x_1}$. LOD, limit of detection; NW, nonweighted; W, weighted [Colour figure can be viewed at wileyonlinelibrary.com]

parameters μ_{x_1} and σ_{x_1} of the covariate distribution. It is clear from the tables that the naive method (N) provides systematic biases and large MSEs for all estimators. Furthermore, both bias and MSE tend to increase when the proportion of the LOD increases. For example, when estimating β_1 at the sample size $n = 200$, it appears from Table 1 that the naive method gives biases of -0.1805 (36% relative bias) and -0.2367 (47% relative bias) for the LOD proportions 0.3 and 0.5, respectively. The corresponding biases from the nonweighted method (NW) are -0.0027 and -0.0042 , and that from the weighted method (W) are -0.0028 and -0.0044 , which indicate that the NW and W methods are roughly unbiased. Both NW and W methods also appear to be equally efficient for large samples in terms of the MSEs. For small samples, however, the W method appears to give smaller biases and MSEs of the regression estimators, as compared to the NW method. For example, when estimating β_1 at the sample size $n = 100$ and with the LOD proportion 0.5, the NW method gives a

TABLE 1 Empirical biases and mean squared errors (MSEs) of estimators of regression parameters β_0 and β_1 for two proportions (0.3, 0.5) of left-censored (limit of detection (LOD)) covariate x_{i1} . Response $y_i \sim \text{ind. Bernoulli}(\mu_i)$ with the logit link $\log(\mu_i/(1 - \mu_i)) = \beta_0 + \beta_1 x_{i1}$ and covariate $x_{i1} \sim \text{ind. Normal}(\mu_{x_1}, \sigma_{x_1}^2)$. Parameters are fixed at $\beta_0 = -2$, $\beta_1 = 0.5$, $\mu_{x_1} = 0$, and $\sigma_{x_1}^2 = 1$. (N = Naive method; NW = Nonweighted method; W = Weighted method)

<i>n</i>	LOD	Method	Bias		MSE	
			$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$
100	0.3	N	0.0482	-0.1500	0.1343	0.1050
		NW	-0.0878	0.0049	0.1436	0.1366
		W	-0.0876	0.0047	0.1436	0.1365
	0.5	N	0.1344	-0.2347	0.2936	0.1308
		NW	-0.1012	-0.0131	0.3031	0.2125
		W	-0.0987	-0.0122	0.2513	0.1953
200	0.3	N	0.1248	-0.1805	0.0707	0.0626
		NW	-0.0201	-0.0027	0.0606	0.0560
		W	-0.0200	-0.0028	0.0605	0.0559
	0.5	N	0.2145	-0.2367	0.1015	0.0760
		NW	-0.0215	-0.0042	0.0619	0.0629
		W	-0.0213	-0.0044	0.0619	0.0629
500	0.3	N	0.1298	-0.1758	0.0371	0.0420
		NW	-0.0239	0.0134	0.0237	0.0213
		W	-0.0239	0.0134	0.0237	0.0213
	0.5	N	0.2213	-0.2303	0.0694	0.0604
		NW	-0.0234	0.0109	0.0242	0.0233
		W	-0.0233	0.0108	0.0242	0.0233

bias of -0.0131 and an MSE of 0.2125 , whereas the W method provides a slightly smaller bias of -0.0122 and a smaller MSE of 0.1953 .

Table 2 presents the empirical biases and MSEs of the estimators of the regression coefficients β_0 , β_1 , and β_2 for the binary regression model with two left-censored covariates x_1 and x_2 . Table A2 in the Appendix presents the empirical results for the nuisance parameters μ_{x_1} , μ_{x_2} , σ_{x_1} , σ_{x_2} , and $\rho_{x_1, x_2} = \text{corr}(x_1, x_2)$ of the bivariate normal distribution for the left-censored covariates. We assume that the binary response y_i has the success probability μ_i , which is related to the covariates x_{i1} and x_{i2} by the logit link function $\log(\mu_i/(1 - \mu_i)) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$. The values of the parameters were fixed at $\beta_0 = -2$, $\beta_1 = 0.5$, $\beta_2 = 1$, $\mu_{x_1} = 0$, $\mu_{x_2} = 2$, $\sigma_{x_1}^2 = 1$, $\sigma_{x_2}^2 = 2$, and $\sigma_{x_1, x_2} = 0.5$. Here, also, we observe that, unlike the NW and W methods, the naive method (N) provides systematically large biases and MSEs for all regression coefficients and nuisance parameters considered. For example, when $n = 200$ and $\text{LOD} = 0.3$, for estimating the regression parameter β_1 , as shown in Table 2, the naive method (N) provides a bias of -0.2169 (43% relative bias) and an MSE of 0.0613 , whereas the nonweighted (NW) provides a bias of 0.0172 and an MSE of 0.0497 and the weighted (W) method provides a bias of 0.0170 and an MSE of 0.0497 . The NW and W methods appear to be almost equally efficient in terms of biases and MSEs for all the model parameters.

It is interesting to note that, in some cases the naive method provides lower MSEs of the regression parameters, as compared to the other two methods. For example, in Table 1, for $n = 100$ and $\text{LOD} = 0.3$, when estimating the slope parameter β_1 , the naive method (N) provides an MSE of 0.1050 , which is smaller than MSEs 0.1366 and 0.1365 obtained by the nonweighted (NW) and weighted (W) methods, respectively. This artificially smaller MSE obtained by the naive method is due to the fact that the method generally provides systematic biases that lead to “shrinkage estimators” of the model parameters. Here, it is clear that the naive method (N) provides a large bias of -0.1500 (30 relative bias)%, as compared to very small biases of 0.0049 and 0.0047 as obtained by the nonweighted (NW) and weighted (W) methods, respectively. When the sample size n and percentage of LOD both increase, the naive method (N) appears to provide worst results in terms of much larger systematic biases and larger MSEs as compared to the other two methods. For example, in Table 1, for $n = 500$ and $\text{LOD} = 0.5$, when estimating the slope parameter β_1 , the naive method (N) provides a bias of -0.2303 (46% relative bias) and an MSE of 0.0604 , whereas the nonweighted (NW) provides a bias of 0.0109 and an MSE of 0.0233 and the weighted (W) method provides a bias of 0.0108 and an MSE of 0.0233 .

TABLE 2 Empirical biases and mean squared errors (MSEs) of estimators of regression parameters β_0 , β_1 , and β_2 for two proportions (0.3, 0.5) of left-censored (limit of detection (LOD)) covariates x_1 and x_2 . Response $y_i \sim \text{ind. Bernoulli}(\mu_i)$ with the logit link $\log(\mu_i/(1 - \mu_i)) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$ and (x_{i1}, x_{i2}) are bivariate normal with means (μ_{x_1}, μ_{x_2}) and covariance terms $\sigma_{x_1}^2$, $\sigma_{x_2}^2$, and $\sigma_{x_1 x_2}$. Parameters are fixed at $\beta_0 = -2$, $\beta_1 = 0.5$, $\beta_2 = 1$, $\mu_{x_1} = 0$, $\mu_{x_2} = 2$, $\sigma_{x_1}^2 = 1$, $\sigma_{x_2}^2 = 2$, and $\sigma_{x_1 x_2} = 0.5$. (N = Naive method; NW = Nonweighted method; W = Weighted method)

n	LOD	Method	Bias			MSE		
			$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
100	0.3	N	0.4022	-0.2084	-0.1077	0.6663	0.0733	0.0971
		NW	-0.1525	0.0366	0.0710	0.3784	0.1030	0.0681
		W	-0.1517	0.0362	0.0708	0.3783	0.1030	0.0681
	0.5	N	2.1311	-0.2130	-0.5643	4.6472	0.0715	0.3268
		NW	-0.2261	0.0532	0.1020	0.5507	0.1387	0.1006
		W	-0.2232	0.0529	0.1011	0.5485	0.1386	0.1001
200	0.3	N	0.4850	-0.2169	-0.1478	0.4529	0.0613	0.0572
		NW	-0.0749	0.0172	0.0328	0.1450	0.0497	0.0260
		W	-0.0746	0.0170	0.0327	0.1450	0.0497	0.0260
	0.5	N	2.1427	-0.2276	-0.5810	4.6362	0.0644	0.3407
		NW	-0.0920	0.0117	0.0399	0.2152	0.0625	0.0376
		W	-0.0908	0.0114	0.0396	0.2150	0.0624	0.0376
500	0.3	N	0.5578	-0.2262	-0.1789	0.3923	0.0564	0.0451
		NW	-0.0234	0.0016	0.0124	0.0514	0.0178	0.0096
		W	-0.0232	0.0016	0.0123	0.0514	0.0178	0.0096
	0.5	N	2.1497	-0.2300	-0.5854	4.6388	0.0576	0.3439
		NW	-0.0356	0.0040	0.0173	0.0774	0.0221	0.0135
		W	-0.0352	0.0039	0.0172	0.0774	0.0221	0.0135

4.1 | Results for other response and covariate distributions

So far, we have studied estimators in logistic regression models for binary discrete outcomes with left-censored covariates that were assumed to be normally distributed. In this section, we extend our simulation study by considering other distributions in the exponential family for modeling the outcomes and covariates, which include the Gaussian distribution for continuous outcomes and gamma distribution for left-censored covariates. Specifically, we extend our simulations using two additional models.

- $y_i | x_{i1} \sim \text{ind. Bernoulli}(\mu_i)$; $\log\{\mu_i/(1 - \mu_i)\} = \beta_0 + \beta_1 x_{i1}$; $x_{i1} \sim \text{ind. Gamma}(\gamma_{x_1}, \lambda_{x_1})$;
- $y_i | x_{i1} \sim \text{ind. Normal}(\mu_i, \sigma^2)$; $\mu_i = \beta_0 + \beta_1 x_{i1}$; $x_{i1} \sim \text{ind. Normal}(\mu_{x_1}, \sigma_{x_1}^2)$,

for $i = 1, \dots, n$. For the logistic model iii, the regression parameters were fixed at $\beta_0 = -1$ and $\beta_1 = 1$, and the shape and scale (rate) parameters of the gamma distribution for the covariate x_{i1} were fixed at $\gamma_{x_1} = 4$ and $\lambda_{x_1} = 2$, respectively. For the Gaussian model iv, the linear regression parameters were fixed at $\beta_0 = 1$ and $\beta_1 = 2$, the variance parameter at $\sigma^2 = 1$, and the mean and variance of the normal distribution for the covariate x_{i1} at $\mu_{x_1} = 2$ and $\sigma_{x_1}^2 = 1$, respectively. As before, the response y_i was considered to be fully observed, and the covariates were considered left censored due to the LOD. We chose the LOD values of the covariate in such a way that the overall proportion of the left-censored covariates was either 0.3 or 0.5 for each of the aforementioned two models. The simulations were carried out for each combination of the sample sizes $n = 100, 200, 500$ for the logistic model iii). For the linear model iv, the simulations were based on three combinations of sample sizes $n = 40, 60, 100$. Each simulation run was based on a series of 1000 replicates of datasets.

Table 3 presents empirical biases and MSEs for the estimators of the regression parameters β_0 and β_1 for the binary regression model iii and for a single left-censored gamma covariate x_1 with two proportions of LOD, ie, 0.3 and 0.5. Table A3 in the Appendix presents the corresponding empirical results for the nuisance parameters γ_{x_1} and λ_{x_1} of the gamma distribution for the left-censored covariate x_1 . As before, it is clear from the tables that the naive method (N) generally provides systematic biases and large MSEs for all estimates of the model parameters. Moreover, both bias and MSE tend to increase when the proportion of LOD increases. For example, when estimating β_1 at sample size $n = 200$, it appears from Table 3 that the naive method (N) gives biases of -0.1203 (12% relative bias) and -0.1435 (14% relative bias) for the LOD proportions 0.3 and 0.5, respectively. The corresponding biases from the nonweighted method (NW) are 0.0328 and 0.0353, and that from the weighted method (W) are 0.0326 and 0.0351, respectively, which indicate that

TABLE 3 Empirical biases and mean squared errors (MSEs) of estimators of regression parameters β_0 and β_1 for two proportions (0.3, 0.5) of left-censored (limit of detection (LOD)) covariate x_1 . Response $y_i \sim \text{ind. Bernoulli}(\mu_i)$ with the logit link $\log(\mu_i/(1 - \mu_i)) = \beta_0 + \beta_1 x_{1i}$ and covariate $x_{1i} \sim \text{ind. Gamma}(\gamma_{x_1}, \lambda_{x_1})$. Parameters are fixed at $\beta_0 = -1$, $\beta_1 = 1$, $\gamma_{x_1} = 4$, and $\lambda_{x_1} = 2$. (N = Naive method; NW = Nonweighted method; W = Weighted method)

<i>n</i>	LOD	Method	Bias		MSE	
			$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$
100	0.3	N	0.3178	-0.1083	0.3317	0.0869
		NW	-0.0595	0.0554	0.3771	0.1233
		W	-0.0583	0.0549	0.3769	0.1233
	0.5	N	0.4093	-0.1330	0.4072	0.1003
		NW	-0.0738	0.0562	0.4817	0.1482
		W	-0.0724	0.0557	0.4811	0.1479
200	0.3	N	0.3158	-0.1203	0.2092	0.0479
		NW	-0.0428	0.0328	0.1732	0.0530
		W	-0.0422	0.0326	0.1730	0.0529
	0.5	N	0.4117	-0.1435	0.2805	0.0627
		NW	-0.0507	0.0353	0.2223	0.0724
		W	-0.0501	0.0351	0.2225	0.0725
500	0.3	N	0.3220	-0.1304	0.1523	0.0312
		NW	-0.0252	0.0161	0.0766	0.0223
		W	-0.0250	0.0160	0.0766	0.0223
	0.5	N	0.4215	-0.1543	0.2224	0.0394
		NW	-0.0247	0.0161	0.0863	0.0262
		W	-0.0245	0.0160	0.0863	0.0262

the NW and W methods are roughly unbiased. Both NW and W methods also appear to be equally efficient for large samples in terms of the MSEs. For small samples, however, the W method appears to give smaller biases and MSEs of the regression estimators, as compared to the NW method. For example, when estimating β_1 at sample size $n = 100$ and with the LOD proportion 0.5, the NW method gives a bias of 0.0562 and an MSE of 0.1482, whereas the W method provides a slightly smaller bias of 0.0557 and a smaller MSE of 0.1479.

Table 4 presents empirical biases and MSEs for the estimators of the regression parameters β_0 and β_1 for the linear regression model iv and for a single left-censored normal covariate x_1 with two proportions of LOD, ie, 0.3 and 0.5. Table A4 in the Appendix presents the corresponding empirical results for the nuisance parameters μ_{x_1} and σ_{x_1} of the normal distribution for the left-censored covariate x_1 . Here, also, it is clear from the tables that the naive method (N) provides systematic biases and large MSEs for all parameter estimates. For example, when estimating β_0 at sample size $n = 100$, it appears from Table 4 that the naive method (N) gives biases of 0.2230 (22% relative bias) and 0.5446 (54% relative bias) for the LOD proportions 0.3 and 0.5, respectively. The corresponding biases from the nonweighted method (NW) are -0.0246 and -0.0285, and that from the weighted method (W) are -0.0212 and -0.0282, respectively, which indicate that the NW and W methods are roughly unbiased. Both NW and W methods are equally efficient for larger samples. For smaller samples, the W method generally provides smaller biases in the regression estimators, as compared to the NW method. For example, when estimating β_1 at sample size $n = 40$ and with the LOD proportion 0.3, the NW method gives a bias of 0.0155, whereas the W method provides a slightly smaller bias of 0.0138.

5 | APPLICATION: MIREC COHORT STUDY

The MIREC study¹⁷ recruited around 2000 pregnant women between 2008 and 2011 from 10 sites across Canada. Participant inclusion criteria were the ability to consent and to communicate in English or French, age 18 years or older, less than 14 weeks gestation, willing to provide a sample of cord blood, and planning on delivering at a local hospital. Women with a certain medical history were excluded from the study. Biospecimens were collected during each trimester, at delivery, and in the early postnatal period. Questionnaires were administered during the first and third trimesters to collect demographic, lifestyle, medical history, use of natural health products and medications, and potential sources of

TABLE 4 Empirical biases and mean squared errors (MSEs) of estimators of β_0 , β_1 , and σ for two proportions (0.3, 0.5) of left-censored (limit of detection (LOD)) covariate x_1 . Response $y_i \sim \text{ind. Normal}(\mu_i, \sigma^2)$ with $\mu_i = \beta_0 + \beta_1 x_{1i}$ and covariate $x_{1i} \sim \text{ind. Normal}(\mu_{x_1}, \sigma_{x_1}^2)$. Parameters are fixed at $\beta_0 = 1$, $\beta_1 = 2$, $\sigma = 1$, $\mu_{x_1} = 2$, and $\sigma_{x_1} = 1$. (N = Naive method; NW = Nonweighted method; W = Weighted method)

<i>n</i>	LOD	Method	Bias			MSE		
			$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\sigma}$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\sigma}$
40	0.3	N	0.2129	-0.0765	0.1069	0.2700	0.0446	0.0319
		NW	-0.0370	0.0155	-0.0402	0.2815	0.0465	0.0173
		W	-0.0323	0.0138	-0.0393	0.2831	0.0467	0.0173
	0.5	N	0.5575	-0.1846	0.2747	0.5645	0.0771	0.1061
		NW	-0.0611	0.0288	-0.0426	0.4550	0.0674	0.0228
		W	-0.0587	0.0281	-0.0410	0.4575	0.0678	0.0228
60	0.3	N	0.2388	-0.0884	0.1245	0.1983	0.0334	0.0292
		NW	-0.0152	0.0049	-0.0255	0.1693	0.0298	0.0110
		W	-0.0122	0.0038	-0.0249	0.1690	0.0297	0.0110
	0.5	N	0.5439	-0.1812	0.2876	0.4648	0.0607	0.1042
		NW	-0.0348	0.0161	-0.0346	0.3543	0.0517	0.0155
		W	-0.0315	0.0150	-0.0336	0.3547	0.0520	0.0154
100	0.3	N	0.2230	-0.0806	0.1368	0.1369	0.0216	0.0274
		NW	-0.0246	0.0089	-0.0121	0.0978	0.0162	0.0068
		W	-0.0212	0.0076	-0.0118	0.0984	0.0163	0.0068
	0.5	N	0.5446	-0.1826	0.2946	0.4029	0.0504	0.0990
		NW	-0.0285	0.0111	-0.0258	0.2008	0.0290	0.0087
		W	-0.0282	0.0110	-0.0252	0.2028	0.0293	0.0087

exposure data. A validated food frequency questionnaire was administered in the second trimester, along with blood and spot urine collection, blood pressure, clinical laboratory tests, and anthropometric measurements. In this study, we examined the association between the participants' birth outcomes (BOs) (eg, low birth weight and spontaneous abortion) and levels of chemical mixtures. We consider a dichotomous response variable y , representing the participants' health outcome, and covariates (x_1, \dots, x_p) , measuring the levels of chemical concentrations. Considering a variable with high proportion of LOD observations provides little information on the model fitting, we therefore excluded those with more than 70% nondetects.

There were 38 available chemicals with less than 70% nondetects in the MIREC database and we used a backward elimination procedure based on likelihood deviances to choose the covariates (including the chemical mixtures) in our regression models that were found to be associated with each of the seven binary health outcomes, ie, (i) BO (0 = delivery of a live birth; 1 = spontaneous abortion); (ii) glucose tolerance outcome (OGTT_1) (0 = normal; 1 = gestational diabetes mellitus (GDM)); (iii) glucose tolerance outcome (OGTT_2) (0 = normal; 1 = impaired glucose tolerance (IGT)); (iv) whether an infant's weight was considered as low birth weight (birth weight \leq 2500 grams) (LBW) (0 = normal; 1 = low birth weight); (v) whether an infant is considered large for gestational age (LGA) (0 = normal; 1 = large); (vi) whether an infant is considered small for gestational age (SGA) (0 = normal; 1 = small); and (vii) whether it is a preterm birth for singleton live births PreB (0 = gestational age \geq 37 weeks, 1 = gestational age \leq 37 weeks). Note that GDM and IGT were categorized in accordance with guidelines from the Canadian Diabetes Association and the Society of Obstetricians and Gynaecologists of Canada, as described in the work of Shapiro et al.²¹ The LGA and SGA categories were \leq 10th and \geq 90th percentiles, respectively, as described in the work of Thomas et al.²² The description, sources, and units of the chemical concentrations and their summary statistics are presented in Table 5. It is clear from the table that some of the chemical mixtures included measurements with a high proportion of detection limits. For example, the chemical dimethylphosphate (DMP) was measured with 21% LOD values, cotinine with 46% LOD values, and dimethylarsinic acid (DMAA) with 14% LOD values. The chemical concentrations were found to be positively skewed with a large variability in the measurements. To reduce the variability, we took the natural logarithm of the chemical concentrations and used them as covariates in the regression model. The histogram plots (not shown here) of the log-transformed values were found to be roughly symmetric and bell shaped, and we assumed that they followed a multivariate normal distribution with unknown mean and variance parameters, which are to be estimated along with the regression parameters by the proposed joint likelihood method as described earlier.

TABLE 5 Definitions, sources, and units of chemical concentrations in MIREC study with summary statistics

Abbreviation	Description	Matrix	Units	% \leq LOD	MEAN	STD
<i>Plasticisers</i>						
MBzP	Mono benzyl phthalate	Urine	$\mu\text{g/L}$	0.50%	12.19	25.39
MEOHP	Mono-(2-ethyl-5-oxohexyl) phthalate	Urine	$\mu\text{g/L}$	0.28%	15.16	47.96
MEHHP	Mono-(2-ethyl-5-hydroxyhexyl) phthalate	Urine	$\mu\text{g/L}$	0.62%	23.52	74.10
<i>Perfluoroalkyl substances (PFASs)</i>						
PFOA	Perfluorooctanoic acid	Plasma	$\mu\text{g/L}$	0.15%	1.95	1.24
PFHxS	Perfluorohexane sulfonate	Plasma	$\mu\text{g/L}$	4.12%	1.46	1.88
<i>PCBs</i>						
BPC170	2,2',3,3',4,4',5-heptachlorobiphenyl	Plasma	$\mu\text{g/L}$	46.82%	0.02	0.02
<i>Organophosphate Pesticides (OPs)</i>						
DMP	Dimethylphosphate	Urine	$\mu\text{g/L}$	20.83%	5.26	8.72
<i>Organochlorine Pesticides (OCs)</i>						
OXYCHLOR	Oxychlordan	Plasma	$\mu\text{g/L}$	7.81%	0.01	0.01
TRANSONA	Trans-nonachlor	Plasma	$\mu\text{g/L}$	15.87%	0.02	0.02
<i>Arsenic species</i>						
ASAL	Arsenobetaine	Urine	$\mu\text{mol/L}$	51.16%	0.12	0.69
DMAA	Dimethylarsinic acid	Urine	$\mu\text{mol/L}$	14.12%	0.05	0.07
<i>Smoking Biomarker</i>						
COTISE	Cotinine	Plasma	ng/mL	46.08%	5.98	27.52

Note that the substitution was applied on \leq LOD observations. LOD, limit of detection.

We fitted a binary regression model for each of the seven binary outcomes, ie, BO, OGTT_1, OGTT_2, LBW, LGA, SGA, and PreB, which were described as functions of the log-transformed values of the chemical mixtures used in the MIREC study. As the values of the chemical concentrations were left-censored due to the limits of detection, we obtained the estimates of the model parameters based on the proposed maximum likelihood method by addressing the issue of left-censoring in the chemical mixtures. To model the binary indicators of nondetects, we used Bahdaur models with simple “exchangeable” correlation structures. Note that, when describing the associations among the health outcomes and chemical mixtures, we adjusted the logistic regression models for other demographic variables including being a first-time mother, mother’s smoking status, and prepregnancy BMI. Table 6 presents estimates of the regression parameters and their corresponding standard errors obtained by each of the three estimation methods, ie, N, NW, and W, as considered earlier. Moreover, Table A5 in Appendix presents estimates of the nuisance parameters with their corresponding standard errors. The three methods appear to provide somewhat similar conclusions about the regression coefficients. The naive method (N), however, produces slightly different estimates than those obtained by the other two methods. For example, when the response is the BO, the naive method produces the estimated values 0.198 and -0.715 for the effects of the chemical mixtures COTISE and PFHxS, respectively, whereas the weighted method produces somewhat different estimated values of 0.174 and -0.770 , respectively. Here, the discrepancies among the estimated regression coefficients are due to the fact the chemical COTISE contains a large proportion of left-censored values (46% LOD) in its measurements. The naive method generally produces systematic biases in the regression estimators in such a case, as we have observed in the simulation study earlier.

It is clear from Table 6 that some of the health outcomes were associated with the chemical concentrations considered in the models. For example, higher concentrations of oxychlordan (OXYCHLOR) was associated with an increased risk of a low birth weight infant. Furthermore, higher concentrations of arsenobetaine (ASAL) was associated with an increased risk of a small for SGA. In particular, given a fixed value of cotinine, for every one unit increase in the logarithmic value of oxychlordan, the odds of having a low birth weight infant increases by 1.61 ($= \exp(0.477)$) times, as obtained by the proposed method (W). Similarly, given a fixed value of perfluorooctanoic acid (PFOA), for every one unit increase in the logarithmic value of arsenobetaine, the odds of having a small for gestational age baby increases by 1.08 ($= \exp(0.173)$) times. Among the two chemicals trans-nonachlor (TRANSONA) and 2,2',3,3',4,4',5-heptachlorobiphenyl (BPC170) considered in the logistic regression model for preterm birth (infants born before 37 weeks), TRANSONA appears to be significantly positively associated with the preterm birth (PreB), but the chemical BPC170 negatively associated with the preterm birth.

TABLE 6 Estimates of regression parameters (standard errors in parentheses) from logistic regression fits to MIREC data. (N = Naive method; NW = Nonweighted method; W = Weighted method)

y	x_1	x_2	Method	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$
BO	COTISE	PFHxS	N	-3.918 _(0.301)	0.199 _(0.058)	-0.831 _(0.284)
			NW	-3.917 _(0.301)	0.174 _(0.061)	-0.874 _(0.287)
			W	-3.937 _(0.294)	0.177 _(0.063)	-0.876 _(0.288)
OGTT_1	DMAA	DMP	N	-0.960 _(0.664)	0.582 _(0.167)	-0.574 _(0.201)
			NW	-1.185 _(0.617)	0.533 _(0.164)	-0.565 _(0.165)
			W	-1.179 _(0.616)	0.535 _(0.164)	-0.567 _(0.164)
OGTT_2	MEOHP	MEHHP	N	-3.905 _(0.405)	-1.931 _(0.781)	1.819 _(0.755)
			NW	-3.909 _(0.403)	-2.017 _(0.790)	1.891 _(0.761)
			W	-3.908 _(0.403)	-2.012 _(0.790)	1.887 _(0.761)
LBW	OXYCHLOR	COTISE	N	-0.143 _(0.750)	0.523 _(0.172)	0.069 _(0.033)
			NW	-0.413 _(0.757)	0.474 _(0.169)	0.040 _(0.039)
			W	-0.411 _(0.754)	0.477 _(0.169)	0.038 _(0.039)
LGA	MBzP	DMAA	N	-3.261 _(0.43)	0.183 _(0.065)	-0.252 _(0.101)
			NW	-3.355 _(0.413)	0.179 _(0.067)	-0.269 _(0.091)
			W	-3.352 _(0.413)	0.179 _(0.067)	-0.269 _(0.091)
SGA	PFOA	ASAL	N	-1.933 _(0.316)	0.303 _(0.208)	0.130 _(0.060)
			NW	-1.794 _(0.331)	0.296 _(0.196)	0.174 _(0.055)
			W	-1.794 _(0.331)	0.295 _(0.196)	0.173 _(0.055)
PreB	TRANSONA	BPC170	N	-1.186 _(0.784)	0.448 _(0.175)	-0.099 _(0.185)
			NW	-1.874 _(0.697)	0.445 _(0.175)	-0.265 _(0.139)
			W	-1.879 _(0.700)	0.444 _(0.175)	-0.265 _(0.140)

Note: Bold faced numbers indicate significant coefficients from t-tests of nonzero coefficients at 5% level of significance. The seven binary health outcomes are defined as follows: (i) BO: birth outcome (0 = delivery of a live birth; 1 = spontaneous abortion), (ii) OGTT_1: glucose tolerance outcome (0 = normal; 1 = GDM), (iii) OGTT_2: glucose tolerance outcome (0 = normal; 1 = IGT), (iv) LBW: whether an infant's weight was considered as low birth weight (birth weight \leq 2500 grams) (0 = normal; 1 = low birth weight), (v) LGA: whether an infant is considered large for gestational age (0 = normal; 1 = large), (vi) SGA: whether an infant is considered small for gestational age (0 = normal; 1 = small), and (vii) PreB: whether it is a preterm birth for singleton live births (0 = gestational age \geq 37 weeks; 1 = gestational age \leq 37 weeks). Based on Akaike information criterion values for model selection, the adjusted covariate prepregnancy BMI is chosen for all outcome except for outcome SGA, where both the first-time mother and prepregnancy BMI are chosen by the model selection criteria.

6 | DISCUSSION

We often encounter problems of nondetects in clinical and environmental studies, where it is necessary to address the issue of nondetects for a valid statistical inference. We have developed and studied a novel method for analyzing data in the framework of generalized linear models with covariates subject to detection limits. The finite-sample properties of the proposed estimators are investigated using Monte Carlo simulations, where we have shown that our proposed method generally provides efficient estimators in terms of smaller biases and smaller MSEs as compared to its other competitors. Among the methods studied, the substitution method is the simplest and it is also easy to implement. However, this naive method provides estimators that are generally biased and inefficient, and hence is not recommended.

The proposed W method requires an assumption on the joint density function $f_{\mathbf{x}}(\mathbf{x}|\boldsymbol{\alpha})$ of the vector of covariates \mathbf{x} , where, in practice, the “true density” is unknown. We have investigated the robustness of the proposed method to violations of the distributional assumptions. In fact, when the shapes of the individual covariate distributions are roughly symmetric and bell shaped, their joint distribution can be safely approximated by a multivariate normal distribution. A slight deviation from normality (eg, when the true distribution is t) may not have much impact on the regression estimators. In the case of asymmetric distributions of the covariates, we considered using independent gamma distributions, which were found to be useful for robustly modeling the asymmetry in the distribution. For a completely unknown covariate distribution, one can also consider approximating the density function by a nonparameteric approach. This, however, is beyond the scope of the current study.

We have presented an application of the proposed method by analyzing actual data from the MIREC cohort study, where we investigated the effects of exposure to chemicals on the health outcomes of pregnant women and their newborns in

Canada. We have found some associations between chemicals and health conditions of pregnant women and their infants. Specifically, higher concentrations of arsenobetaine, cotinine, dimethylarsinic acid, mono-(2-ethyl-5-hydroxyhexyl) phthalate (MEHHP), oxychlorodane, mono benzyl phthalate (MBzP), perfluorooctanoic acid (PFOA), and trans-nonachlor appeared to be statistically associated with higher risks of adverse health outcomes. On the other hand, higher concentrations of perfluorohexane sulfonate (PFHxS), DMP, mono-(2-ethyl-5-oxohexyl) phthalate (MEOHP), and DMAA were statistically associated with lower risks of spontaneous abortion, GDM, IGT, and delivering a large infant, respectively.

Our Monte Carlo study indicates that the proposed weighted (W) method is generally more efficient than the non-weighted (NW) method when the sample size is small. The two methods appear to be equally efficient for larger samples as considered in the simulations. Further study of the models (not shown here) showed that, when a model contains multiple covariates with different proportions of nondetects, the W method generally performs better than the NW method irrespective of the sample size. This behavior is also observed in the MIREC data analysis, where measurements were obtained from a large group of $n = 1983$ individuals. As can be seen from Table 6, when fitting the response OGTT_2 as a function of the chemicals MEOHP and MEHHP with different proportions of nondetects (28% LOD for MEOHP and 62% LOD for MEHHP), the proposed W method provides estimates of the regression coefficients with smaller standard errors as compared to the NW method.

In observational epidemiological studies, in addition to a group of main factors of interest, it is common to adjust for potential confounding factors. In order for a variable to be considered as a confounder²³ (i) the variable must be independently associated with the outcome (ie, be a risk factor); (ii) the variable must be associated with the exposure under study in the source population, that is, it must be unequally distributed between exposure groups; and (iii) it should not lie on the causal pathway between exposure and disease. In the MIREC analysis, we studied the effects of the chemical contaminants on the health outcomes based on logistic regression models, where we considered adjusting the models for the effects of some demographic variables, which included the binary indicators being a first-time mother (parity), mother's smoking status, and prepregnancy BMI. The variables to be included in the models were chosen by the likelihood-based Akaike information criterion for model selection, rather than the Change-in-Estimate criterion²⁴ for identifying confounders. The effects of the contaminants examined in the paper remain after adjusting for those variables. Further research needs to be done to evaluate the best approach for identifying and controlling for potential confounders in models with multiple chemical exposures.

Note that, for describing the joint distribution of the binary indicators $\mathbf{v} = (v_1, \dots, v_p)'$ of nondetects, we have used a multivariate Bahadur model for correlated binary data. The Bahadur model is attractive in that, under this model, the marginal distribution of the individual binary outcome v_j is a simple Bernoulli distribution with the probability of success π_j . In fact, when the associations among the binary indicators are not strong enough, the joint Bahadur distribution may be simply approximated by the product of individual Bernoulli distributions of the binary indicators (v_1, \dots, v_p) . The choice of the association parameters for the Bahadur model is a practical issue. For left-censored covariates as considered in this article, we recommend a simple "working correlation" structure, such as the exchangeable correlation structure, for the multivariate binary outcomes. We, however, have not studied in detail yet how the proposed W method under a "misspecified Bahadur model" would compare with the NW method, where the NW method provides consistent estimators irrespective of the distribution of the binary indicators of nondetects. From a limited simulation study (not shown here), we found that, in the case of weak correlations among the binary indicators, the assumption of "working independence" also leads to efficient estimators of the model parameters.

An important feature of the chemical exposure data is that measurements on the exposures included a number of extreme observations. It would be interesting to investigate how the classical estimators are influenced by the extreme observations or "outliers" in the data. In the presence of "influential outliers" in the data, a robust method may be explored to bound the influence of such outliers. Work remains to be done in this direction. We intend to develop a robust method for censored data in future research.

ACKNOWLEDGEMENTS

The authors thank the referees and the associate editor for very helpful comments and suggestions that led to significant improvements of the manuscript. We also thank all the MIREC participants and the staff at the coordinating center and each recruitment site, as well as the MIREC study group. The MIREC study was funded by the Health Canada's Chemicals Management Plan, the Canadian Institute of Health Research (grant # MOP - 81285), and the Ontario Ministry of the

Environment. Sanjoy Sinha is grateful for the support provided by a grant from the Natural Sciences and Engineering Research Council of Canada.

ORCID

Wan-Chen Lee  <http://orcid.org/0000-0002-8798-5695>

REFERENCES

1. Cole SR, Chu H, Nie L, Schisterman EF. Estimating the odds ratio when exposure has a limit of detection. *Int J Epidemiol*. 2009;38:1674-1680.
2. Thompson M, Nelson KP. Linear regression with Type I interval- and left-censored response data. *Environ Ecol Stat*. 2003;10:221-230.
3. Helsel DR. *Statistics for Censored Environmental Data Using Minitab and R*. 2nd ed. Hoboken, NJ: John Wiley & Sons; 2012.
4. Helsel DR. Fabricating data: how substituting values for nondetects can ruin results, and what can be done about it. *Chemosphere*. 2006;65:2434-2439.
5. Herring AH. Nonparametric Bayes shrinkage for assessing exposures to mixtures subject to limits of detection. *Epidemiology*. 2010;21:S71-S76.
6. May RC, Ibrahim JG, Chu H. Maximum likelihood estimation in generalized linear models with multiple covariates subject to detection limits. *Statist Med*. 2011;30:2551-2561.
7. Sattar A, Sinha SK, Morris NJ. A parametric survival model when a covariate is subject to left-censoring. *J Biom Biostat*. 2012;S3:002.
8. Sattar A, Sinha SK, Wang X, Li Y. Frailty models for pneumonia to death with a left-censored covariate. *Statist Med*. 2015;34:2266-2280.
9. Bernhardt PW, Wang HJ, Zhang D. Statistical models for generalized linear models with covariates subject to detection limits. *Stat Biosci*. 2015;7:68-89.
10. Holstein CA, Griffin M, Hong J, Sampson PD. Statistical method for determining and comparing limits of detection of bioassays. *Anal Chem*. 2015;87:9795-9801.
11. LaFleur B, Lee W, Billhiemer D, Lockhart C, Liu J, Merchant N. Statistical methods for assays with limits of detection: serum bile acid as a differentiator between patients with normal colons, adenomas, and colorectal cancer. *J Carcinog*. 2011;10:12.
12. Lubin JH, Colt JS, Camann D, et al. Epidemiologic evaluation of measurement data in the presence of detection limits. *Environ Health Perspect*. 2004;112:1691-1696.
13. Wood MD, Beresford NA, Coplestone D. Limit of detection values in data analysis: Do they matter? *Radioprotection*. 2011;46(6):S85-S90.
14. Richardson DB, Ciampi A. Effects of exposure measurement error when an exposure variable is constrained by a lower limit. *Am J Epidemiol*. 2003;157:355-363.
15. Schisterman E, Vexler A, Whitcomb B, Liu A. The limitations due to exposure detection limits for regression models. *Am J Epidemiol*. 2006;163:374-383.
16. 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;21:S17-S24.
17. Arbuckle TE, Fraser WD, Fisher M, et al. Cohort profile: the maternal-infant research on environmental chemicals research platform. *Paediatr Perinat Epidemiol*. 2013;27:415-425.
18. McCullagh P, Nelder JA. *Generalized Linear Models*. 2nd ed. London, UK: Chapman & Hall/CRC; 1989.
19. Bahadur RR. A representation of the joint distribution of responses to n dichotomous items. In: *Studies in Item Analysis and Prediction*. Stanford, CA: Stanford University Press; 1961:158-168.
20. McCulloch CE. Maximum likelihood algorithms for generalized linear mixed models. *J Am Stat Assoc*. 1997;92:162-170.
21. Shapiro GD, Dodds L, Arbuckle TE, et al. Exposure to phthalates, bisphenol A and metals in pregnancy and the association with impaired glucose tolerance and gestational diabetes mellitus: the MIREC study. *Environ Int*. 2015;83:63-71.
22. Thomas S, Arbuckle TE, Fisher M, Fraser WD, Ettinger A, King W. Metals exposure and risk of small-for-gestational age birth in a Canadian birth cohort: the MIREC study. *Environ Res*. 2015;140:430-439.
23. Rothman KJ. *Modern Epidemiology*. Boston, MA: Little, Brown & Co; 1986.
24. Lee PH. Is a cutoff of 10% appropriate for the change-in-estimate criterion of confounder identification? *J Epidemiol*. 2014;24(2):161-167.

How to cite this article: Lee W-C, Sinha SK, Arbuckle TE, Fisher M. Estimation in generalized linear models under censored covariates with an application to MIREC data. *Statistics in Medicine*. 2018;37:4539-4556. <https://doi.org/10.1002/sim.7942>

APPENDIX

Tables A1–A5: Simulation results for estimators of nuisance parameters

TABLE A1 Empirical biases and mean squared errors (MSEs) of estimators of nuisance parameters μ_{x_1} and σ_{x_1} for two proportions (0.3, 0.5) of left-censored (limit of detection (LOD)) covariate x_1 . Response $y_i \sim \text{ind. Bernoulli}(\mu_i)$ with the logit link $\log(\mu_i/(1 - \mu_i)) = \beta_0 + \beta_1 x_{1i}$ and covariate $x_{1i} \sim \text{ind. Normal}(\mu_{x_1}, \sigma_{x_1}^2)$. Parameters are fixed at $\beta_0 = -2$, $\beta_1 = 0.5$, $\mu_{x_1} = 0$, and $\sigma_{x_1}^2 = 1$. (N = Naive method; NW = Nonweighted method; W = Weighted method)

n	LOD	Method	Bias		MSE	
			$\hat{\mu}_{x_1}$	$\hat{\sigma}_{x_1}$	$\hat{\mu}_{x_1}$	$\hat{\sigma}_{x_1}$
100	0.3	N	-0.4809	0.5979	0.2552	0.3625
		NW	-0.0021	-0.0015	0.0109	0.0079
		W	-0.0031	-0.0008	0.0111	0.0080
	0.5	N	-0.8493	0.6959	0.7491	0.4867
		NW	-0.0031	-0.0020	0.0152	0.0120
		W	-0.0045	-0.0013	0.0153	0.0120
200	0.3	N	-0.4801	0.5988	0.2437	0.3611
		NW	-0.0006	-0.0039	0.0059	0.0040
		W	-0.0011	-0.0036	0.006	0.0040
	0.5	N	-0.8515	0.6982	0.7396	0.4887
		NW	-0.0013	-0.0038	0.0076	0.0058
		W	-0.0022	-0.0034	0.0076	0.0058
500	0.3	N	-0.4805	0.6030	0.2357	0.3646
		NW	-0.0003	-0.0004	0.0021	0.0016
		W	-0.0003	-0.0004	0.0022	0.0016
	0.5	N	-0.8511	0.7017	0.7300	0.4929
		NW	-0.0013	0.0002	0.0029	0.0024
		W	-0.0014	0.0003	0.0030	0.0025

TABLE A2 Empirical biases and mean squared errors (MSEs) of estimators of nuisance parameters μ_{x_1} , μ_{x_2} , σ_{x_1} , σ_{x_2} , and $\rho_{x_1, x_2} = \text{corr}(x_1, x_2)$ for two proportions (0.3, 0.5) of left-censored (limit of detection (LOD)) covariates x_1 and x_2 . Response $y_i \sim \text{ind. Bernoulli}(\mu_i)$ with the logit link $\log(\mu_i/(1 - \mu_i)) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$ and (x_{1i}, x_{2i}) are bivariate normal with means (μ_{x_1}, μ_{x_2}) and covariance terms $\sigma_{x_1}^2$, $\sigma_{x_2}^2$, and σ_{x_1, x_2} . Parameters are fixed at $\beta_0 = -2$, $\beta_1 = 0.5$, $\beta_2 = 1$, $\mu_{x_1} = 0$, $\mu_{x_2} = 2$, $\sigma_{x_1}^2 = 1$, $\sigma_{x_2}^2 = 2$, and $\sigma_{x_1, x_2} = 0.5$. (N = Naive method; NW = Nonweighted method; W = Weighted method)

n	LOD	Method	Bias					MSE				
			$\hat{\mu}_{x_1}$	$\hat{\sigma}_{x_1}$	$\hat{\mu}_{x_2}$	$\hat{\sigma}_{x_2}$	$\hat{\rho}_{x_1, x_2}$	$\hat{\mu}_{x_1}$	$\hat{\sigma}_{x_1}$	$\hat{\mu}_{x_2}$	$\hat{\sigma}_{x_2}$	$\hat{\rho}_{x_1, x_2}$
100	0.3	N	-0.4801	0.5899	-0.4243	0.7348	-0.0321	0.2555	0.3558	0.2622	0.5895	0.0101
		NW	-0.0052	-0.0020	-0.0019	-0.0087	-0.0006	0.0108	0.0078	0.0447	0.0272	0.0098
		W	-0.0063	-0.0014	-0.0039	-0.0072	-0.0006	0.0113	0.0079	0.0467	0.0283	0.0098
	0.5	N	-0.8432	0.6846	-1.6719	1.3656	-0.0549	0.7416	0.4751	2.9311	1.8900	0.0130
		NW	-0.0054	-0.0034	0.0013	-0.0170	-0.0015	0.0150	0.0124	0.0637	0.0483	0.0135
		W	-0.0072	-0.0026	-0.0019	-0.0157	-0.0016	0.0151	0.0125	0.0645	0.0487	0.0135
200	0.3	N	-0.4831	0.6019	-0.4362	0.7542	-0.0327	0.2463	0.3649	0.2309	0.5900	0.0055
		NW	-0.0024	-0.0001	-0.0066	-0.0068	0.0006	0.0057	0.0041	0.0221	0.0132	0.0048
		W	-0.0031	0.0003	-0.0074	-0.0062	0.0006	0.0058	0.0041	0.0228	0.0135	0.0048
	0.5	N	-0.8542	0.6999	-1.7114	1.3949	-0.0512	0.7440	0.4911	2.9878	1.9514	0.0072
		NW	-0.0046	0.0013	-0.0096	-0.0061	0.0036	0.0076	0.0062	0.0312	0.0265	0.0064
		W	-0.0057	0.0018	-0.0110	-0.0056	0.0035	0.0078	0.0062	0.0316	0.0266	0.0064
500	0.3	N	-0.4797	0.6032	-0.4373	0.7671	-0.0333	0.2352	0.3649	0.2066	0.5974	0.0029
		NW	0.0004	-0.0002	-0.0053	0.0018	-0.0013	0.0023	0.0015	0.0084	0.0053	0.0020
		W	0.0003	-0.0001	-0.0060	0.0024	-0.0013	0.0023	0.0015	0.0086	0.0056	0.0020
	0.5	N	-0.8502	0.7021	-1.7059	1.4027	-0.0551	0.7286	0.4935	2.9343	1.9696	0.0049
		NW	-0.0004	0.0004	-0.0046	0.0001	-0.0014	0.0030	0.0024	0.0123	0.0093	0.0026
		W	-0.0007	0.0006	-0.0050	0.0003	-0.0014	0.0030	0.0024	0.0125	0.0093	0.0026

TABLE A3 Empirical biases and mean squared errors (MSEs) of estimators of nuisance parameters γ_{x_1} and λ_{x_1} for two proportions (0.3, 0.5) of left-censored (limit of detection (LOD)) covariate x_1 . Response $y_i \sim \text{ind. Bernoulli}(\mu_i)$ with the logit link $\log(\mu_i/(1 - \mu_i)) = \beta_0 + \beta_1 x_{1i}$ and covariate $x_{1i} \sim \text{ind. Gamma}(\gamma_{x_1}, \lambda_{x_1})$. Parameters are fixed at $\beta_0 = -1$, $\beta_1 = 1$, $\gamma_{x_1} = 4$, and $\lambda_{x_1} = 2$. (N = Naive method; NW = Nonweighted method; W = Weighted method)

n	LOD	Method	Bias		MSE	
			$\hat{\gamma}_{x_1}$	$\hat{\lambda}_{x_1}$	$\hat{\gamma}_{x_1}$	$\hat{\lambda}_{x_1}$
100	0.3	N	-0.9195	-0.3853	0.9429	0.1796
		NW	0.1620	0.0830	0.6034	0.1541
		W	0.1516	0.0791	0.6007	0.1527
	0.5	N	-0.7879	-0.2588	0.6895	0.1065
		NW	0.1507	0.0723	0.9154	0.2169
		W	0.1431	0.0696	0.9114	0.2160
200	0.3	N	-0.9551	-0.4011	0.9605	0.1771
		NW	0.0769	0.0444	0.2878	0.0753
		W	0.0734	0.0431	0.2913	0.0758
	0.5	N	-0.8024	-0.2643	0.6781	0.0893
		NW	0.0775	0.0415	0.4442	0.1030
		W	0.0744	0.0404	0.4432	0.1027
500	0.3	N	-0.9688	-0.4142	0.9577	0.1774
		NW	0.0416	0.0200	0.1124	0.0277
		W	0.0396	0.0193	0.1139	0.0278
	0.5	N	-0.8171	-0.2777	0.6811	0.0842
		NW	0.0329	0.0159	0.1748	0.0390
		W	0.0320	0.0155	0.1755	0.0390

TABLE A4 Empirical biases and mean squared errors (MSEs) of estimators nuisance parameters μ_{x_1} and σ_{x_1} for two proportions (0.3, 0.5) of left-censored (limit of detection (LOD)) covariate x_1 . Response $y_i \sim \text{ind. Normal}(\mu_i, \sigma^2)$ with $\mu_i = \beta_0 + \beta_1 x_{1i}$ and covariate $x_{1i} \sim \text{ind. Normal}(\mu_{x_1}, \sigma_{x_1}^2)$. Parameters are fixed at $\beta_0 = 1$, $\beta_1 = 2$, $\sigma = 1$, $\mu_{x_1} = 2$, and $\sigma_{x_1} = 1$. (N = Naive method; NW = Nonweighted method; W = Weighted method)

n	LOD	Method	Bias		MSE	
			$\hat{\mu}_{x_1}$	$\hat{\sigma}_{x_1}$	$\hat{\mu}_{x_1}$	$\hat{\sigma}_{x_1}$
40	0.3	N	-0.0337	-0.0206	0.0258	0.0091
		NW	-0.0016	-0.0210	0.0271	0.0183
		W	-0.0053	-0.0197	0.0279	0.0186
	0.5	N	-0.0979	-0.0239	0.0353	0.0083
		NW	0.0005	-0.0221	0.0335	0.0227
		W	-0.0013	-0.0221	0.0343	0.0227
60	0.3	N	-0.0370	-0.0176	0.0187	0.0059
		NW	-0.0031	-0.0190	0.0188	0.0122
		W	-0.0053	-0.0181	0.0192	0.0122
	0.5	N	-0.0981	-0.0134	0.0255	0.0053
		NW	-0.0032	-0.0056	0.0231	0.0174
		W	-0.0052	-0.0053	0.0238	0.0174
100	0.3	N	-0.0333	-0.0092	0.0114	0.0035
		NW	0.0015	-0.0111	0.0112	0.0072
		W	-0.0014	-0.0100	0.0118	0.0073
	0.5	N	-0.1026	-0.0128	0.0204	0.0032
		NW	-0.0039	-0.0073	0.0141	0.0094
		W	-0.0042	-0.0074	0.0144	0.0094

TABLE A5 Estimates of nuisance parameters (standard errors in parentheses) for left-censored covariates in logistic regression fits to MIREC data. (N = Naive method; NW = Nonweighted method; W = Weighted method)

y	x ₁	x ₂	Method	$\hat{\mu}_{x_1}$	$\hat{\sigma}_{x_1}$	$\hat{\mu}_{x_2}$	$\hat{\sigma}_{x_2}$
BO	COTISE	PFHxS	N	-3.316 _(0.063)	2.571 _(0.045)	0.042 _(0.019)	0.770 _(0.013)
			NW	-3.578 _(0.082)	2.637 _(0.059)	0.039 _(0.019)	0.776 _(0.014)
			W	-2.372 _(0.066)	3.056 _(0.071)	0.051 _(0.017)	0.766 _(0.012)
OGTT_1	DMAA	DMP	N	-3.450 _(0.028)	0.911 _(0.020)	1.102 _(0.030)	0.976 _(0.021)
			NW	-3.534 _(0.032)	1.045 _(0.025)	0.955 _(0.038)	1.191 _(0.031)
			W	-3.532 _(0.027)	1.044 _(0.023)	0.951 _(0.031)	1.194 _(0.028)
OGTT_2	MEOHP	MEHHP	N	1.813 _(0.035)	1.160 _(0.025)	2.176 _(0.037)	1.224 _(0.026)
			NW	1.804 _(0.035)	1.174 _(0.023)	2.170 _(0.037)	1.235 _(0.026)
			W	1.804 _(0.035)	1.174 _(0.022)	2.172 _(0.037)	1.232 _(0.025)
LBW	OXYCHLOR	COTISE	N	-4.375 _(0.013)	0.533 _(0.009)	-3.325 _(0.063)	2.562 _(0.045)
			NW	-4.395 _(0.014)	0.573 _(0.011)	-3.584 _(0.082)	2.623 _(0.058)
			W	-4.411 _(0.013)	0.586 _(0.010)	-2.377 _(0.066)	3.049 _(0.071)
LGA	MBzP	DMAA	N	1.657 _(0.033)	1.301 _(0.024)	-3.456 _(0.023)	0.895 _(0.016)
			NW	1.650 _(0.034)	1.313 _(0.024)	-3.539 _(0.027)	1.028 _(0.021)
			W	1.652 _(0.033)	1.311 _(0.022)	-3.542 _(0.023)	1.030 _(0.019)
SGA	PFOA	ASAL	N	0.500 _(0.016)	0.592 _(0.011)	-4.047 _(0.042)	1.594 _(0.030)
			NW	0.501 _(0.016)	0.589 _(0.011)	-4.664 _(0.078)	2.340 _(0.071)
			W	0.499 _(0.016)	0.594 _(0.011)	-4.676 _(0.056)	1.030 _(0.019)
PreB	TRANSONA	BPC170	N	-3.971 _(0.013)	0.527 _(0.009)	-4.237 _(0.014)	0.557 _(0.010)
			NW	-4.023 _(0.015)	0.607 _(0.012)	-4.557 _(0.026)	0.890 _(0.023)
			W	-4.015 _(0.013)	0.602 _(0.011)	-4.537 _(0.019)	0.882 _(0.022)

Abbreviations: ASAL, arsenobetaine; BO, birth outcome; DMAA, dimethylarsinic acid; DMP, dimethylphosphate; COTISE, cotinine; LBW, low birth weight; LGA, large for gestational age; MBzP, mono benzyl phthalate; MEHHP, mono-(2-ethyl-5-hydroxyhexyl) phthalate; MEOHP, mono-(2-ethyl-5-oxohexyl) phthalate; PFOA, perfluorooctanoic acid; PreB, preterm birth; OXYCHLOR, oxychlorane; SGA, small for gestational age; TRANSONA, trans-nonachlor.