




# Statistical Methods for Modeling Exposure Variables Subject to Limit of Detection

Eunsil Seok<sup>1</sup> · Akhgar Ghassabian<sup>1,2</sup> · Yuyan Wang<sup>1</sup> · Mengling Liu<sup>1</sup> 

Received: 27 December 2022 / Revised: 19 August 2023 / Accepted: 12 October 2023 /

Published online: 28 November 2023

© The Author(s) under exclusive licence to International Chinese Statistical Association 2023

## Abstract

Environmental health research aims to assess the impact of environmental exposures, making it crucial to understand their effects due to their broad impacts on the general population. However, a common issue with measuring exposures using bio-samples in laboratory is that values below the limit of detection (LOD) are either left unreported or inaccurately read by machines, which subsequently influences the analysis and assessment of exposure effects on health outcomes. We address the challenge of handling exposure variables subject to LOD when they are treated as either covariates or an outcome. We evaluate the performance of commonly-used methods including complete-case analysis and fill-in method, and advanced techniques such as multiple imputation, missing-indicator model, two-part model, Tobit model, and several others. We compare these methods through simulations and a dataset from NHANES 2013–2014. Our numerical studies show that the missing-indicator model generally yields reasonable estimates when considering exposure variables as covariates under various settings, while other methods tend to be sensitive to the LOD-missing proportions and/or distributional skewness of exposures. When modeling an exposure variable as the outcome, Tobit model performs well under Gaussian distribution and quantile regression generally provides robust estimates across various shapes of the outcome's distribution. In the presence of missing data due to LOD, different statistical models should be considered for being aligned with scientific questions, model assumptions, requirements of data distributions, as well as their interpretations. Sensitivity analysis to handle LOD-missing exposures can improve the robustness of model conclusions.

**Keywords** Environmental exposure · Missing data · Multiple imputation · NHANES · Tobit model · Two-part model

## 1 Introduction

Environmental exposures are ubiquitous and crucial to individual and population health. Assessing how exposures are associated with biomedical features is of great interest in clinical research, public health, and health policy. One commonly encountered issue is that measurements of exposures are often subject to the limitation of machine-reading sensors. Specifically, it is constrained by a value, which is known as the limit of detection (LOD), and any observation below the LOD value is left either undetermined or deemed with low accuracy. As a result, data obtained from such measurements are either recorded as LOD values, imprecise machine reading values, or remain missing. Such measurement issues can result in misleading or erroneous scientific conclusions, and research found that different methods of handling the LOD issue in exposure variables could yield inconsistent results [1, 2]. Indeed, this issue widely exists beyond environmental health research and has been prevalent in various fields such as genetics, clinical diagnosis, and biomarker studies [3–6].

Various analytical methods have been proposed to handle measurements affected by LOD, and reviews and methodological literature have considered LOD-missing variable as a covariate [1, 4, 7, 8] or an outcome [9–12]. The simplest approach is complete-case analysis, using only samples above the LOD, but often leads to a significant loss of data and/or biased samples. Single value substitution methods by replacing missing data with a single fixed value (e.g., LOD, LOD/2, or LOD/ $\sqrt{2}$ ) can underestimate data variability [13] and may introduce bias [14]. Alternative substituting methods such as using conditional expectations  $\mathbb{E}(X | X \leq \text{LOD})$  [15] and  $\mathbb{E}(X | X > \text{LOD})$  [16] have been suggested. Another effective approach is multiple imputations, which generate completed datasets with plausible values randomly imputed for missing observations and then combine results from these datasets [17]. Baccarelli et al. [18] proposed an alternative multiple imputation method by imputing values based on a conditioned parametric distributional assumption, ensuring all imputed values strictly fall below the LOD. However, Arunajadai and Rauh [19] showed that this method may result in attenuated estimates.

Furthermore, several advanced statistical models and approaches have been suggested to handle such LOD-missing variables. When such variable is treated as covariate, the maximum likelihood (ML) approach assumes a parametric distribution for the LOD-missing variable and estimates its association with the outcome by maximizing the observed-data likelihood directly [4, 5] or using the EM algorithm [3]. Chiou et al. [6] and Ortega-Villa et al. [1] considered a missing-indicator model which includes an indicator variable of whether the exposure is observed or not to represent the binary effect of above and below the LOD, as well as a continuous exposure variable to characterize the effect above the LOD.

In addition, it is of scientific interest to comprehend the factors influencing the variations in exposure levels among individuals [2, 20–22], and thus an exposure variable is modeled as an outcome (*i.e.*, dependent variable). Two different perspectives of treating the missingness can be taken in constructing the statistical

models. The two-part model treats the LOD-missing variable by separately modeling its probability of being above the LOD and its continuous relationship with other covariates when it is observed. On the other hand, Tobit model [23] considers LOD-missing variable as left-censored and estimates the effects of covariates on the underlying true exposure-outcome. A parallel perspective arises when the focus shifts to the impact of covariates on a specific quantile of the exposure-outcome. In this context, the two-part quantile model [24] differentiates between missing and observed values, whereas quantile regression [25] handles the LOD-missing variable by treating it as left-censored.

In this paper, we aim to compare various methods of modeling exposure variables subject to LOD either as covariates or an outcome. The comparisons are based on the real application using the National Health and Nutrition Examination Survey (NHANES) data and a series of extensive simulation studies. The remainder of the paper is organized as follows. In Sect. 2, we describe the details of the motivated NHANES dataset. Section 3 gives the different models for treating exposure variables as covariates or an outcome. Section 4 presents data analysis performed on the NHANES dataset and illustrates results. In Sect. 5, we effectively compare and evaluate the performances of the methods through a series of simulation studies under known data-generating settings. Finally, Sect. 6 provides a summary of the strengths and weaknesses of these methods under different scenarios, along with recommendations.

## 2 Dataset

NHANES [26] is a program designed to assess the health and nutritional status of adults and children in the United States. Assessments consist of interviews and physical examinations to obtain demographic, socioeconomic, dietary, and health-related information. We used data from NHANES 2013–2014 and merged tables on demographic, physical examination, laboratory measures, and questionnaires. This resulted in a dataset with three chemical exposure variables, body mass index (BMI) as the health outcome variable, and seven variables representing individual characteristics as covariates. We excluded samples with missingness in BMI or other individual characteristics, resulting in a total of 2046 participants for further analysis. The data flow diagram for deriving the dataset is in Figure A1.

### 2.1 Exposure Variables

Bisphenols are phenolic organic compounds, commonly used in the manufacturing of daily used items for the hardening of plastic and are easily dissolved in foods and drinks. Accordingly, humans always have a concentration of bisphenols in their body fluids such as urine or blood, even without intentional exposure to those. Nevertheless, bisphenols, particularly bisphenol A (BPA), are known to have adverse health effects in humans (e.g., obesity [27, 28]). In recent years, the use of BPA in baby products has been prohibited due to the concern about its potential toxicity

[29] and BPA is increasingly replaced by structurally similar chemicals, such as bisphenol F (BPF) and bisphenol S (BPS). However, these replacements are also known to be detrimental to human health [30, 31].

Our analysis uses urinary concentrations of BPA, BPS, and BPF from the laboratory dataset. The measurements were adjusted for the urinary creatinine concentration to account for variations in urinary dilution, and levels are reported in the unit of  $\mu\text{g/L}$ . The LOD value was 0.2  $\mu\text{g/L}$  for BPA, 0.1  $\mu\text{g/L}$  for BPS, and 0.2  $\mu\text{g/L}$  for BPF. The percentages of missing data due to values below the LOD were 4.59% for BPA, 10.17% for BPS, and 34.36% for BPF. Table 1 provides descriptive statistics of exposure variables including median, interquartile range (IQR), LOD values, and missing rates.

We log-transformed the bisphenol data first due to the skewed distribution and standardized all exposure measurements to achieve comparable interpretations of their effects per standard deviation (SD) change across exposures. Figure A2 provides histograms of original (left) and log-transformed and standardized (right) exposures.

## 2.2 Health Outcome

The primary health outcome variable is BMI from the examination dataset. The body measures data were collected in the Mobile Examination Center (MEC) by trained health technicians, and BMI was calculated as weight in kilograms divided by the square of height in meters.

## 2.3 Covariates

We incorporated seven covariates including an individual's gender, age, race/ethnicity, education level, income to poverty ratio, the number of days of moderate workout in a week, and urinary creatinine level. Race/ethnicity was grouped as non-Hispanic Whites, Mexican American, other Hispanic, non-Hispanic Black, non-Hispanic Asian, and Others. Socioeconomic status was indicated using the Income to Poverty Ratio (IPR), calculated by dividing the total family income by the poverty threshold. Moderate workout days per week were obtained from answers to the question "How much time do you spend doing moderate-intensity activities at work on a typical day?" in the questionnaire data. Table 2 summarizes characteristics of included subjects using mean with SD for continuous variables, and count with percent for categorical variables. Particularly, we used median and IQR for the urinary creatinine level due to its right-skewed distribution.

**Table 1** Descriptive statistics of exposure variables: median, IQR, LOD value, and below LOD rate

Exposure	Median	IQR	LOD	% < LOD
BPA ( $\mu\text{g/L}$ )	1.4	2.0	0.2	4.59
BPS ( $\mu\text{g/L}$ )	0.5	0.9	0.1	10.17
BPF ( $\mu\text{g/L}$ )	0.7	1.3	0.2	34.36

**Table 2** Summary of characteristics of included subjects from NHANES 2013–2014 dataset

Variable	Total ( $N = 2046$ )
Gender, n(%)	
Male, baseline	971 (47.46)
Female	1075 (52.54)
Age (years) at screening, mean (SD)	42.63 (20.77)
Race/ethnicity, n(%)	
Non-Hispanic White, baseline	812 (39.69)
Mexican American	298 (14.57)
Other Hispanic	187 (9.14)
Non-Hispanic Black	451 (22.04)
Non-Hispanic Asian	227 (11.09)
Other	71 (3.47)
Education level, n(%)	
Less than 9th grade, baseline	297 (14.52)
9–11th grade	417 (20.38)
High school graduate/GED or equivalent	404 (19.74)
Some college or AA degree	527 (25.76)
College graduate and above	401 (19.60)
Income to poverty ratio (IPR), mean (SD)	2.38 (1.65)
Moderate work days per week, mean (SD)	1.43 (2.25)
Urinary creatinine level, median (IQR)	116 (109)
Body Mass Index (BMI, $kg/m^2$ ), mean (SD)	28.15 (7.17)

### 3 Methods

In this section, we first consider each exposure as a single covariate for its marginal effect, and then all three exposures simultaneously as multiple covariates to study the joint effects. Furthermore, we examine methods to model the exposures as the outcome of interest. We denote the continuous outcome by  $Y \in \mathbb{R}$ ,  $T^* \in \mathbb{R}^q$  for the underlying  $q$  true exposure variables, and  $Z \in \mathbb{R}^p$  as  $p$ -dimensional confounders. Subject to the LOD, we only observe the exposure variable  $T$  as the left-censored version of  $T^*$  and specify it as follows.

$$T_j = \begin{cases} T_j^* & T_j^* > \text{LOD}_j \\ \text{LOD}_j & T_j^* \leq \text{LOD}_j \end{cases}, \quad j = 1, \dots, q. \quad (1)$$

The observed dataset consists of  $\{(y_i, t_i, z_i)\}_{i=1}^N$  of  $N$  independent and identically distributed (i.i.d.) samples, which are realizations of random variables  $(Y, T, Z)$ . We slightly abuse the notation by using  $T$  for both observed and standardized exposure values.

### 3.1 Single Exposure as a Covariate

Consider a linear model characterizing the relationship between the continuous health outcome (e.g., BMI) and a single exposure, *i.e.*,  $q = 1$ , as

$$Y = \beta_0 + \beta_1 T^* + \alpha^T Z + \epsilon, \quad (2)$$

where  $\epsilon$  is a random error i.i.d. from  $N(0, \sigma_y^2)$  and independent of  $T^*$  and  $Z$ . The parameters  $\beta_0 \in \mathbb{R}$ ,  $\beta_1 \in \mathbb{R}$ , and  $\alpha \in \mathbb{R}^p$  represent the intercept, exposure's and covariates' effects on the mean of the outcome, respectively. We refer to this model as the single exposure model. It is often employed as an initial screening step for potentially inter-correlated exposure markers and is used to assess the impact of missingness resulting from LOD on estimating the marginal effect of each individual exposure. By focusing specifically on single exposures, we can gain insights into the challenges and approaches involved in estimating the effects of covariates while accounting for missing due to LOD. Note that the model can be readily formulated as a generalized linear model to handle other types of outcomes, such as binary data or count data. We examine five methods of handling missingness due to LOD under this context.

- (i) Complete-case analysis uses only the samples with complete information in all variables for the analysis, which subsequently leads to reduced sample size. Though this approach is simple and straightforward, it could yield biased results when data violates the missing completely at random (MCAR) assumption [32]. Also, the model could become unstable when the sample size is small.
- (ii) Fill-in method is to substitute missing values with a single fixed value, such as  $\text{LOD}/\sqrt{2}$ , and then proceed with the model fitting.
- (iii) Maximum-likelihood (ML) approach assumes that the exposure variable follows a parametric distribution, e.g., normal distribution as

$$T^* | Z \sim N(\gamma^T Z, \sigma_t^2), \quad (3)$$

where coefficients  $\gamma$  represent covariates' effects on the exposure. Thus, under models (2) and (3) and parametric assumption for  $\epsilon$ , the likelihood function  $L(\theta)$  of parameter vector  $\theta = (\beta_0, \beta_1, \alpha^T, \gamma^T, \sigma_y^2, \sigma_t^2)^T$  can be specified as

$$L(\theta) = \prod_{i=1}^N \left\{ f(Y_i = y_i | T_i^* = t_i, Z_i = z_i) f(T_i^* = t_i | Z_i = z_i) \right\}^{(1-\delta_i)} \times \left\{ \int_{-\infty}^{\text{LOD}} f(Y_i = y_i | T_i^* = u, Z_i = z_i) f(T_i^* = u | Z_i = z_i) du \right\}^{\delta_i}, \quad (4)$$

where  $f(\cdot)$  is the probability density function and  $\delta_i = I(t_i \leq \text{LOD})$  is the indicator of whether the exposure is below the LOD or not. Further details on the log-likelihood function are in Appendix B. The likelihood maximization can be achieved through the utilization of the `optim` function in R or via the EM algorithm [3]. When dealing with a convex likelihood function, the EM algorithm

is renowned for its ability to find solutions, but it may have a relatively slow convergence rate. Thus, we chose to directly use the `optim` function in R that provides a more streamlined implementation for our ML approach, allowing for direct utilization of the likelihood function and offering faster convergence in optimization procedures. We obtained the initial values of parameter estimation from the complete-case analysis and supplied them to the `optim` function.

- (iv) Missing-indicator model [6] includes an indicator variable to denote whether an exposure variable is observed or not, along with the continuous exposure when it is above the LOD. Ortega-Villa et al. [1] showed that the missing-indicator model performs well in analyzing exposures subject to LOD. The model is specified as:

$$Y = \beta_0 + \beta_1(1 - \delta)(T - \text{LOD}) + \beta_2\delta + \alpha^T Z + \epsilon,$$

where  $\delta$  is as previously defined in (4). When the exposure is below the LOD, its effect on the outcome is characterized by  $\beta_2$ , and the effect of a unit change in the exposure above the LOD is represented by  $\beta_1$ .

- (v) Multiple imputation method [17] generates multiple completed datasets, each with different plausible values imputed for the missing observations, and then combines multiple results from those datasets. Multiple imputation consists of three steps: (1) the imputation step: plausible values are sampled from a suitable distribution to fill in the missing observations, resulting in  $m$  completed datasets; (2) the completed-data analysis step: model (2) is fitted separately on each of the  $m$  completed datasets, and (3) the pooling step: the  $m$  results are combined to yield the overall estimate which is the average of individual estimates. The overall standard error takes into account the uncertainty of both the sampled values and the imputation process. Multiple imputation is known to produce asymptotically unbiased estimates and standard errors and is asymptotically efficient [33]. To implement multiple imputation, we use the `mice` function in the `mice` package in R [34] with the number of imputations,  $m = 5$ .

### 3.2 Multiple Exposures as Covariates

When the joint effects of multiple exposures are of interest, we include multiple exposure variables in one model simultaneously and name this model the multiple exposures model. We consider the linear relationship between a continuous health outcome and multiple exposures,  $T^* = (T_1^*, \dots, T_q^*)$ ,

$$Y = \beta_0 + \beta_1^T T^* + \alpha^T Z + \epsilon.$$

where  $Y$ ,  $Z$ , and  $\epsilon$  are as defined in model (2).

To handle multiple exposure variables subject to LOD in one model, we consider four methods: (i) complete-case analysis, (ii) fill-in method, (iii) missing-indicator model, and (iv) multiple imputation. We forewent the ML approach due to its rapidly increasing complexity and its heavy reliance on parametric distributional

assumptions. Except for the multiple imputation, other methods are similar to what we have discussed in the single exposure model in Sect. 3.1.

**Multiple imputation** method remains the same three steps in principle as Sect. 3.1 (v). However, for multiple missing exposures, the first imputation step becomes more complicated and challenging. There are two general approaches for imputing multivariate data: joint modeling (JM, Schafer [35]) and multiple imputation by chained equations (MICE, Van Buuren and Oudshoorn [36]). The JM requires a parametric multivariate distribution for the missing data, while MICE (also known as a fully conditional specification FCS, Van Buuren et al. [37]) relaxes this assumption by defining the imputation model on a variable-by-variable basis through a set of conditional densities. The idea of splitting a  $q$ -dimensional problem into  $q$  one-dimensional problems allows MICE to flexibly handle different types of variables since the appropriate model can be selected for different types of variables. Moreover, a small number of iterations (e.g., 5 to 20) is known to be sufficient [17, 38]. Due to its versatility and practical advantages, MICE is rapidly emerging as a commonly used method for handling missing data. Further mathematical details of MICE are in Van Buuren et al. [37]. We implement the MICE method using the `mice` package in R [34], with the number of imputations  $m = 5$ .

### 3.3 Modeling Exposure as Outcome

When considering the LOD-missing variable as an outcome, we can view this from two different perspectives: zero-inflated outcome or left-censored outcome. If non-detected values are recorded as the LOD value as in (1), the distribution of the exposure variable would have a peak at the LOD value. Without loss of generality, assuming the LOD value to be zero, the LOD-missing variable can be viewed as zero-inflated data. On the other hand, if we regard non-detected values as missing and left-censored at the LOD value, the LOD-missing variable can be viewed as left-censored data.

To investigate the effects of covariates on the LOD-missing exposure-outcome, e.g., BPA, we consider six methods. For (i) complete-case analysis, we fit a regression model using only subjects whose all data are observed. For (ii) fill-in method, we replace the missing outcome with a single constant value of  $\text{LOD}/\sqrt{2}$  for each exposure and then fit a regression model.

(iii) Tobit model [23, 39] is designed to estimate linear relationships between variables when there is either left- or right-censoring in the outcome variable, and can be specified as follows,

$$T^* = \alpha_0 + \alpha_1^T Z + \epsilon$$

where  $T^*$  and  $Z$  are the same as earlier, and  $\alpha_1$  captures the effects of  $Z$  on the underlying exposure variable  $T^*$ , not the observed outcome  $T$ . Under the assumption of a Gaussian error  $\epsilon$ , which is i.i.d. from  $N(0, \sigma^2)$ , we have underlying  $T^*$  has a normal distribution,  $T^* | Z \sim N(\alpha^T Z, \sigma^2)$ . Tobit model uses the maximum likelihood approach to estimate the parameters. We use the `tobit` function in the package `AER` [40].



(iv) Two-part model treats the observed exposure data subject to LOD as zero-inflated non-negative continuous or semi-continuous data. Such data often have two distinct features: (1) a large portion of zero values, and (2) right-skewed and heteroscedasticity for positive continuous values. In contrast to Tobit model which considers the true exposure-outcome as continuous and left-censored at the LOD value, the two-part model treats the observed data as a mixture, separating the zero and positive values explicitly by two parts. The two models are specified as follows. For some function  $h : \mathbb{R}^p \rightarrow [0, 1]$  and parameters  $\alpha_0 \in \mathbb{R}$  and  $\alpha_1 \in \mathbb{R}^p$ ,

$$\begin{aligned} \text{Part I: } P(T > 0 \mid Z) &= h(Z) \\ \text{Part II: } E(T \mid T > 0, Z) &= \alpha_0 + \alpha_1^T Z. \end{aligned} \quad (5)$$

The first part can be a logistic regression to model the probability of the outcome being positive, and the second part assumes a linear regression for the positive values. Note that fitting the conditional linear regression in the second part gives the same fitted model as complete-case analysis.

(v) Quantile regression [25] is to quantify the associations between covariates and specific quantiles of the exposure-outcome when the distribution of exposure-outcome is non-Gaussian or the research interest is on specific quantiles of the outcome. The advantages of quantile regression include its ability to handle outcome without requiring distributional assumptions, robustness to outliers, and invariance to monotone transformations. The conditional  $\tau$ -th quantile,  $Q_{T|Z}(\tau) = \inf\{t : F_{T|Z}(t) \geq \tau\}$  with  $F_{T|Z}$  being the cumulative distribution function of  $T$  given  $Z$ , is specified as

$$Q_{T|Z}(\tau) = \alpha_{\tau 0} + \alpha_{\tau}^T Z$$

where parameters  $\alpha_{\tau} \in \mathbb{R}^p$  characterize the effects of covariates on  $\tau$ th quantile of the exposure-outcome. The interpretation of estimated  $\alpha_{\tau j}$ , i.e., the  $j$ -th element of  $\alpha_{\tau}$ , when  $\tau = 0.5$  is that for a unit increase in  $Z_j$ , there is a  $\alpha_{\tau j}$  increase in the predicted median of the exposure  $T$  given other elements of  $Z$  fixed. To implement quantile regression in R, we use the `rq` function in the `quantreg` package [41].

(vi) Two-part quantile model [24] is a natural extension of quantile regression to handle zero-inflated outcome. Specifically, in addition to modeling the probability of being above the LOD, linear quantile regression is employed to model the positive part, where the quantile levels are adjusted by subject-specific zero inflation rates. To be specific, for some functions  $h : \mathbb{R}^p \rightarrow [0, 1]$  and  $\alpha : [0, 1] \rightarrow \mathbb{R}^p$ ,

$$\begin{aligned} \text{Part I: } P(T > 0 \mid Z) &= h(Z) \\ \text{Part II: } Q_{T|Z}(\tau) &= I\{\tau > 1 - h(Z)\} \cdot [\alpha_{\tau 0} + \alpha^T\{\Gamma(Z, h, \tau)\}Z] \end{aligned} \quad (6)$$

where  $\Gamma(Z, h, \tau) := \max\left\{\frac{\tau - (1 - h(Z))}{h(Z)}, 0\right\}$ . Similar to the two-part model, the first part of the model is a logistic regression model.

To implement two-part quantile model, we use functions from the author's GitHub [42]. The function `proposed.nonsmooth` fits the proposed

zero-inflated quantile model and the function `AQE` calculates Average Quantile Effect (AQE), which is the marginal effect of a covariate on the quantile of the outcome. To obtain 95% confidence intervals, we use bootstrap approach with 500 bootstrap replicates.

Note that Tobit model and quantile regression consider non-detected values as missing, while two-part model and two-part quantile model treat that as a zero-inflated outcome which clearly can be separated into two parts: zero and positive parts. Furthermore, Tobit model and two-part model are interested in the effect of a covariate on the mean of the outcome, whereas quantile regression and two-part quantile model are designed to investigate the effect of a covariate on a certain quantile of the outcome distribution.

All statistical analyses were performed using statistical software R (V4.0.0), and R codes are provided in an R Markdown document in Appendix C.

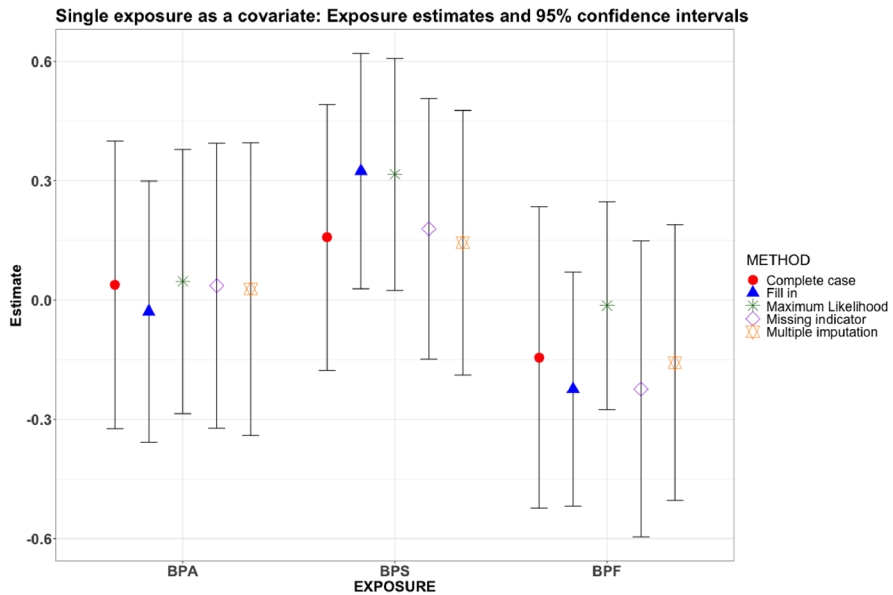
## 4 NHANES Analysis

In this section, we illustrate and compare results from different approaches in Sect. 3 using the dataset from NHANES. Before the analysis, we standardized all continuous confounders. Note that NHANES used a probability-based sampling method to represent the non-institutionalized U.S. population [43]; however, we chose not to consider sampling weights in our analysis to focus on the primary goal of comparing the methods for handling the LOD issue. By disregarding the sampling weights, our results may not estimate the true effect of exposure on a health outcome or the impact of a covariate on an exposure-outcome in the general US population.

### 4.1 Single Exposure as a Covariate

Figure 1 presents the results from the model and methods presented in Sect. 3.1, and corresponding numerical values are in Table A1 in Appendix A. When the missing rate was low, as in BPA missing around 5%, the estimated effects were similar across all five methods. When the missing rate increased to around 10% as in BPS, complete-case analysis, missing-indicator model, and multiple imputation gave similar estimates. In contrast, the fill-in method and ML method gave noticeably different results and indicated statistical significance. A potential explanation for these observations is that BPS has nonlinear effects on BMI, which lead to the effects of BPS being different above the LOD versus below the LOD. Therefore, the fill-in method and ML method that assumed a homogeneous linear effect across the entire range of exposure showed different estimates to the other methods that mainly focused on modeling the effects above the LOD.

When the missing rate was as high as 35% in BPF, coefficient estimates of BPF varied more evidently across methods. First, the complete-case analysis had the widest confidence interval which would be due to the reduction in sample size. The multiple imputation gave a similar result to that of the complete-case analysis but with a smaller standard error. The missing-indicator model gave similar estimates



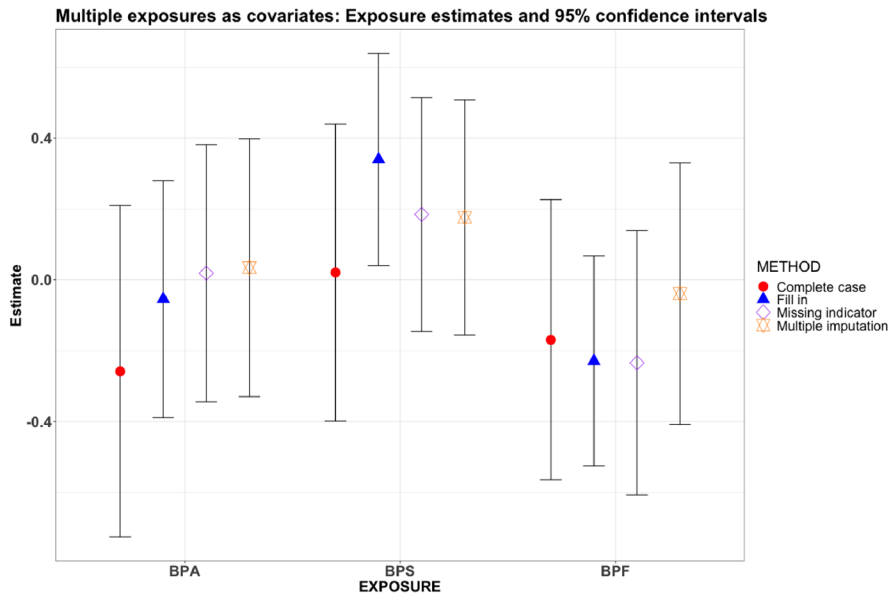
**Fig. 1** Estimated effects with 95% confidence intervals of chemical exposures in a single exposure model. x-axis indicates the exposure in the model and y-axis indicates the estimated coefficients

as the fill-in method but had a larger standard error, which was expected because its coefficient characterized the partial effect of the exposure above the LOD. Furthermore, the ML method performed differently than the other four methods, probably indicating that the Gaussian assumption for log-BPF was incorrect as indicated in Figure A2.

In summary, when a single exposure variable subject to LOD is included as a covariate in the model, we observed that the missing-indicator model and multiple imputation were robust to missing proportions and modeling assumptions. It is noteworthy that the two methods yield different interpretations. For example, the estimate of BPS of 0.1791 from the missing-indicator model means that for one SD increase in log-BPS when its value is above the LOD, there is a 0.1791 increase in the average value of BMI. On the other hand, the interpretation for the estimate of BPS of 0.1445 from the multiple imputation is: for one SD increase in log-BPS would lead 0.1445 increase in the predicted value of BMI on average. This result was closely in line with the prior study of Ortega-Villa et al. [1] that recommended missing-indicator model for practical use.

## 4.2 Multiple Exposures as Covariates

Figure 2 shows coefficient estimates of exposures with 95% their confidence intervals from the methods in Sect. 3.2 and corresponding numerical values are in Table A2 in Appendix A. For the complete-case analysis that removed 818



**Fig. 2** Estimated effects with 95% confidence intervals of chemical exposures in multiple exposures model. x-axis indicates the exposure and y-axis indicates the estimated coefficients

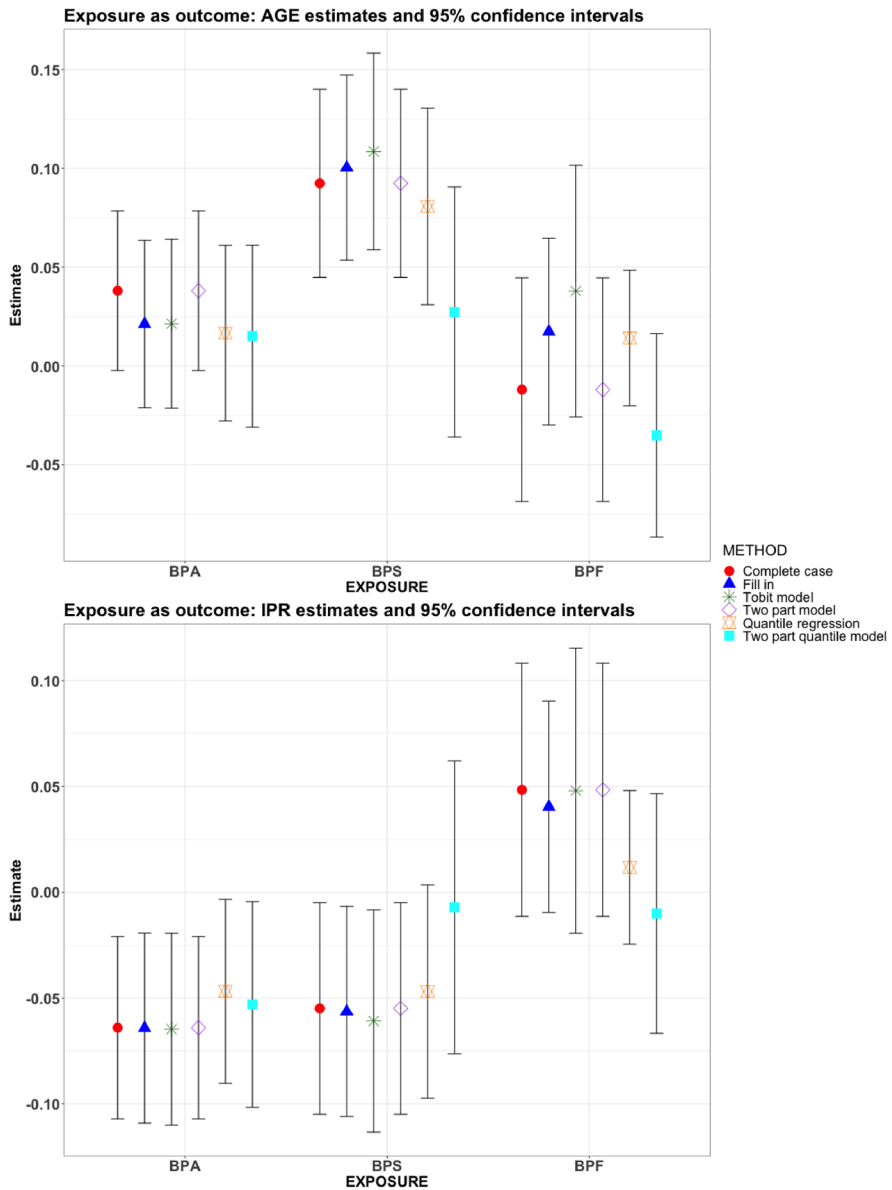
(39.98%) observations due to having at least one missing among the three exposures, estimates for BPA and BPS were noticeably smaller, and confidence intervals were wider compared to other methods. These might be due to the high missing rate as we considered multiple exposures simultaneously. In contrast, the fill-in method underestimated variability by using a single fixed value for all missing observations. It gave the smallest standard error for all three exposures among the four methods.

The missing-indicator model and multiple imputation again showed quite similar patterns of estimates. Considering their different interpretations, we recommend using the missing-indicator model if the interest is in identifying the effect of high exposure values on the outcome. On the other hand, when investigating the effect of small or entire possible values of exposure, the multiple imputation method delivers the proper interpretation.

In summary, when multiple exposures subject to LOD were modeled as covariates in a model simultaneously, the complete-case analysis could suffer a significant loss of efficiency due to the reduction of sample size. The Fill-in method might underestimate the variability by using a single value to substitute the missing values. On the other hand, the missing-indicator model and multiple imputation gave similar results when the missing rate was relatively low but would diverge as the missing rate got severe. However, they estimate different effects, hence we encourage to use of either method based on the scientific question.

### 4.3 Modeling Exposure as Outcome

Figure 3 summarizes the coefficient estimates of Age (top) and IPR (bottom) with 95% confidence intervals when modeling each of the three exposures (BPA, BPS, BPF),



**Fig. 3** Estimated effects of Age (top) and IPR (bottom) with 95% confidence intervals when the exposure is an outcome in a model. x-axis indicates the exposure and y-axis indicates the estimated value when other covariates are fixed

BPS, and BPF) as the outcome using the six methods discussed in Sect. 3.3. We present the results of  $\tau = 0.5$  which models the median of the outcome for quantile regression and the two-part quantile model. Corresponding numerical results are in Table A3 in Appendix A. Additional plots for the rest of the covariates in the model are in Figures A3–A14.

When the exposure-outcome has a low missing rate, such as BPA, estimated effects of both AGE and IPR were consistent across the six methods. In addition, standard errors were quite similar regardless of method, and their significance remained the same for all methods. Note that, results from the complete-case analysis and two-part model were exactly the same because the Part II of the two-part model (5) is an analysis without missing data.

When the missing rate of the exposure-outcome increased to around 10% as in BPS, estimated effects of AGE and IPR were mostly similar to each other except for the two-part quantile model. The interpretation of the four mean-regression methods was different from quantile regression. For example, for Tobit model, the estimated effect of IPR  $- 0.0608$  was interpreted as: for one SD increase in IPR would lead to a 0.0608 decrease in the predicted mean of log-BPS on average. On the other hand, the estimated effect of IPR  $- 0.0469$  from quantile regression meant: for one SD increase in IPR, there is a 0.0469 decrease in the predicted median of log-BPS. Estimates of Age and IPR from the two-part quantile model were different from other models and had larger standard errors.

When missing increased as high as 35% in BPF, estimated effects of AGE were variable across methods even though none was statistically significant. As the missing rate got severe, the complete-case analysis and two-part model could have lost efficiency by discarding a large number of samples. Furthermore, the fill-in method underestimated variability by using a single fixed value for 35% of the samples. On the other hand, Tobit model gave relatively consistent results as the model considered the outcome left-censored. For IPR, estimated effects from both quantile models were quite different from the other four methods, which was expected due to the skewed distribution of BPF even after log-transformation and standardization as in Figure A2. As we mentioned earlier, these approaches are interested in the effect of a covariate on the median of outcome and thus could be more robust to missing. Moreover, quantile regression had a smaller standard error than the two-part quantile model.

Besides the comparison of the estimated effects, it is noteworthy that Tobit model is a more suitable approach than the two-part model in the context when a LOD value is specified. Although we might record non-detected observations as the LOD value, they are actually some other values less than the LOD. On the other hand, for other data types, such as medical expenses or alcohol consumption, the two-part model is more appropriate because a value of zero is a true zero indicating no medical service usage or abstinence. In those cases, Tobit model is not appropriate because there is no clinically meaningful definition of the LOD.

## 5 Simulations

In this Section, we thoroughly evaluate the performance of the methods by conducting extensive simulation studies under various known data-generating settings that closely mimic real-world data. Section 5.1 examines the LOD-missing exposure variables as covariates, and we investigated two scenarios: (S1) a linear relationship between exposure and health outcome across the entire range of exposure variables and (S2) a relationship that was linear but present only above the LOD and null below the LOD. In each of these scenarios, we assessed both the single exposure model and the multiple exposures model. When modeling the LOD-missing variable as an outcome in Sect. 5.2, we considered normal and skewed distributions, respectively. For all scenarios, we generated the dataset with a sample size of 2000 and repeated the simulations 500 times. For each method, we reported the mean of coefficient estimates (EST) with standard errors (SE), and coverage probability (CP) of 95% confidence intervals.

### 5.1 Simulation 1: Covariates Models

Two exposure variables  $T_1$  and  $T_2$  were first generated from a multivariate normal distribution with shared  $\mu = 3$ ,  $\sigma^2 = 1$ , and correlation  $\rho = 0.6$ . We further transformed  $T_2 = \exp(T_2/3)$  to introduce skewness into  $T_2$ . We considered different values of LOD to achieve different proportions of data falling below the LOD around 5%, 10%, and 30%, respectively (details are listed in Table A4 in Appendix A).

#### 5.1.1 Scenario 1. Linear Relationship

Firstly, we assumed a linear relationship between exposures and the outcome across the entire range of exposure variables. The continuous health outcome was generated from either  $Y = 0.7 + 0.3T_j + \epsilon$ ,  $j = 1, 2$  for the single exposure covariate model, or  $Y = 0.7 + 0.3T_1 + 0.5T_2 + \epsilon$  for multiple covariates model with  $\epsilon \sim N(0, 1)$ .

**5.1.1.1 Single Exposure as a Covariate** We compared the five methods of modeling single exposure variable subject to LOD described in Sect. 3.1, and Table 3 presents the results. All five methods showed satisfactory results for normally distributed exposure variable  $T_1$ , especially when the missing rate was low. Particularly, the ML method outperformed as the distributional assumption for the method was correct. In addition, we observed that standard errors increased when the missing rate was higher. For the skewed exposure variable  $T_2$ , the ML method showed biases due to the violation of the normality assumption. The fill-in method showed biased estimates and decreased coverage probability particularly when the missing rate was high. Conversely, the other three methods, namely complete-case

**Table 3** Results from five methods in single exposure model when the exposure has a normal distribution (top) or a skewed distribution (bottom) under Scenario 1

Missing rate	5%		10%		30%	
Method	EST (SE)	CP	EST (SE)	CP	EST (SE)	CP
$T_1$						
Complete-case analysis	0.2995 (0.0257)	0.970	0.2997 (0.0278)	0.972	0.2997 (0.0386)	0.932
Fill-in with $\text{LOD}/\sqrt{2}$	0.2999 (0.0225)	0.962	0.2980 (0.0224)	0.960	0.2894 (0.0226)	0.954
ML approach	0.2994 (0.0225)	0.962	0.2994 (0.0226)	0.962	0.2998 (0.0235)	0.964
Missing-indicator*	0.2995 (0.0257)	0.970	0.2997 (0.0278)	0.972	0.2997 (0.0386)	0.932
Multiple imputation	0.3040 (0.0261)	0.954	0.3059 (0.0284)	0.966	0.3100 (0.0404)	0.920
$T_2$						
Complete-case analysis	0.3002 (0.0244)	0.952	0.3006 (0.0258)	0.954	0.3007 (0.0314)	0.964
Fill-in with $\text{LOD}/\sqrt{2}$	0.2921 (0.0222)	0.932	0.2865 (0.0218)	0.894	0.2751 (0.0212)	0.778
ML approach	0.2917 (0.0222)	0.932	0.2857 (0.0218)	0.888	0.2679 (0.0208)	0.650
Missing-indicator <sup>a</sup>	0.3002 (0.0244)	0.952	0.3006 (0.0258)	0.954	0.3007 (0.0314)	0.964
Multiple imputation	0.3023 (0.0246)	0.954	0.3037 (0.0260)	0.958	0.3063 (0.0320)	0.948

<sup>a</sup> Effect of a chemical exposure when the value is above the LOD value

analysis, missing-indicator model, and multiple imputations, were robust with respect to the distribution of exposure variables and provided robust results in both coefficient estimates and coverage probability.

**5.1.1.2 Multiple Exposures as Covariates** Table 4 presents the results of the four methods for modeling multiple exposures. The complete-case analysis showed increased standard error as the missing rate increases due to the reduced sample size and the correlation between exposures by considering two exposure variables simultaneously in a model. The fill-in method showed biased estimates for  $T_2$  even if only one exposure has a low missing rate of 5%, and its coverage probability was as low as 0.53. The multiple imputation method also showed large bias even when both missing rates are low, which might be attributed to the correlation between the two exposure variables as well as the violation of the normality assumption for  $T_2$ . On the other hand, the missing-indicator model gave quite consistent estimated values even if the missing rate increased, and its coverage probability remained around 0.95.

### 5.1.2 Scenario 2. No Linear Relationship Below the LOD

In practice, exposures may only exhibit effects across a certain threshold, and Scenario 2 mimics such a setting where the effect of exposures on the outcome only presents above the LOD and is null below the LOD. Therefore, the continuous health outcome was generated from the following linear models:

$$Y = 0.7 + 0.3(T_j - \text{LOD}_j)I(T_j > \text{LOD}_j) + \epsilon$$



**Table 4** Results from four methods in multiple exposures model under Scenario 1

Method	Missing rate		$T_1$		$T_2$	
	$T_1$ (%)	$T_2$ (%)	EST (SE)	CP	EST (SE)	CP
Complete-case analysis	5	5	0.2997 (0.0313)	0.956	0.5004 (0.0294)	0.946
		10	0.2998 (0.0319)	0.954	0.5010 (0.0305)	0.948
		30	0.2998 (0.0352)	0.950	0.5009 (0.0355)	0.952
	10	5	0.2996 (0.0332)	0.962	0.5004 (0.0297)	0.946
		10	0.2997 (0.0337)	0.966	0.5010 (0.0307)	0.948
		30	0.2994 (0.0367)	0.948	0.5009 (0.0357)	0.952
	30	5	0.2994 (0.0439)	0.948	0.5005 (0.0318)	0.946
		10	0.2995 (0.0442)	0.942	0.5009 (0.0326)	0.954
		30	0.2996 (0.0464)	0.942	0.5010 (0.0371)	0.956
Fill-in with $\text{LOD}/\sqrt{2}$	5	5	0.2984 (0.0278)	0.944	0.4891 (0.0274)	0.918
		10	0.3001 (0.0278)	0.946	0.4786 (0.0269)	0.868
		30	0.3161 (0.0275)	0.912	0.4504 (0.0259)	0.534
	10	5	0.2955 (0.0277)	0.940	0.4980 (0.0274)	0.926
		10	0.2971 (0.0277)	0.942	0.4803 (0.0269)	0.878
		30	0.3127 (0.0274)	0.924	0.4520 (0.0259)	0.554
	30	5	0.2799 (0.0276)	0.892	0.5042 (0.0271)	0.938
		10	0.2812 (0.0276)	0.898	0.4935 (0.0267)	0.936
		30	0.2942 (0.0275)	0.950	0.4646 (0.0259)	0.720
Missing-indicator <sup>a</sup>	5	5	0.2989 (0.0308)	0.950	0.5002 (0.0291)	0.942
		10	0.2997 (0.0308)	0.954	0.4997 (0.0301)	0.956
		30	0.3101 (0.0308)	0.952	0.4945 (0.0347)	0.956
	10	5	0.2982 (0.0328)	0.960	0.5007 (0.0292)	0.942
		10	0.2986 (0.0328)	0.962	0.5000 (0.0302)	0.950
		30	0.3065 (0.0328)	0.964	0.4951 (0.0348)	0.962
	30	5	0.2922 (0.0431)	0.950	0.5084 (0.0293)	0.932
		10	0.2924 (0.0432)	0.948	0.5054 (0.0304)	0.952
		30	0.2940 (0.0435)	0.948	0.4968 (0.0354)	0.956
Multiple imputation	5	5	0.3278 (0.0314)	0.850	0.5211 (0.0297)	0.890
		10	0.3443 (0.0314)	0.700	0.5226 (0.0310)	0.878
		30	0.3960 (0.0313)	0.138	0.5300 (0.0367)	0.858
	10	5	0.3284 (0.0338)	0.866	0.5322 (0.0298)	0.834
		10	0.3438 (0.0341)	0.744	0.5338 (0.0313)	0.794
		30	0.3962 (0.0342)	0.188	0.5398 (0.0372)	0.812
	30	5	0.3305 (0.0467)	0.876	0.5737 (0.0301)	0.284
		10	0.3421 (0.0475)	0.800	0.5757 (0.0317)	0.316
		30	0.3908 (0.0486)	0.544	0.5804 (0.0384)	0.462

<sup>a</sup> Effect of a chemical exposure when the value is above the LOD value

**Table 5** Results from four methods in multiple exposures model under Scenario 2

Method	Missing rate		$T_1$		$T_2$	
	$T_1(\%)$	$T_2(\%)$	EST (SE)	CP	EST (SE)	CP
Complete-case analysis	5	5	0.2997 (0.0313)	0.956	0.5004 (0.0294)	0.946
		10	0.2998 (0.0319)	0.954	0.5010 (0.0305)	0.948
		30	0.2998 (0.0352)	0.950	0.5009 (0.0355)	0.952
	10	5	0.2996 (0.0332)	0.962	0.5004 (0.0297)	0.946
		10	0.2997 (0.0337)	0.956	0.5010 (0.0307)	0.948
		30	0.2994 (0.0367)	0.948	0.5009 (0.0357)	0.952
	30	5	0.2994 (0.0439)	0.948	0.5005 (0.0318)	0.946
		10	0.2995 (0.0442)	0.942	0.5009 (0.0326)	0.954
		30	0.2996 (0.0464)	0.942	0.5010 (0.0371)	0.956
Fill-in with $\text{LOD}/\sqrt{2}$	5	5	0.2814 (0.0278)	0.902	0.4813 (0.0274)	0.888
		10	0.2797 (0.0278)	0.892	0.4620 (0.0270)	0.706
		30	0.2760 (0.0275)	0.880	0.3983 (0.0260)	0.026
	10	5	0.2687 (0.0277)	0.812	0.4822 (0.0274)	0.894
		10	0.2671 (0.0277)	0.790	0.4628 (0.0270)	0.722
		30	0.2630 (0.0275)	0.738	0.3995 (0.0260)	0.032
	30	5	0.2135 (0.0276)	0.108	0.4832 (0.0271)	0.898
		10	0.2132 (0.0276)	0.112	0.4629 (0.0267)	0.710
		30	0.2091 (0.0275)	0.084	0.3996 (0.0259)	0.024
Missing-indicator <sup>a</sup>	5	5	0.2994 (0.0308)	0.952	0.5005 (0.0291)	0.942
		1	0.2994 (0.0308)	0.952	0.5009 (0.0301)	0.954
		30	0.2993 (0.0307)	0.948	0.5011 (0.0345)	0.960
	10	5	0.2995 (0.0327)	0.960	0.5005 (0.0292)	0.942
		10	0.2995 (0.0327)	0.960	0.5009 (0.0302)	0.948
		30	0.2995 (0.0327)	0.964	0.5011 (0.0347)	0.956
	30	5	0.2995 (0.0430)	0.950	0.5004 (0.0292)	0.944
		10	0.2994 (0.0430)	0.952	0.5008 (0.0303)	0.948
		30	0.2994 (0.0431)	0.950	0.5010 (0.0351)	0.954
Multiple imputation	5	5	0.3203 (0.0310)	0.920	0.5153 (0.0294)	0.908
		10	0.3299 (0.0309)	0.846	0.5161 (0.0306)	0.922
		30	0.3460 (0.0303)	0.668	0.5171 (0.0355)	0.904
	10	5	0.3201 (0.0332)	0.922	0.5213 (0.0293)	0.892
		10	0.3293 (0.0333)	0.860	0.5219 (0.0307)	0.880
		30	0.3465 (0.0327)	0.704	0.5215 (0.0357)	0.894
	30	5	0.3197 (0.0450)	0.904	0.5314 (0.0292)	0.810
		10	0.3250 (0.0448)	0.890	0.5334 (0.0305)	0.812
		30	0.3416 (0.0446)	0.810	0.5301 (0.0360)	0.860

<sup>a</sup> Effect of a chemical exposure when the value is above the LOD value

for  $j = 1, 2$  for a single exposure model and

$$Y = 0.7 + 0.3(T_1 - \text{LOD}_1)I(T_1 > \text{LOD}_1) + 0.5(T_2 - \text{LOD}_2)I(T_2 > \text{LOD}_2) + \epsilon$$

for multiple exposures model, where  $\epsilon \sim N(0, 1)$ . The results are shown in Table A5 for single exposure model and Table 5 for multiple exposures model.

**5.1.2.1 Single Exposure as a Covariate** From Table A5, we observed that regardless of the distribution of the exposures, the complete-case analysis and the missing-indicator model exhibited superior performance compared to other methods, with desired coverage probability close to 0.95. This can be attributed to the fact that the linear relationship between exposure and outcome is present only above the LOD, and these two methods disregard observations below the LOD. In contrast, the fill-in method and ML method showed attenuated coefficients even with a low missing rate of 5%. When the missing rate became more severe, biases for both methods increased and their coverage probability decreased dramatically even for normally distributed  $T_1$ .

**5.1.2.2 Multiple Exposures as Covariates** Similar to the results observed in single exposure models, the complete-case analysis and the missing-indicator model outperformed other methods in the multiple exposures model as Table 5 presents. The fill-in method exhibited biased estimates when the missing rate for both exposures was as low as 5%, and the bias became more severe even when only one of the missing rates increased. Regarding coverage probability, the values deviated significantly from the desired value of 0.95 in most cases and reached less than 0.10 for both exposures when they are missing around 30%. A result similar to that observed with a single exposure, but more severe, may have arisen from a correlation between the two exposure variables since the model includes both exposure variables at the same time.

Multiple imputation also displayed biases for most of the combinations of the missing rates, which may arise from the distribution and correlation between two exposure variables. Again, note that this approach estimates the effect of exposure on the outcome across the entire range of the exposure. As a result, the implications of multiple imputation differ from those of the complete-case analysis and the missing-indicator model.

## 5.2 Simulation 2. Outcome Models

In this Section, we assess the performance of the six methods of handling the LOD-missing variable as an outcome from Sect. 3.3. We considered two scenarios where the variable followed a normal distribution and a skewed distribution, respectively. The exposure-outcomes were generated as follows:  $T_j^* = 0.7 + 0.3Z_1 + 0.5Z_2 + \epsilon_j$ ,  $j = 1, 2$ , where two covariates  $Z_1 \sim N(0, 0.25^2)$  and  $Z_2 \sim N(0, 0.5^2)$ . The random error  $\epsilon_1 \sim N(0, 2^2)$  generated the normally distributed outcome variable  $T_1^*$ , and  $\epsilon_2 \sim \chi^2(3)$  generated the skewed outcome variable  $T_2^*$ . Similar to covariates models, we used the same three approximate missing rates around 5%, 10%, and 30% due to LOD. Note that we used  $\tau = 0.5$  for quantile regression and two-part quantile model to model the conditional median of the outcome.

**Table 6** Results from six methods in outcome model when the exposure-outcome is normally distributed

Missing rate		5%		10%		30%	
Method		EST (SE)	CP	EST (SE)	CP	EST (SE)	CP
$Z_1$	Complete-case analysis	0.2459(0.1653)	0.950	0.2114 (0.1590)	0.922	0.1481 (0.1509)	0.836
	Fill-in with $\text{LOD}/\sqrt{2}$	0.2763 (0.1657)	0.952	0.2562 (0.1561)	0.958	0.2060 (0.1309)	0.886
	Tobit model	0.3041 (0.1794)	0.954	0.3046 (0.1804)	0.942	0.3053 (0.1869)	0.954
	Two-part model	0.2459 (0.1653)	0.950	0.2114 (0.1590)	0.922	0.1481 (0.1509)	0.836
	Quantile regression <sup>a</sup>	0.3094 (0.2237)	0.946	0.3094 (0.2237)	0.946	0.3058 (0.2201)	0.942
	Two-part quantile model <sup>a,b</sup>	0.2611 (0.1914)	–	0.1891 (0.1908)	–	0.1636 (0.1998)	–
$Z_2$	Complete-case analysis	0.4043 (0.0828)	0.774	0.3513 (0.0797)	0.518	0.2484 (0.0757)	0.084
	Fill-in with $\text{LOD}/\sqrt{2}$	0.4549 (0.0829)	0.910	0.4222 (0.0781)	0.816	0.3382 (0.0655)	0.294
	Tobit model	0.5007 (0.0898)	0.936	0.5003 (0.0903)	0.940	0.4990 (0.0937)	0.940
	Two-part model	0.4043 (0.0828)	0.774	0.3513 (0.0797)	0.518	0.2484 (0.0757)	0.084
	Quantile regression <sup>a</sup>	0.4993 (0.1120)	0.940	0.4993 (0.1120)	0.940	0.4936 (0.1100)	0.942
	Two-part quantile model <sup>a,b</sup>	0.4365 (0.0971)	–	0.3127 (0.0969)	–	0.2801 (0.1016)	–

<sup>a</sup> Effect of a variable on median of the outcome while others give those on mean of the outcome

<sup>b</sup> Standard error of estimates from 500 bootstrap replicates due to time limit

### 5.2.1 Case 1: Normally Distributed Outcome $T_1$

Table 6 presents the results from the six methods described in Sect. 3.3. For all missing rates, Tobit model and quantile regression method outperformed others in terms of coefficient estimates and coverage probability. Even when the missing rate was around 30%, the coverage probability of the Tobit model remained around the desired level of 0.95 for both covariates. In contrast, the complete-case analysis and filled-in method showed biased estimates and poor coverage probability for  $Z_2$ . Furthermore, they got severe for both covariates as the missing rate increased. The two-part model and two-part quantile model showed attenuated estimates. As noted in Sect. 3.3, the two-part models estimate the effects of the covariates on the outcome conditional on the outcome being above the LOD. Thus, estimated coefficients are not directly comparable with true values of 0.3 and 0.5.

### 5.2.2 Case 2: Skewed Outcome $T_2$

Table 7 shows the results for the skewed exposure-outcome  $T_2$ . In this case, we observed that the fill-in method surprisingly performed well up to the missing rate around 10%. When further examining the distribution of the outcome by filling the missing value with the single fixed value of  $\text{LOD}/\sqrt{2}$ , we found it had little impact on the original skewed distribution (see Figure A15). In contrast, Tobit model displayed biases and they got severe as the missing rate increased. Its poor performance

**Table 7** Results from six methods in outcome model when the exposure-outcome is skewed

Missing rate	5%		10%		30%	
Method	EST (SE)	CP	EST (SE)	CP	EST (SE)	CP
$Z_1$						
Complete-case analysis	0.1358 (0.2226)	0.894	0.1046 (0.2279)	0.848	0.0293 (0.2501)	0.812
Fill-in with LOD/ $\sqrt{2}$	0.3058 (0.2191)	0.950	0.2925 (0.2197)	0.954	0.2446 (0.2179)	0.944
Tobit model	0.3328 (0.2274)	0.944	0.3718 (0.2383)	0.938	0.3727 (0.2777)	0.940
Two-part model	0.1358 (0.2226)	0.894	0.1046 (0.2279)	0.848	0.0293 (0.2501)	0.812
Quantile regression <sup>a</sup>	0.2974 (0.2382)	0.946	0.2945 (0.2380)	0.938	0.2878 (0.2344)	0.954
Two-part quantile model <sup>a,b</sup>	0.3644 (0.2469)	–	0.4326 (0.2498)	–	0.3538 (0.2226)	–
$Z_2$						
Complete-case analysis	0.2635 (0.1123)	0.442	0.1667 (0.1146)	0.184	0.0772 (0.1251)	0.074
Fill-in with LOD/ $\sqrt{2}$	0.4977 (0.1096)	0.958	0.4831 (0.1099)	0.944	0.4187 (0.1088)	0.898
Tobit model	0.5872 (0.1140)	0.898	0.6154 (0.1195)	0.824	0.6290 (0.1390)	0.862
Two-part model	0.2635 (0.1123)	0.442	0.1667 (0.1146)	0.184	0.0772 (0.1251)	0.074
Quantile regression <sup>a</sup>	0.4993 (0.1191)	0.948	0.4946 (0.1190)	0.956	0.4845 (0.1170)	0.958
Two-part quantile model <sup>a,b</sup>	0.5781 (0.1236)	–	0.6786 (0.1243)	–	0.6657 (0.1213)	–

<sup>a</sup> Effect of a variable on median of the outcome while others give those on mean of the outcome

<sup>b</sup> Standard error of estimates from 500 bootstrap replicates due to time limit

may be due to the normal assumption violation. Furthermore, the complete-case analysis and two-part model showed dramatically biased results. As for the coverage probability, it reached the value less than 0.10 for  $Z_2$  when the missing rate is as high as 30%.

Quantile regression performed well and produced reliable estimates close to the true value of 0.3 and 0.5 for  $Z_1$  and  $Z_2$ , respectively, even when the missing rate increased up to 30%. However, note that quantile regression estimates the effect of covariates on the median of the outcome. The two-part quantile model showed biased effects of covariates on the outcome compared to quantile regression. This may be attributed to the nature of the two-part quantile model, which estimates the marginal effect of the covariates on the outcome adjusted for the conditional outcome being above the LOD.

## 6 Conclusions

In this paper, we investigate and compare various methods to model exposure variables subject to LOD by leveraging real-world data from the NHANES dataset and a series of simulation studies. We consider both cases of exposure variables as covariates and an outcome in a model. For the single exposure model, we consider complete-case analysis, fill-in method, ML approach, missing-indicator model, and multiple imputation. For the multiple exposures model with multiple exposures as covariates simultaneously in a model, we consider complete-case

analysis, fill-in method, missing-indicator model, and multiple imputation. Treating an exposure variable as an outcome in a model, we compare complete-case analysis, fill-in method, Tobit model, two-part model, quantile regression, and two-part quantile model. From the comparison of various methods, we provide practical valuable insights in using the statistical models to address the LOD-missing data in environmental health research projects.

In summary, our numerical investigations demonstrate that method performance varies contingent upon distinct settings and modeling presumptions related to LOD-influenced exposure variables. When the missing rate is low, all methods perform similarly. As the missing rate increases, the fill-in method and ML method could be sensitive to modeling assumptions of linearity and Gaussian distribution. The complete-case analysis could suffer a significant loss of power. The missing-indicator model generally delivers robust results, and the multiple imputation method also performs well in single exposure models or when multiple exposures are not highly correlated. But note that they bear different interpretations, and thus the choice of methods should be guided by scientific questions. When an exposure variable is modeled as an outcome, Tobit model provides a suitable interpretation given a specific LOD value. Furthermore, if the exposure-outcome distribution is skewed or the scientific interest is in tail quantiles rather than in the center of the outcome distribution, quantile regression or two-part quantile regression would be useful.

Generally, we suggest paying attention to the distribution of exposure variables and the trend of exposure effects in addition to the missing proportion due to LOD. If the exposure variable deviates from the normal distribution or the assumed parametric distribution, the ML approach is incorrect and would provide biased results. On the other hand, the missing-indicator model is least affected by modeling assumptions and shows better stability. Also, estimated effects from mean regression and median regression would diverge as the normal distribution assumption does not meet.

In conclusion, the methods exemplified in this paper are a collection that can be useful for researchers to: 1) Assess the effects of one or more chemical exposures on a health outcome; 2) Evaluate the effect of a covariate on a chemical exposure; 3) Address the relationship between exposures and other variables in various situations such as exposure measurements are handled as a mixture or left-censored data, or one's interest lies in modeling a quantile of the outcome rather than the mean of the outcome.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s12561-023-09408-3>.

**Funding** This research was partially supported by NIH R01ES032808, NIH R01ES032826, and NIH UH3OD023305, and CDC/NIOSH grant U01OH012637.

## Declarations

**Conflict of interest** The authors declare that they have no conflict of interest.


## References

- Ortega-Villa AM, Liu D, Ward MH, Albert PS (2021) New insights into modeling exposure measurements below the limit of detection. *Environ Epidemiol* 5(1):e116
- He H, Mi X, Tang W, Kelly T, Shen H, Deng H, Du Y (2020) Statistical issues on analysis of censored data due to detection limit. *Int J Stat Probab* 9(4):49–61
- Lin DY, Zeng D, Couper D (2020) A general framework for integrative analysis of incomplete multiomics data. *Genet Epidemiol* 44(7):646–664
- Lynn HS (2001) Maximum likelihood inference for left-censored HIV RNA data. *Stat Med* 20(1):33–45
- May RC, Ibrahim JG, Chu H (2011) Maximum likelihood estimation in generalized linear models with multiple covariates subject to detection limits. *Stat Med* 30(20):2551–2561
- Chiou SH, Betensky RA, Balasubramanian R (2019) The missing indicator approach for censored covariates subject to limit of detection in logistic regression models. *Ann Epidemiol* 38:57–64
- Nie L, Chu H, Liu C, Cole SR, Vexler A, Schisterman EF (2010) Linear regression with an independent variable subject to a detection limit. *Epidemiology* 21(suppl 4):S17
- Cole SR, Chu H, Nie L, Schisterman EF (2009) Estimating the odds ratio when exposure has a limit of detection. *Int J Epidemiol* 38(6):1674–1680
- Lubin JH, Colt JS, Camann D, Davis S, Cerhan JR, Severson RK, Hartge P (2004) Epidemiologic evaluation of measurement data in the presence of detection limits. *Environ Health Perspect* 112(17):1691–1696
- Neelon B, O'Malley AJ, Smith VA (2016) Modeling zero-modified count and semicontinuous data in health services research part 1: background and overview. *Stat Med* 35(27):5070–5093
- Liu L, Shih YCT, Strawderman RL, Zhang D, Johnson BA, Chai H (2019) Statistical analysis of zero-inflated nonnegative continuous data: a review. *Stat Sci* 34(2):253–279
- Bernhardt PW, Wang HJ, Zhang D (2015) Statistical methods for generalized linear models with covariates subject to detection limits. *Stat Biosci* 7:68–89
- Dziura JD, Post LA, Zhao Q, Fu Z, Peduzzi P (2013) Strategies for dealing with missing data in clinical trials: from design to analysis. *Yale J Biol Med* 86(3):343
- Helsel DR (2005) More than obvious: better methods for interpreting nondetect data. *Environ Sci Technol* 39(20):419A–423A
- Richardson DB, Ciampi A (2003) Effects of exposure measurement error when an exposure variable is constrained by a lower limit. *Am J Epidemiol* 157(4):355–363
- Schisterman EF, Vexler A, Whitcomb BW, Liu A (2006) The limitations due to exposure detection limits for regression models. *Am J Epidemiol* 163(4):374–383
- Rubin DB (1987) Multiple imputation for nonresponse in surveys. Wiley, New York
- Baccarelli A, Pfeiffer R, Consonni D, Pesatori AC, Bonzini M, Patterson DG Jr, Landi MT (2005) Handling of dioxin measurement data in the presence of non-detectable values: overview of available methods and their application in the seveso chloracne study. *Chemosphere* 60(7):898–906
- Arunajadai SG, Rauh VA (2012) Handling covariates subject to limits of detection in regression. *Environ Ecol Stat* 19(3):369–391
- Liu H, Campana AM, Wang Y, Kannan K, Liu M, Zhu H, Ghassabian A (2021) Organophosphate pesticide exposure: demographic and dietary predictors in an urban pregnancy cohort. *Environ Pollut* 283:116920
- Tyrrell J, Melzer D, Henley W, Galloway TS, Osborne NJ (2013) Associations between socioeconomic status and environmental toxicant concentrations in adults in the USA: NHANES 2001–2010. *Environ Int* 59:328–335
- Vrijheid M, Martinez D, Aguilera I, Ballester F, Basterrechea M, Esplugues A, Sunyer J (2012) Socioeconomic status and exposure to multiple environmental pollutants during pregnancy: evidence for environmental inequity? *J Epidemiol Commun Health* 66(2):106–113
- Tobin J (1958) Estimation of relationships for limited dependent variables. *Econometrica* 26:24–36
- Ling W, Cheng B, Wei Y, Willey JZ, Cheung YK (2022) Statistical inference in quantile regression for zero-inflated outcomes. *Stat Sin* 32:1411–1433
- Koenker R, Bassett G Jr (1978) Regression quantiles. *Econometrica* 46:33–50
- Centers for Disease Control and Prevention (CDC) and National Center for Health Statistics (NCHS) (2014) National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department

- of Health and Human Services, Centers for Disease Control and Prevention, 2013–2014. Available at: <https://www.cdc.gov/nchs/nhanes/continuousnhanes/overview.aspx?BeginYear=2013>
27. Trasande L, Attina TM, Blustein J (2012) Association between urinary bisphenol a concentration and obesity prevalence in children and adolescents. *J Am Med Assoc* 308(11):1113–1121
  28. Carwile JL, Michels KB (2011) Urinary bisphenol a and obesity: Nhanes 2003–2006. *Environ Res* 111(6):825–830
  29. Food and Drug Administration (FDA) (2014) Bisphenol A (BPA): Use in food contact application. Available at: <https://www.fda.gov/newsevents/publichealthfocus/ucm064437.htm>
  30. Rochester JR, Bolden AL (2015) Bisphenol s and f: a systematic review and comparison of the hormonal activity of bisphenol a substitutes. *Environ Health Perspect* 123(7):643–650
  31. Eladak S, Grisin T, Moison D, Guerquin MJ, N'Tumba-Byn T, Pozzi-Gaudin S, Habert R (2015) A new chapter in the bisphenol a story: bisphenol s and bisphenol f are not safe alternatives to this compound. *Fertil Steril* 103(1):11–21
  32. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Carpenter JR (2009) Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ*. <https://doi.org/10.1136/bmj.b2393>
  33. White IR, Royston P, Wood AM (2011) Multiple imputation using chained equations: issues and guidance for practice. *Stat Med* 30(4):377–399
  34. Van Buuren S, Groothuis-Oudshoorn K (2011) Mice: multivariate imputation by chained equations in r. *J Stat Softw* 45:1–67
  35. Schafer JL (1997) Analysis of incomplete multivariate data. CRC Press, Boca Raton
  36. Van Buuren S, Oudshoorn CG (2000) Multivariate imputation by chained equations. TNO, Leiden
  37. Van Buuren S, Brand JP, Groothuis-Oudshoorn CG, Rubin DB (2006) Fully conditional specification in multivariate imputation. *J Stat Comput Simul* 76(12):1049–1064
  38. Jakobsen JC, Gluud C, Wetterslev J, Winkel P (2017) When and how should multiple imputation be used for handling missing data in randomised clinical trials—a practical guide with flowcharts. *BMC Med Res Methodol* 17(1):1–10
  39. Amemiya T (1984) Tobit models: a survey. *J Econometr* 24(1–2):3–61
  40. Kleiber C, Zeileis A, Zeileis MA (2020) Package 'aer'. R package version 12(4)
  41. Koenker R, Portnoy S, Ng PT, Zeileis A, Grosjean P, Ripley BD (2018) Package 'quantreg'. Cran R-project. org
  42. Ling W (2022) Statistical inference in quantile regression for zero-inflated outcomes. <https://github.com/wdl2459/ZIQ>. GitHub
  44. Zipf G, Chiappa M, Porter KS, Ostchega Y, Lewis BG, Dostal J (2013) Health and nutrition examination survey plan and operations, 1999–2010. National Center for Health Statistics. *Vital Health Stat* 1(56)

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

## Authors and Affiliations

Eunsil Seok<sup>1</sup> · Akhgar Ghassabian<sup>1,2</sup> · Yuyan Wang<sup>1</sup> · Mengling Liu<sup>1</sup> 

✉ Mengling Liu  
mengling.liu@nyulangone.org

<sup>1</sup> Department of Population Health, New York University Grossman School of Medicine, New York, NY 10016, USA

<sup>2</sup> Department of Pediatrics, New York University Grossman School of Medicine, New York, NY 10016, USA