International Journal of Statistics and Probability; Vol. 9, No. 4; July 2020
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Statistical Issues on Analysis of Censored Data Due to Detection Limit

Hua He[1], Xuenan Mi[1] Jerry Cornell[2], Wan Tang[2], Tanika Kelly[1], Hui Shen[2], Hongwen Deng[2] & Yan Du[2]

[1] Department of Epidemiology, USA

[2] Department of Biostatistics and Data Science, Tulane University School of Public Health and Tropical Medicine, New Orleans, LA, USA

Correspondence: Hua He, Department of Epidemiology, USA

Received: April 22, 2020 Accepted: May 20, 2020 Online Published: June 14, 2020

doi:10.5539/ijsp.v9n4p49 URL: https://doi.org/10.5539/ijsp.v9n4p49

Abstract

Measures of substance concentration in urine, serum or other biological matrices often have an assay limit of detection. When concentration levels fall below the limit, the exact measures cannot be obtained, and thus are left censored. Common practice for addressing the censoring issue is to delete or 'fill-in' the censored observations in data analysis, which often results in biased or non-efficient estimates. Assuming the concentration or transformed concentration follows a normal distribution, a Tobit regression model can be applied. When the study population is heterogeneous, for example due to the existence of a latent group of subjects who lack the substance, the problem becomes more challenging. In this paper, we conduct intensive simulation studies to investigate the statistical issues in analyzing censored data and compare different methods in which the data are treated either as a dependent variable or an independent variable. We also analyze triclosan data in the NHANES study and metabolites data in the Bogalusa Heart Study to illustrate the issues. Some guidelines for analyzing such censored data are provided.

Keywords: censored data, censored normal distribution, detection limit, heterogeneous population, joint modeling, mixture model, relative efficiency, Tobit regression model

1. Introduction

Measures of substance concentration in urine, serum or other biological matrices that fall below assay limit of detection are common in epidemiological and medical research (Nassan et al., 2017; Ferrero et al., 2017; Østergren et al., 2017; Kim et al., 2018; Zhao et al., 2018; Gomez et al., 2019; Maule et al., 2019). When the concentration levels are under the detection limit (DL), denoted as L, accurate measures cannot be obtained. Instead, their values are only partially known and left censored. For example, triclosan, collected in NHANES studies, is a broad-spectrum antimicrobial chemical and widely used in household and health care related products. Currently the DL for urine triclosan concentration is 2.3 ng/ml, meaning only triclosan concentrations greater than or equal to 2.3 ng/ml can be detected. For triclosan concentrations lower than 2.3 ng/ml, their values are censored. Instead of precise measures of the triclosan concentration levels, the values are only partially known, namely somewhere between 0 and 2.3. The censoring generates informative missing data that need to be properly handled in order to obtain valid inference and achieve efficiency.

A simple approach for handling such censored data is complete-case analysis (Little and Rubin, 2002). This approach discards all subjects who have censored data and fits regression models by using only subjects whose data are uncensored. This method is often invalid because it violates the missing completely at random assumption and is potentially highly inefficient even with moderate censoring.

Another common approach is the 'fill-in' method where the censored observations are replaced by a constant such as L, or $\frac{1}{2}L$ (Olson, 1993; LaFleur et al., 2011; Slymen et al., 1994; Gleit, 1985; Newman et al., 1989). This approach is widely applied in epidemiological and medical research because of its simplicity of implementation, but often leads to biased estimates.

When the underlying measures follow a single normal distribution, the data then follow a censored normal distribution due to DL. Tobit regression models can be applied to model such censored dependent variables (Tobin, 1958). When the underlying measures are not from a normal distribution, data transformation such as log-transformation can be employed first, and a Tobit regression model can then be applied on the transformed data. The Tobit regression model has been widely applied in some fields such as economics (Rosen, 1976; Keeley et al., 1978; McDonald and Moffitt, 1980; Amemiya, 1984; Zhou et al., 2018; Al-Hanawi et al., 2018; Deng et al., 2019; Leng et al., 2019; Hezaveh and Cherry, 2019); however, its application is not well-recognized in epidemiological and medical research despite the universality of

censored data due to DL.

We run into issues when censored observations are from heterogeneous populations, such as when exposure and non-exposure become factors. Subjects from the non-exposure population are not exposed at all, so their concentration level is 0 and thus censored, while subjects from the exposure population will have a concentration level greater than 0. If their levels are lower than L, they are censored as well. Only those subjects who are exposed and also have concentration levels greater than or equal to L can be observed. In the above triclosan example, subjects having triclosan in their urine, namely the exposed group, but with concentration levels lower than L are censored. However, when there is a subgroup of subjects who do not have triclosan at all, i.e, are not exposed to triclosan, their measures are of course under L and thus are censored as well. Because measures from the non-exposed group are always censored, they are considered latent. If the data from the exposure population follow a censored normal distribution, data from the whole sample follow a mixture of censored normal and degenerate distributions. In this case, if a Tobit model is applied, biased estimates can occur. To model data from such heterogeneous populations, mixture model techniques should be applied to achieve unbiased estimates(Moulton and Halsey, 1995; Taylor et al., 2001; Reisetter et al., 2017).

When censored variables are used as predictors, common practices such as deletion or 'fill-in' can be problematic (Zhao et al., 2018; Park et al., 2018; Javad et al., 2018) . To better facilitate information provided by the censored observations in predictors, a joint modeling approach is proposed to aid estimation. The joint modeling has been applied in Lynn (2001); Rigobon and Stoker (2003); Austin and Hoch (2004); Bernhardt et al. (2015), but its application in epidemiological and medical research is very rare. Furthermore, no method is available when censored predictors are from heterogeneous populations.

In this paper, we propose a mixture model for outcomes obtained from heterogeneous populations, and develop joint modeling for censored predictor, either from a single or heterogeneous populations. We conduct intensive simulation studies to investigate statistical issues for different methods and also use two real data examples to illustrate the methods.

2. Methods

2.1 As Dependent Variable

When censored data are used as outcome, Tobit models can be applied. Next, we give a brief review of the Tobit model for the outcome from a single population. Because the Tobit model is inappropriate for modeling outcomes from heterogeneous populations, a mixture model is also proposed for those cases (Moulton and Halsey, 1995; Taylor et al., 2001; Reisetter et al., 2017).

2.1.1 Tobit Model for Single Population

The Tobit model is a linear regression model for data from censored normal distribution. Consider an independent sample $(\mathbf{x}_i, z_i)$, $i = 1, ..., n$, where n is the sample size, $\mathbf{x}_i = (x_{i1}, x_{i2,}..., x_{ip})$ is a p-dimensional covariate, and z_i is the outcome for the i^{th} subject. The z_i is obtained based on an underlying outcome variable z_i^*, which is assumed to have a linear relationship with $\mathbf{x}_i$ through

$$z_i^* = \mathbf{x}_i^\top \beta + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2). \tag{1}$$

Let L be the lower DL, the outcome z_i is defined as

$$z_i = \begin{cases} z_i^* & \text{if } z_i^* \geq L, \\ \text{censored} & \text{if } z_i^* < L. \end{cases} \tag{2}$$

Under (1), the z_i follows a censored normal distribution with likelihood given by

$$f(Z_i = z_i) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(z_i-\mu_i)^2}{2\sigma^2}\right) & \text{if } z_i \geq L, \\ \Phi\left(\frac{L-\mu_i}{\sigma}\right) & \text{if } z_i \text{ censored}, \end{cases} \tag{3}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal and $\mu_i = \mathbf{x}_i^\top \beta$ is the conditional mean of the outcome. The outcome z_i can be modeled by a Tobit regression model, denoted as Tobit(μ_i, σ^2, L), given by

$$z_i \mid \mathbf{x}_i \sim \text{i.d. Tobit}\left(\mu_i, \sigma^2, L\right), \quad \mu_i = \mathbf{x}_i^T \beta. \tag{4}$$

Let c_i indicate whether z_i is censored or not with $c_i = 1$ for censored and $c_i = 0$ for $z_i \geq L$. The likelihood function

consisting censored and non-censored subjects is

$$L = \prod_{i=1}^{n} \left[\Phi\left(\frac{L-\mu_i}{\sigma} \right) \right]^{c_i} \left[\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(z_i-\mu_i)^2}{2\sigma^2} \right) \right]^{(1-c_i)}. \tag{5}$$

The parameters β and σ can be estimated based on maximum likelihood estimate (MLE) approach, i.e., maximizing the likelihood (5) for given (4).

The Tobit regression model is applied for outcomes from censored normal distribution. When the underlying measures are not normally distributed, data transformation can be applied first, and a Tobit regression model can then be applied on the transformed data.

2.1.2 Mixture Model for Heterogenous Populations

For data from heterogeneous populations, such as concerning exposure and non-exposure, Tobit models are not appropriate. Instead, models which can separate the two populations are needed.

Let ω be the prevalence of the non-exposure, so the whole population consists of $100\omega\%$ non-exposure and $100(1-\omega)\%$ exposure. If data from the exposure population follow a censored normal distribution (3), the data from the whole population have a mixture of censored normal distribution with probability $(1-\omega)$ and degenerate distributions with probability ω, given by

$$f(Z_i = z_i) = \begin{cases} (1-\omega)\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(z_i-\mu_i)^2}{2\sigma} \right) & \text{if exposure and } z_i \geq L, \\ (1-\omega)\Phi\left(\frac{L-\mu_i}{\sigma} \right) & \text{if exposure and } z_i < L, \\ \omega & \text{if non-exposure} \end{cases} \tag{6}$$

If the likelihood of non-exposure is predicted by some covariates, say $\mathbf{u}_i$, and modeled by a logit model, then the exposure outcome is modeled by a Tobit model (4), so the data from the heterogeneous populations can then be modeled by a mixture of Tobit and logit models, denoted as mTobit$\left(\omega_i, \mu_i, \sigma^2, L\right)$, given by

$$z_i \mid \mathbf{x}_i, \mathbf{u}_i \sim \text{i.d. mTobit}\left(\omega_i, \mu_i, \sigma^2, L\right), \quad \text{logit}\left(\omega_i\right) = \mathbf{u}_i^{\top}\beta_{\omega}, \quad \mu_i = \mathbf{x}_i^T\beta. \tag{7}$$

The mTobit model is composed of a logit model for the likelihood of non-exposure, and a Tobit model for the correlates of the exposure. The covariates $\mathbf{u}_i$ and $\mathbf{x}_i$ in (7) can be the same or different. If the likelihood of non-exposure doesn't depend on any covariates, no link function is needed. The likelihood function of the mTobit model is given by

$$L = \prod_{i=1}^{n} \left[\omega + (1-\omega)\Phi\left(\frac{L-\mu_i}{\sigma} \right) \right]^{c_i} \left[(1-\omega)\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(z_i-\mu_i)^2}{2\sigma^2} \right) \right]^{(1-c_i)}. \tag{8}$$

Again, the parameters β_{ω}, β and σ can be estimated by an MLE approach based on the likelihood (8) and model (7)

Under the Tobit model, for a given L, the proportion of the data under the DL is $E\left[\Phi\left(\frac{L-\mu_i}{\sigma}\right)\right]$, while under the mTobit model, the proportion of the data under the DL becomes $\omega + (1-\omega)E\left[\Phi\left(\frac{L-\mu_i}{\sigma}\right)\right]$, which is always greater than $E\left[\Phi\left(\frac{L-\mu_i}{\sigma}\right)\right]$ for $\omega > 0$. Therefore, ω is a parameter indicating excessive observations under the DL. Therefore, when the data exhibit more censored observations than what would be expected under Tobit model, the mTobit can be applied to address the heterogeneous issue.

2.2 As Independent Variable

Censored data are also often used as predictors to examine their relationship with some health-related outcomes. Because of the limitations of deletion and 'fill-in' methods, a joint modeling approach is developed to address the censoring issues. Similar to the censored data being treated as the outcome, we consider now censored predictors from either a single exposure or heterogeneous populations. To illustrate the issues of censored predictors, for simplicity, we consider the censored predictor as the only predictor for continuous outcomes as similar issues occur for other types of outcomes.

2.2.1 Predictor From Single Population

Suppose the underlying exposure $Z_i^* \sim N(\mu, \sigma_z^2)$, the continuous outcome Y_i is associated with Z_i^* through

$$y_i = \beta_0 + \beta_1 z_i^* + \varepsilon_i, \text{ where } \varepsilon_i \sim N(0, \sigma^2), \tag{9}$$

Due to DL, z_i^* is observed only if $z_i^* \geq L$, i.e., the predictor z_i is obtained based on (2), and the distribution of z_i is given by (3). Let $F(\cdot)$, $f(\cdot)$ and $g(\cdot)$ be the distribution of (Y, Z), $Y|Z$ and Z, separately. In joint modeling, we have

$$F(Y_i = y_i, Z_i = z_i) = f(Y_i = y_i|Z_i = z_i)g(Z_i = z_i). \tag{10}$$

For uncensored subjects, based on (9) and (3), we have

$$f(Y_i = y_i|Z_i = z_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{[y_{i-} (\beta_0 + \beta_1 z_i)]^2}{2\sigma^2}\right),$$
$$g(Z_i = z_i) = \frac{1}{\sqrt{2\pi}\sigma_z} \exp\left(-\frac{(z_i - \mu)^2}{2\sigma_z^2}\right).$$

Thus the joint likelihood for the uncensored subjects is given by

$$F(Y_i = y_i, Z_i = z_i) = \frac{1}{2\pi\sigma\sigma_z} \exp\left(-\frac{[y_{i-} (\beta_0 + \beta_1 z_i)]^2}{2\sigma^2} - \frac{(z_{i-}\mu)^2}{2\sigma_z^2}\right). \tag{11}$$

While for censored subjects with $Z_i < L$, the joint likelihood

$$F(Y_i = y_i, Z_i < L) = \int_{-\infty}^{L} F(Y_i = y_i, Z_i = t)dt = \int_{-\infty}^{L} f(Y_i = y_i|Z_i = t)g(Z_i = t)dt \tag{12}$$
$$= \int_{-\infty}^{L} \frac{1}{2\pi\sigma\sigma_z} \exp\left(-\frac{[y_{i-} (\beta_0 + \beta_1 t)]^2}{2\sigma^2} - \frac{(t_-\mu)^2}{2\sigma_z^2}\right)dt.$$

So the likelihood for the whole sample can be given by

$$\prod_{i=1}^{n} [F(Y_i = y_i, Z_i < L)]^{c_i} \cdot [F(Y_i = y_i, Z_i = z_i)]^{1-c_i} \tag{13}$$

By plugging (11) and (12) into (13) and applying MLE approach, the parameters $(\beta_0, \beta_1, \mu, \sigma_z^2, \sigma_z^2)$ can be estimated, and inference can be made.

2.2.2 Predictor From Heterogenous Populations

Assume the predictor z is from heterogeneous populations composed of non-exposure with probability ω and exposure with probability $(1 - \omega)$. Suppose the underlying exposure $z_i^* \sim N(\mu, \sigma_z^2)$, and z_i^* is observed only if $z_i^* \geq L$ due to DL. The predictor z_i is a mixture of censored exposure and non-exposure with distribution given in (6). As it is very likely that the relationships of exposure and non-exposure with outcome are different, we assume that relationships of the exposure and non-exposure with outcome y_i is

$$y_i = \begin{cases} \beta_0 + \beta_1 z_i^* + \varepsilon_i, & \text{for exposure ,} \\ \beta_0 + \beta_2 + \varepsilon_i & \text{for non-exposure,} \end{cases} \tag{14}$$

where $\varepsilon_i \sim N(0, \sigma^2)$. In (14), the coefficient β_1 describes the relationship between the exposure and outcome, while β_2 captures the relationship between the non-exposure and outcome.

For uncensored subjects ($c_i = 0$), based on (14) and (6), we have

$$f(Y_i = y_i|Z_i = z_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{[y_{i-} (\beta_0 + \beta_1 z_i)]^2}{2\sigma^2}\right),$$
$$g(Z_i = z_i) = (1 - \omega)\frac{1}{\sqrt{2\pi}\sigma_z} \exp\left(-\frac{(z_i - \mu)^2}{2\sigma_z^2}\right).$$

So the joint likelihood is given by

$$F(Y_i = y_i, Z_i = z_i) = \frac{(1 - \omega)}{2\pi\sigma\sigma_z} \exp\left(-\frac{[y_{i-} (\beta_0 + \beta_1 z_i)]^2}{2\sigma^2} - \frac{(z_{i-}\mu)^2}{2\sigma_z^2}\right). \tag{15}$$

Censored subjects can be from either the non-exposure or exposure population. Assume R_i be the membership indicator

with $R_i = 1(0)$ for non-exposure (exposure) and R_i is known, based on (14) and (6), the joint likelihood for non-exposure is given by

$$F(Y_i = y_i,\ R_i = 0) = \omega \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{[y_{i-} - (\beta_0 + \beta_2)]^2}{2\sigma^2}\right), \tag{16}$$

and the joint likelihood for censored exposure is

$$\begin{aligned} F(Y_i = y_i, Z_i < L, R_i = 1) &= \int_{-\infty}^{L} F(Y_i = y_i, Z_i = t)dt \\ &= \int_{-\infty}^{L} \frac{(1-\omega)}{2\pi\sigma\sigma_z} \exp\left(-\frac{[y_{i-} - (\beta_0 + \beta_1 t)]^2}{2\sigma^2} - \frac{(t-\mu)^2}{2\sigma_z^2}\right)dt. \end{aligned} \tag{17}$$

The joint likelihood for the whole sample including non-exposure, censored exposure and non-censored exposure is given by

$$\prod_{i=1}^{n} [F(Y_i = y_i,\ R_i = 0)]^{R_i} \ [F(Y_i = y_i, Z_i < L)]^{(1-R_i)c_i} \ [F(Y_i = y_i, Z_i = z)]^{(1-R_i)(1-c_i)}. \tag{18}$$

Plugging (15), (16), (17) into (18), and apply MLE approach, the parameters $(\beta_0, \beta_1, \beta_2, \mu, \sigma_z^2, \sigma^2)$ can be estimated and inference can be conducted.

3. Simulation Studies

Simulation studies are conducted to investigate statistical issues of different methods for handling censored data. For censored data treated as the outcome, the data are considered from single exposure and heterogeneous populations separately. Methods considered for handling the censoring issue include deletion, 'fill-in' by L or $\frac{1}{2}L$, and fitting the data with Tobit and mTobit models. For data treated as a predictor, in addition to deletion or 'fill-in' by L or $\frac{1}{2}L$, we also consider joint modeling for continuous outcome.

Different values for L are considered to reflect the varying amount of data censored. For data from heterogeneous populations, different prevalence (ω) of the non-exposure is also considered. In all the simulations, small (200), moderate (500) and large (1000) sample sizes are considered and a Mont Carlo (MC) sample size of 1000 is used.

3.1 As Dependent Variable

For Tobit models, we use the R function 'VGAM' to estimate parameters. As there is no R package available for mTobit models, we use a R function 'optim' to find the MLE of parameters based on (8).

3.1.1 From a Single Exposure Population

Let x, generated from $N(0, 1)$, be a predictor of the underlying exposure z^* through

$$z^* = \beta_0 + \beta_1 x + \epsilon, \quad \epsilon \sim N(0, 0.5), \tag{19}$$

Let $\beta_0 = 0.5$ and $\beta_1 = 1$, and let the outcome z be obtained by censoring z^* at L, where L is set to be $-0.93, -0.44, -0.086$, 0.22 and 0.50 in corresponding to $10\%, 20\%, 30\%, 40\%$ and 50% of data censored, respectively. The outcome z is fitted by 1) deleting all the censored observations; 2) filling-in the censored observations by L or $\frac{1}{2}L$; 3) Tobit model; 4) mTobit model to estimate β_1, which is of our interest.

The mean of the estimated β_1 across 1000 realizations for different methods are provided in Figure 1. Based on the figure, all the methods except the Tobit model yield biased estimates with the deletion method performing the worst. As expected, the bias becomes larger as more data censored. The patterns are similar across different sample sizes. In all the cases, the Tobit model gives unbiased estimates regardless of the proportion of censored data and sample size. For outcome from a single population, the mTobit model doesn't converge for most (75% to 100%) of the cases because of the lack of a non-exposure population ($\omega = 0$).

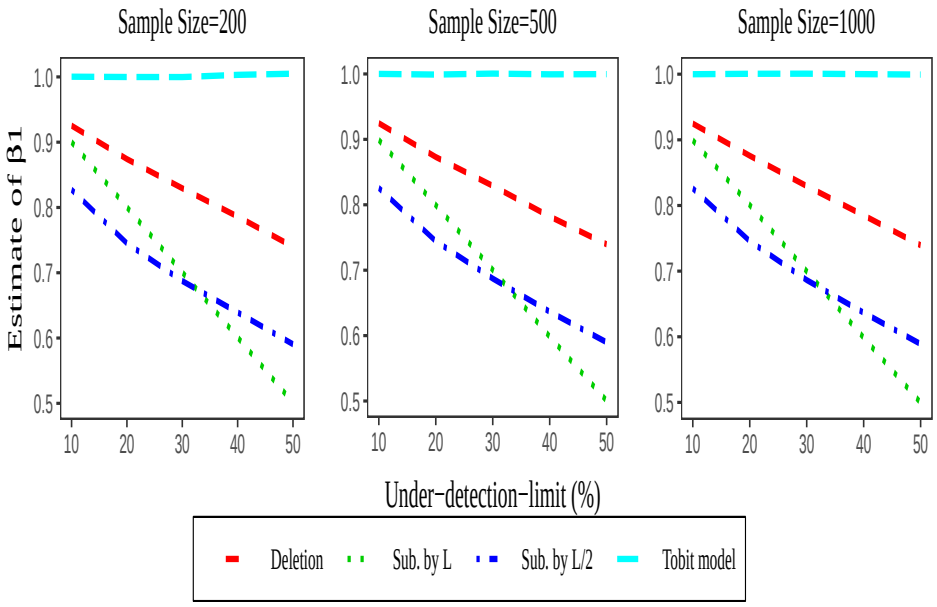

Figure 1. The mean of estimated β_1 (β_1 =1) for different methods when censored dependent variable from a single exposure population

3.1.2 From Heterogeneous Populations

Outcomes from heterogeneous population are generated in two steps. Let ω be the prevalence of non-exposure. In the first step, a membership R_i is generated to indicate whether subjects are exposed based on a Binomial (n, ω) distribution. In the second step, for subjects who are exposed, the underlying exposure z_i^* are generated as in (19). Specifically, the outcome z_i is generated by

$$z_i = \begin{cases} z_i^* & \text{if } R_i = 0 \text{ and } z_i^* \geq L, \\ \text{censored} & \text{if } R_i = 0 \text{ and } z_i^* < L, \\ \text{censored} & \text{if } R_i = 1. \end{cases}$$

Let ω range from 10% to 30%, and L be the same as in Section 3.1.1. The censored observations come from two sources, one from all non-exposed subjects, and the other from exposed subjects with values under L. The same methods as in Section 3.1.1 are considered.

The mean of the estimated β_1 for different L and ω across 1000 realizations are presented in Figure 2. All the methods except the mTobit model yield biased estimates. Unlike the data from a single exposure population, when the outcome is from heterogeneous populations, the Tobit model yields biased estimates with less bias for smaller non-exposure group. In all the cases, only the mTobit model gives unbiased estimates.

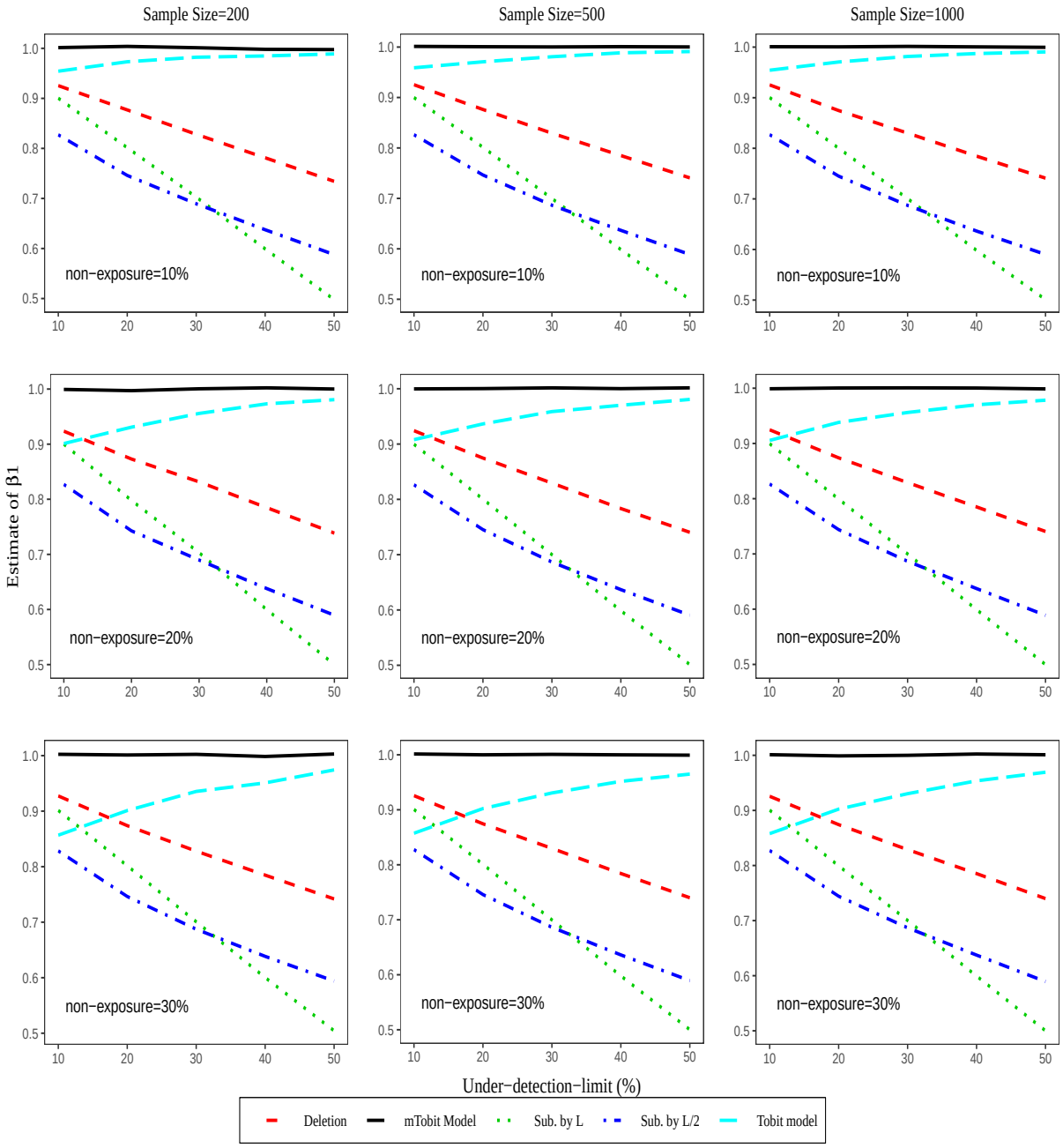

Figure 2. The mean of estimated β_1 (β_1 =1) for different methods when censored dependent variable from heterogenuous populations

3.2 As Independent Variable

3.2.1 From a Single Exposure Population

Let z^* be the underlying exposure and generated similarly as in (19). Specifically, z^* is generated from $N(0.5, 0.5)$, due to DL, with the predictor z censored at L; i.e., $z = z^*$ if $z^* \geq L$; otherwise z is censored.

Suppose the outcome y_i is associated with z_i^* through $y_i = \beta_0 + \beta_1 z_i^* + \varepsilon_i$ with $\varepsilon_i \sim \mathrm{N}(0, 0.5)$, $\beta_0 = -1.0$ and $\beta_1 = 1.0$. The outcome y_i is observed for all subjects, but z_i is censored for some subjects. In addition to deletion and 'fill-in' methods, joint modelling based on (13) is applied to estimate β_1.

Figure 3A summarizes the mean of the estimated β_1 across 1000 realizations. When the censored data are 'filled-in' by L or $\frac{1}{2}L$, the estimates are biased, and the bias becomes worse with more data censored. When the censored observations

are deleted or the data is fitted by joint modeling, the estimates are unbiased. However, the joint modeling yields more efficient estimates, with the relative efficiency ranging from 70% to 30% of that of the joint model when censoring ranges from 10% to 50% (see Figure 3B).

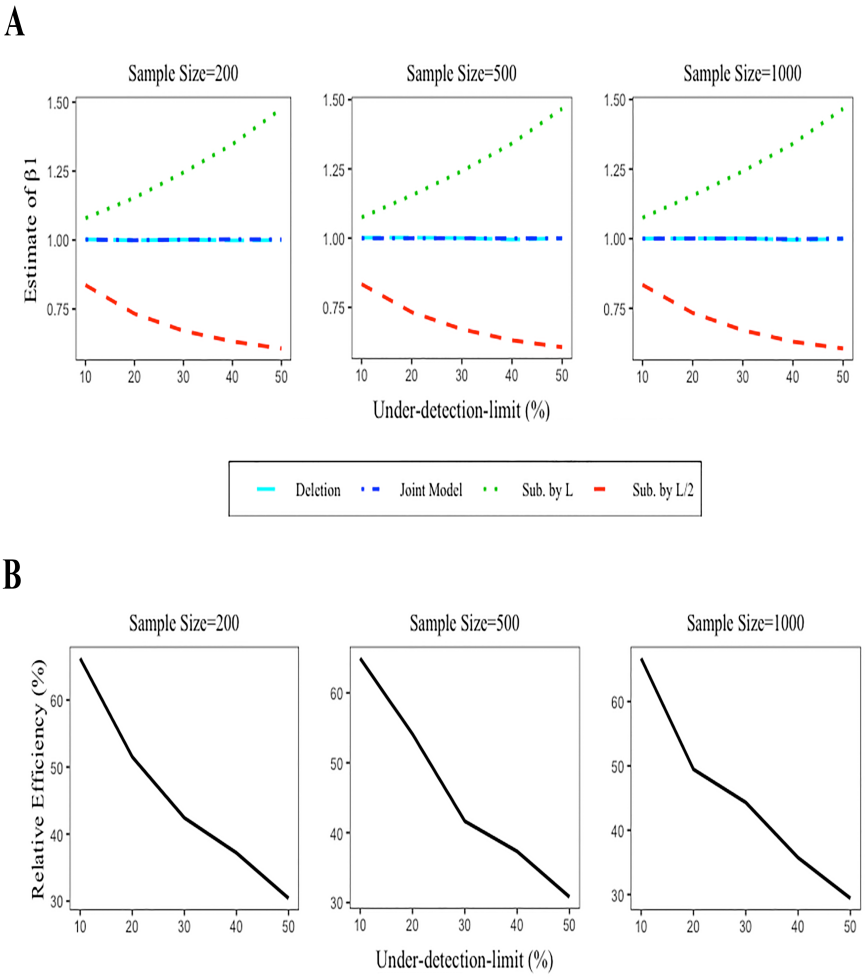

Figure 3. The mean of estimated β_1 (β_1 =1) for different methods when outcome is continuous and the predictor is from a single exposure population (A), and the relative efficiency of Deletion over the Joint Model (B)

3.2.2 From Heterogeneous Populations

The predictor z_i from a heterogeneous population is generated similarly as in Section 3.1.2. Specifically, a binary indicator R_i is first generated from Binomial (n, ω) to indicate if a subject is exposed or not. For subjects exposed, the underlying exposure z^* is generated from $N(0.5, 0.5)$, but censored at L, i.e.,

$$z_i = \begin{cases} z_i^* & \text{if } R_i = 0 \text{ and } z_i^* \geq L \\ \text{censored} & \text{if } R_i = 0 \text{ and } z_i^* < L \\ \text{censored} & \text{if } R_i = 1 \end{cases}$$

L is still set to be $-0.93, -0.44, -0.086, 0.22$ and 0.50 to correspond to $10\%, 20\%, 30\%, 40\%$ and 50% of data censored for the exposure population, respectively. We assume outcome y is associated with the exposure and non-exposure through

$$y_i = \begin{cases} \beta_0 + \beta_1 z_i + N(0, 0.5), & \text{for exposure ,} \\ \beta_0 + \beta_2 + N(0, 0.5) & \text{for non-exposure,} \end{cases} \tag{20}$$

to reflect different relationships of exposure and non-exposure with outcome, where $\beta_0 = -1.0, \beta_1 = 1.0$ and $\beta_2 = -1.0$.

We consider two scenarios. In the first scenario, the predictor is treated from a single population, i.e, all the censored observations are treated as the same and from the exposure population. In the second scenario, we assume that the membership is known, and we use the non-exposure ($R_i = 1$) to estimate β_2, but use the data from exposure population ($R_i = 1$) to estimate β_1 based on joint modeling.

The estimated β_1 are summarized in Figure 4A. The 'filled-in' methods yield biased estimates, and results in greater bias for larger non-exposure populations and increased censored data from the exposure population. The joint modelling produces biased estimates if the censored data is treated as the same, i.e., the censored observations are not differentiated between exposure and non-exposure, while when the membership is known, i.e., if the censored observations from the exposure group can be differentiated from non-exposure, the joint modelling yields unbiased estimates. The deletion method also gives unbiased estimates as deleting the censored observations doesn't change the relationship between the exposure and the outcome. However, the joint modelling is more efficient than the deletion method. The efficiency of the deletion is 70% to 30% of that under joint modeling when the censored data for the exposure population ranges from 10% to 50% (Figure 4B). Similar patterns are found for sample sizes 200 and 500, see Figure S1 and Figure 2 in the online supplementary material. Another advantage is that the joint modelling can estimate the relationship between the non-exposure and outcome, which is often important, both conceptually and clinically.

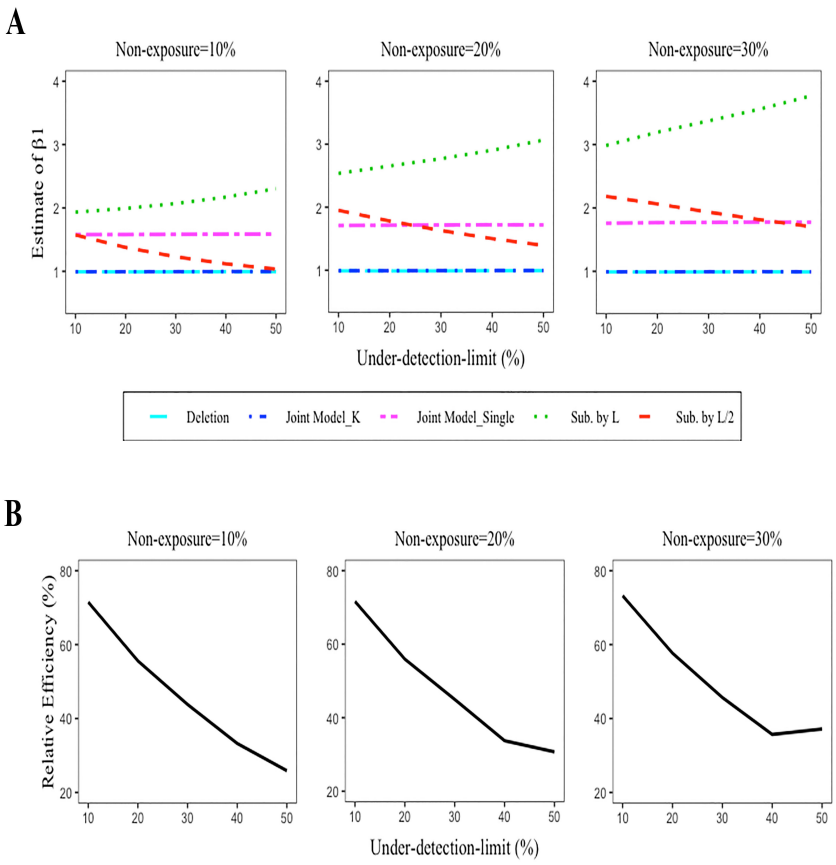

Figure S1. The mean of estimated β_1 (β_1 =1) for different methods when outcome is continuous and the predictor is from heterogenous populations (A), and the relative efficiency of Deletion over Joint Model_K (B) with sample size 200

Note: _K: the membership is known.

_Single: the predictor is treated as from a single population

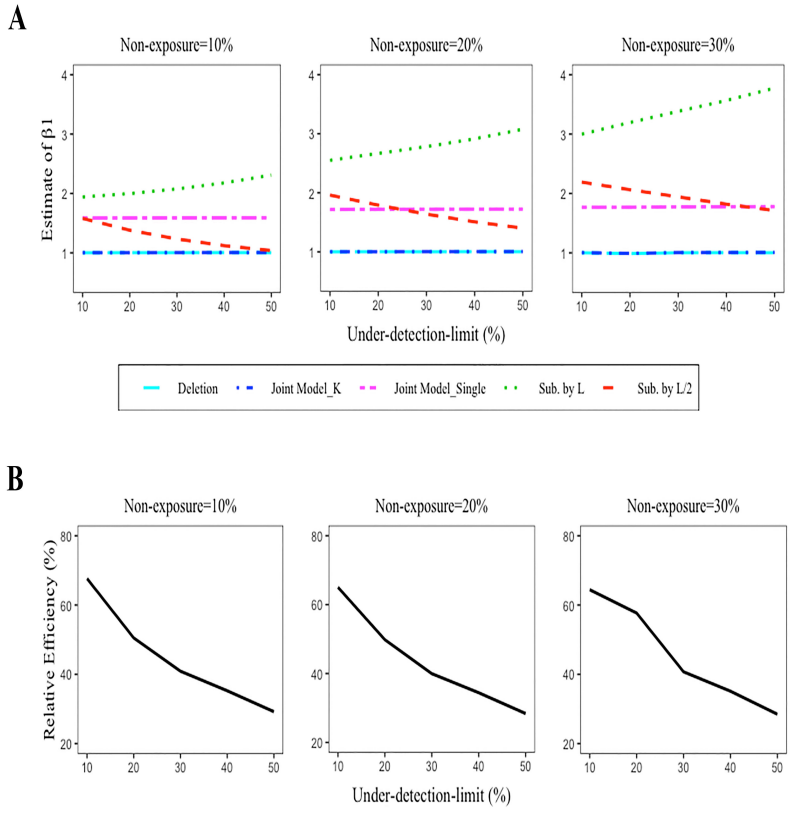

Figure S2. The mean of estimated β_1 (β_1 =1) for different methods when outcome is continuous and the predictor is from heterogenous populations (A), and the relative efficiency of Deletion over Joint Model_K (B) with sample size 500

Note: _K: the membership is known.

Single: the predictor is treated as from a single population.

4. Case Studies

Next we use two examples to illustrate the methods, the urine triclosan exposure and participants' BMI in NHANES 2003-2010 Study and the serum metabolites and BMI in the Bogalusa Heart Study. The two examples used here are not intended to make formal inferences, but are for the purpose of illustration.

4.1 NHANES 2003-2010 Study

NHANES is a continuous program that examines a nationally representative sample of about 5000 persons each year to assess health and nutritional status of adults and children in the general population of the USA (http://www.cdc.gov/nchs/nhanes.htm). Demographic, socioeconomic, dietary, and health-related data were collected via interviews. Blood and urine samples were also collected for laboratory testing. In four surveys conducted between 2003 and 2010, urinary triclosan concentration was measured in a random sample of 3659 children (6-19 years old) and 6566 adults (20 years or older). Of these, 2898 children and 5066 adults had detectable levels of urinary triclosan, which means that there are about 22% participates with their triclosan concentration undetected, and thus censored.

We examine the relationships of urine triclosan with age and BMI by treating triclosan as the outcome variable, with gender, race, cotinine and creatinine controlled, i.e., we consider

$$\text{Triclosan} \sim \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Gender} + \beta_3 \text{Race} + \beta_4 \text{BMI} + \beta_5 \text{Cotinine} + \beta_6 \text{Creatinine}.$$

Because of the skewed distribution of triclosan, a logarithm transformation is first applied.

Table 1 summaries the point estimates and the associated p-value from different methods. The deletion and 'fill-in' methods detect significant association between age and triclosan, however no association is detected between BMI and

triclosan for all the methods. Based on the simulation studies, the Tobit model is preferred to the deletion and 'fill-in' methods. The mTobit model doesn't converge, which is likely due to lack of non-exposure population. This is not surprising as the study participants are 6 years or older, it's very likely everyone has been exposed to triclosan to some degree. This also coincide with the results provided in (He et al., ress) that no non-exposure population was detected in urine triclosan measurement.

Table 1. The estimated associations of triclosan with age/BMI for different methods for handling the censored data

Method	Variable			
	age		BMI	
	Estimate	P-value	Estimate	P-value
Deletion	0.0066	< 0.001	-0.0054	0.0589
Sub. by L	0.0028	0.0016	-0.0004	0.8821
Sub. by $\frac{1}{2}L$	0.0024	0.0107	0.0003	0.9138
Tobit model	0.0016	0.1495	0.0020	0.5575
mTobit model	Does not converge			

4.2 Bogalusa Heart Study

Founded in 1973, the Bogalusa Heart Study (BHS) focuses on the early natural history of cardiovascular disease since childhood. More details about the BHS can be found here (https://www.clersite.org/bogalusaheartstudy/). In the current BHS study, a total of 1202 metabolites for 1261 unique BHS participants from the 2013-2016 visit cycle are obtained after quality control procedures and data cleaning. Among the 1202 metabolites, 167 have a below detection rate > 50% and 1035 have below detection rate ≤ 50%. Among the 1035 metabolites, 398, 401, 77, 56, 52 and 51 metabolites had 0%, <10%, 10-20%, 20-30%, 30-40% and 40-50% of below detection rate, respectively. We examine associations between metabolites and BMI with metabolites treated as outcome. For illustration purposes, we only consider 12 metabolites with known super-pathway and under-detection rate between 45% and 50%. The model considered is

$$\text{Metabolites} \sim \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Gender} + \beta_3 \text{Race} + \beta_4 \text{BMI}.$$

The results are summarized in Table 2. Based on the results, the mTobit model does not converge except for two metabolites, which implies that the two metabolites likely have a non-exposure group. These findings are consistent with the results presented in (He et al., ress), where the tests indicate that there is a non-exposure group in metabolites. For the two metabolites, the estimated associations are different between Tobit and mTobit model, but the results from mTobit model should be preferred because of the existence of non-exposure.

Table 2. The estimated associaton between BMI and metabolites with different methods for handling the censored data

	Deletion		Sub. by L		Sub. by $\frac{1}{2}L$		Tobit model		mTobit model	
Metab.	Est.	P-value	Est.	P-value	Est.	P-value	Est.	P-value	Est.	P-value
Y_16378[a]	0.0153	0.2186	0.0128	0.0539	0.0134	0.0465	0.0329	0.0032	-	-
Y_12249[a]	-0.0235	0.0830	-0.0093	0.2485	-0.0091	0.2584	0.0005	0.9729	-	-
Y_11590[a]	0.0021	0.5625	-0.0057	0.0285	-0.0062	0.0205	-0.0145	0.0015	-	-
Y_11250[b]	-0.0061	0.8733	-0.0197	0.3573	-0.0201	0.3482	-0.0804	0.0264	-	-
Y_12015[b]	-0.1553	0.0893	-0.1001	0.0516	-0.1003	0.0513	-0.1984	0.0215	-	-
Y_13185[c]	-0.0099	0.0114	-0.0096	0.0001	-0.0106	0.0001	-0.0182	0.0000	-	-
Y_13641[c]	-0.0124	0.0001	-0.0135	0.0000	-0.0150	0.0000	-0.0257	0.0000	-0.0267	0.0000
Y_14601[d]	0.0007	0.8227	-0.0015	0.5462	-0.0016	0.5367	-0.0036	0.4223	-0.0004	0.9151
Y_16089[d]	-0.0179	0.0758	-0.0131	0.0252	-0.0132	0.0248	-0.0218	0.0380	-	-
Y_16264[d]	0.0218	0.5596	0.0073	0.7213	0.0073	0.7240	-0.0242	0.4896	-	-
Y_12153[d]	-0.0302	0.7635	-0.0121	0.8138	-0.0121	0.8137	-0.0286	0.7496	-	-
Y_848[d]	0.0430	0.7367	-0.0096	0.8879	-0.0096	0.8879	-0.1902	0.1043	-	-

[a]: Amino Acid super-pathway
[b]: Lipid super-pathway
[c]: Peptide super-pathway
[d]: Xenobiotics super-pathway
−: indicates that the mTobit model doesn't converge.

5. Discussion

In this paper, we investigate statistical issues in analyzing censored data due to detection limits via intensive simulation studies. Based on the simulation studies, when the data is treated as a dependent variable, the deletion and 'fill-in' methods yield biased estimate and thus shouldn't be used. Only Tobit and mTobit models should be opted for. If the outcome is from a single exposure population, the Tobit model can be applied, while for outcomes from heterogeneous populations such as exposure and non-exposure, the mTobit model should be used.

When the censored data are treated as predictors, the 'fill-in' method always yield biased estimates and hence shouldn't be used. The deletion method usually produces unbiased estimates, but it is highly inefficient. For censored predictors from a single population, joint modeling yields unbiased estimates, and is much more efficient than the deletion method, and hence should be preferred. For predictors from heterogeneous populations, the joint modeling needs to differentiate the two populations in order to achieve unbiased estimates.

In the paper, the censored predictor is considered as the only predictor. When other predictors/covariates need to be included, and the censored predictor is correlated with other predictors/covariates, joint modeling is much more involved and needs to be developed to achieve both unbiasedness and efficiency. Joint modeling becomes even more challenging for censored predictors from heterogeneous populations with unknown membership, which is a universal case. However, regardless of if the censored data is treated as an outcome or a predictor, a foremost question is to test if the data is from a single exposure population or from heterogeneous populations in order to choose an appropriate method. Available tests for testing if there is a non-exposure population can be found in (He et al., ress), where three tests including Wald test, likelihood ratio test and score test are developed.

Acknowledgements

This work was supported by the NIH under grants R01GM108337 and P20GM109036.

References

Al-Hanawi, M. K., Alsharqi, O., & Vaidya, K. (2018). Willingness to pay for improved public health care services in saudi arabia: a contingent valuation study among heads of saudi households. *Health Economics, Policy and Law*, 1-28.

Amemiya, T. (1984). Tobit models: A survey. *Journal of econometrics, 24*(1-2), 3-61. https://doi.org/10.1016/0304-4076(84)90074-5

Austin, P. C., & Hoch, J. S. (2004). Estimating linear regression models in the presence of a censored independent variable. *Statistics in medicine, 23*(3), 411-429. https://doi.org/10.1002/sim.1601

Bernhardt, P. W., Wang, H. J., & Zhang, D. (2015). Statistical methods for generalized linear models with covariates subject to detection limits. *Statistics in biosciences, 7*(1), 68-89.

Deng, Z., Yan, J., & Sun, P. (2019). Political status and tax haven investment of emerging market firms: Evidence from china. *Journal of Business Ethics*, 1-20. https://doi.org/10.1007/s10551-018-4090-0

Ferrero, A., Esplugues, A., Estarlich, M., Llop, S., Cases, A., Mantilla, E., ... Iñiguez, C. (2017). Infants' indoor and outdoor residential exposure to benzene and respiratory health in a spanish cohort. *Environmental pollution, 222*, 486-494.

Gleit, A. (1985). Estimation for small normal data sets with detection limits. *Environmental science & technology, 19*(12), 1201-1206. https://doi.org/10.1021/es00142a011

Gomez, C., Gonzalez-Riano, C., Barbas, C., Kolmert, J., Ryu, M. H., Carlsten, C., ... Wheelock, C. E. (2019). Quantitative metabolic profiling of urinary eicosanoids for clinical phenotyping. *Journal of lipid research, 60*(6), 1164-1173. https://doi.org/10.1194/jlr.D090571

He, H., Tang, W., Kelly, T., Li, S., & He, J. (in press). Statistical tests for latent class in censored data due to detection limit. *Statistical methods in medical research*.

Hezaveh, A. M., & Cherry, C. R. (2019). Neighborhood-level factors affecting seat belt use. *Accident Analysis & Prevention, 122*, 153-161. https://doi.org/10.1016/j.aap.2018.10.005

Javad, M. T., Vahidinia, A., Samiee, F., Elaridi, J., Leili, M., Faradmal, J., & Rahmani, A. (2018). Analysis of aluminum, minerals and trace elements in the milk samples from lactating mothers in hamadan, iran. *Journal of Trace Elements in Medicine and Biology, 50*, 8-15.

Keeley, M. C., Robins, P. K., Spiegelman, R. G., & West, R. W. (1978). The labor-supply effects and costs of alternative negative income tax programs. *Journal of Human Resources*, 3-36.

Kim, S., Chang, Y., Sung, E., Kang, J. G., Yun, K. E., Jung, H.-S., ... Ryu, S. (2018). Association between sonographically diagnosed nephrolithiasis and subclinical coronary artery calcification in adults. *American Journal of Kidney Diseases, 71*(1), 35-41.

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

Leng, Y., Liu, W., Xiao, N., Li, Y., & Deng, J. (2019). The impact of policy on the intangible service efficiency of the primary health care institution-based on china's health care reform policy in 2009. *International journal for equity in health, 18*(1), 14.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data, Second Edition.* Hoboken: Wiley. https://doi.org/10.1002/9781119013563

Lynn, H. S. (2001). Maximum likelihood inference for left-censored hiv rna data. *Statistics in medicine, 20*(1), 33-45. https://doi.org/10.1002/1097-0258(20010115)20:1¡33::AID-SIM640¿3.0.CO;2-O

Maule, A. L., Scarpaci, M. M., & Proctor, S. P. (2019). Urinary concentrations of permethrin metabolites in us army personnel in comparison with the us adult population, occupationally exposed cohorts, and other general populations. *International Journal of Hygiene and Environmental Health*.

McDonald, J. F., & Moffitt, R. A. (1980). The uses of tobit analysis. *The review of economics and statistics*, 318-321. https://doi.org/10.2307/1924766

Moulton, L. H., & Halsey, N. A. (1995). A mixture model with detection limits for regression analyses of antibody response to vaccine. *Biometrics*, 1570-1578. https://doi.org/10.2307/2533289

Nassan, F. L., Coull, B. A., Gaskins, A. J., Williams, M. A., Skakkebaek, N. E., Ford, J. B., ... Hauser, R. (2017). Personal care product use in men and urinary concentrations of select phthalate metabolites and parabens: results from the environment and reproductive health (earth) study. *Environmental health perspectives, 125*(8).

Newman, M. C., Dixon, P. M., Looney, B. B., & Pinder III, J. E. (1989). Estimating mean and variance for environmental samples with below detection limit observations 1. *JAWRA Journal of the American Water Resources Association, 25*(4), 905-916.

Olson, D. (1993). A simple method for estimation when there is a detection limit. In *Joint Statistical Meeting of American Statistical Society and Biometric Society*.

Østergren, P. B., Kistorp, C., Fode, M., Henderson, J., Bennedbæk, F. N., Faber, J., & Sønksen, J. (2017). Luteinizing hormone-releasing hormone agonists are superior to subcapsular orchiectomy in lowering testosterone levels of men with prostate cancer: results from a randomized clinical trial. *The Journal of urology, 197*(6), 1441-1447.

Park, J., Park, S. K., & Choi, Y.-H. (2018). Environmental pyrethroid exposure and diabetes in us adults. *Environmental Research*. https://doi.org/10.1016/j.envres.2018.12.043

Reisetter, A. C., Muehlbauer, M. J., Bain, J. R., Nodzenski, M., Stevens, R. D., Ilkayeva, O., ... Scholtens, D. M. (2017). Mixture model normalization for non-targeted gas chromatography/mass spectrometry metabolomics data. *BMC bioinformatics, 18*(1), 84.

Rigobon, R., & Stoker, T. M. (2003). Censored regressors and expansion bias. https://doi.org/10.2139/ssrn.475481

Rosen, H. S. (1976). Taxes in a labor supply model with joint wage-hours determination. *Econometrica: Journal of the Econometric Society*, 485-507. https://doi.org/10.2307/1913978

Slymen, D. J., de Peyster, A., & Donohoe, R. R. (1994). Hypothesis testing with values below detection limit in environmental studies. *Environmental science & technology, 28*(5), 898-902. https://doi.org/10.1021/es00054a022

Taylor, D. J., Kupper, L. L., Rappaport, S. M., & Lyles, R. H. (2001). A mixture model for occupational exposure mean testing with a limit of detection. *Biometrics, 57*(3), 681-688.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica: journal of the Econometric Society*, pages 24-36. https://doi.org/10.2307/1907382

Zhao, Q., Shen, H., Su, K.-J., Zhang, J.-G., Tian, Q., Zhao, L.-J., ... Deng, H. W. (2018). Metabolomic profiles associated with bone mineral density in us caucasian women. *Nutrition & metabolism, 15*(1), 57.

Zhou, C., Yu, N. N., & Losby, J. L. (2018). The association between local economic conditions and opioid prescriptions among disabled medicare beneficiaries. *Medical care, 56*(1), 62-68. https://doi.org/10.1097/MLR.0000000000000841