Published in final edited form as:

Environmetrics. 2018 December ; 29(8): . doi:10.1002/env.2536.

Linear regression with left-censored covariates and outcome using a pseudolikelihood approach

Michael P. Jones

Department of Biostatistics, University of Iowa, Iowa City, IA 52242, U.S.A.

Abstract

Environmental toxicology studies often involve sample values that fall below a laboratory procedure's limit of quantification. Such left-censored data give rise to several problems for regression analyses. First, both covariates and outcome may be left censored. Second, the transformed toxicant levels may not be normal but mixtures of normals because of differences in personal characteristics, e.g. exposure history and demographic factors. Third, the outcome and covariates may be linear functions of left-censored variates, such as averages and differences. Fourth, some toxicant levels may be functions of other toxicant levels resulting in a recursive system. In this paper marginal and pseudo-likelihood based methods are proposed for estimation of the means and covariance matrix of variates found in these four settings. Next, linear regression methods are developed allowing outcomes and covariates to be linear combinations of left-censored measures. This is extended to a recursive system of modeling equations. Bootstrap standard errors and confidence intervals are used. Simulation studies demonstrate the proposed methods are accurate for a wide range of study designs and left-censoring probabilities. The proposed methods are illustrated through the analysis of an on-going community-based study of polychlorinated biphenyls, which motivated the proposed methodology.

Keywords

environmental exposure; limit of quantification; polychlorinated biphenyls; toxicology

1 | INTRODUCTION

A common problem in environmental exposure assessment is the occurrence of contaminant levels that fall below a laboratory's ability for accurate measurement. Values below this limit of quantification (LOQ) are known as left-censored observations. This problem occurs when a laboratory assay's LOQ is high relative to typical field exposure, or when the sampling period is short (Hewett, 2006, p. 415). Thousands of chemicals are in use today, with more introduced every year, and laboratory assays of their levels are frequently not quantifiable at the lower levels. This is an important public health and policy issue since values below the LOQ may be toxic or bioaccumulate to toxic levels. Quantification limits are common with air, water, soil/sediment, food, blood, urine and tissue samples. A few examples include herbicides/pesticides/nitrates in the water supply (Holden et al., 1992; Jones et al. 2016); toxic metals (e.g. lead, chromium) in highway runoff (Shumway et al., 2002) and public water supply lines, as in Flint, MI; okadaic acid, arsenic and cadmium in food (EFSA, 2010; The Economist, 2017); endotoxin levels of environments with organic dust (Thorne et al.,

2010); and polychlorinated biphenyls (PCBs) in air and blood samples (Korwel et al., 2005; Marek et al., 2013). The motivation for the proposed methods of this paper derives from an on-going community-based study assessing blood PCB levels and the need to easily handle left-censoring of the predictors as well as the outcome variable in linear regression models.

The distribution of environmental contaminants is typically right skewed and assumed to follow a lognormal distribution for which industrial hygienists and environmental toxicologists historically have estimated summary statistics for a single variate. Although maximum likelihood estimation (MLE) was introduced early by Hald (1949) to handle left censoring due to measurements below the LOQ, the methods were computationally laborious and required special tables. This led to single imputation of left-censored values by the LOQ, LOQ/2 (Nehls & Akland, 1973) and LOQ/ $\sqrt{2}$ (Hornung & Reed, 1990). Many researchers (e.g. Hewett & Ganser, 2007) have studied these imputation methods via simulation, showing that MLE is far more accurate. For regression problems historical development focused first on the problem in which only the outcome variate is censored. Maximum likelihood regression methods for censored outcomes with completely observed covariates have a long successful history in both the statistical (Kalbfleisch & Prentice, 2002) and econometric literature (Tobin, 1958), and yield consistent, asymptotically normal estimates. The problem of longitudinal or spatial outcomes subject to left censoring is more complex but has been tackled by full likelihood (Gadda et al., 2000), Monte-Carlo EM (Hughes, 1999) and Bayesian methods (Fridley & Dixon, 2006).

Recently, there has been increased interest in regression problems having usually one, sometimes more, censored covariates, but assuming a completely observed outcome. Several approaches have been proposed. A simple regression approach is the complete-case analysis that omits subjects with any censored variates. When only the covariates are subject to censoring, the complete-case analysis will produce unbiased parameter estimates so long as the regression error term remains uncorrelated with the observed covariates. The key concern is that efficiency decreases as the sample size decreases which can be substantial with heavy censoring. A second approach is single imputation, which includes the previously mentioned functions of LOQ and the expected value below the LOQ (Richardson & Ciampi, 2003). These tend to exhibit either bias in the parameter estimate, its standard error or both (Nie et al., 2010). A third and more sophisticated approach is multiple imputation. For the outcome Y modeled as a linear function of a right-censored Z and uncensored U , Atem et al. (2017) sampled from the predictive distribution of Z given the values of U and Y , substituting in parameter estimates derived from a complete-case regression of Y on (Z, U) and a Cox proportional hazards regression model fit to Z given U . Their approach is appropriate for right-censored Z , but less so for our left-censored setting, in part because this family does not contain the lognormal and, as they noted, due to their extrapolations beyond the range of uncensored values, especially problematic for heavily censored variates. The fourth approach is the EM algorithm. May et al. (2011) introduced approximate ML estimation based on a weighted Monte-Carlo EM algorithm in which the conditional Gaussian distribution of censored variate(s) given the values of uncensored values, including the outcome, were repeatedly sampled using the Gibbs sampler and the

adaptive rejection metropolis algorithm. Unlike other authors, May et al. (2011) allowed more than one covariate to be censored.

On another note, there has also been recent interest in estimation of the covariance matrix for a large number of left-censored variates. In 2015, our paper (Jones et al.) and two others (Hoffman & Johnson, 2015; Pesonen et al., 2015) tackled this problem. Although the three methods differed somewhat in their approaches, each came to the conclusion that a pairwise likelihood approach to estimating the correlations was the key to finding a computationally efficient method for finding a consistent set of estimators. For the most part, all assumed multivariate normality of variates. However, this assumption is not realistic enough for many real world applications since the variates may only be normal within homogeneous strata of the population defined by both discrete and continuous variables (e.g. gender and age), thereby resulting in a complex mixture of normals.

An important but so far ignored area of research concerns environmental problems in which both the outcome variate and covariates are subject to left censoring, or more generally, are linear functions of censored variates, such as a difference or an average. This is the goal of this paper. In Section 2, we introduce methods for estimating the covariance matrix of left-censored variates that can be modeled in terms of a set of linear equations based on uncensored variates and normal error terms. This is then generalized to covariance matrices of linear combinations of left-censored variates. These estimators are used as building blocks in Section 3 to create multiple linear regression methods in which both the outcome and covariates can be linear functions of left-censored variates. This is further generalized to incorporate a recursive system of regression equations to accommodate censored covariates that are generated from other censored covariates. Section 4 studies in detail three designs one might find in environmental epidemiology, the last of which extends the framework of the recursive system of regressions. This section also reports on simulation studies of the performance of these methods. Section 5 provides an in-depth application of the proposed methods to the PCB study data that motivated our methodology. Section 6 concludes with some overall discussion.

2 | ESTIMATION OF THE COVARIANCE MATRIX

Section 2.1 begins by postulating that the censored variates can be written as a set of linear functions of the uncensored covariates plus normal possibly correlated error terms. We propose an extension to the maximum pairwise pseudo-likelihood technique of Jones et al. (2015) for estimating the covariance matrix. Section 2.2 considers the special case of no uncensored covariates and provides a warning. Section 2.3 describes a straightforward approach to estimation of the covariance matrix for variates that are linear combinations of the censored and uncensored variates.

2.1 | Estimation for uncensored and left-censored variates

Let $\mathbf{D}_i^\circ = (U_{i1}, \dots, U_{iq}, Z_{i1}^\circ, \dots, Z_{ip}^\circ)'$ be the i -th unit's variates of which the Z° 's will be subject to left censoring, whereas the U 's are uncensored and so measured exactly. The special case $q = 0$ is allowed and will be discussed later. Let $\mathbf{L}_i = (L_{i1}, \dots, L_{ip})'$ be the laboratory lower

limits of quantification for measuring the Z_i^o variates. (The lower limit may depend on the i -th unit's sample volume or other characteristics and hence is subscripted by i .) There are three basic assumptions. (A1) The vector of Z^o variates are conditionally multivariate normal given the U variates. It will be convenient for later use to write this out explicitly as

[illegible]

where $(\varepsilon_{i1}, \dots, \varepsilon_{ip})' \sim N_p(\mathbf{0}, \Sigma_\varepsilon)$ for n independent units. Both p and q are fixed. As for regression models in general, one should avoid models with a large number of variates relative to the sample size. No distributional assumptions are made about the U variates. (A2) As usual in regression, we assume the U variates are uncorrelated with the ε error terms. (A3) Z_{ij}° is assumed conditionally independent of L_{ij} given $\mathbf{U}_i$. If necessary, we assume that the Z° variates have already been appropriately transformed so that the error structure is normal. For laboratory data this is often the log transform. Next, let $Z_{ij} = \max(Z_{ij}^\circ, L_{ij})$ and $\eta_{ij} = 1 \left[Z_{ij}^\circ \geq L_{ij} \right]$ be the observed variates, so that Z_{ij} is uncensored if $\eta_{ij} = 1$, but is left censored if $\eta_{ij} = 0$. As an example using the PCB study, Z_{ij} is the log-transformed, possibly left-censored (below LOQ) measurement of PCB congener j in a blood sample from subject i .

The goal of this section is to estimate the marginal mean and marginal covariance matrix of $\mathbf{D}^\circ = (\mathbf{U}, \mathbf{Z}^\circ)$ from the observed data $\mathbf{D} = \{(\mathbf{U}_i, \mathbf{Z}_i, \boldsymbol{\eta}_i) : i = 1, \dots, n\}$ where

$\mathbf{U}_i = (U_{i1}, \dots, U_{iq})'$, $\mathbf{Z}_i = (Z_{i1}, \dots, Z_{ip})'$ and $\boldsymbol{\eta}_i = (\eta_{i1}, \dots, \eta_{ip})'$. We propose a straight-forward two-stage approach using marginal likelihood and pairwise pseudo-likelihood, that is particularly appealing when p is reasonably large.

In Stage 1, the marginal means and variances of the $\mathcal{Z}$'s are estimated. We begin by considering the j^{th} marginal model from (1), namely, $Z_{ij}^o = \phi_{j0} + \phi_j' \mathbf{U}_i + \varepsilon_{ij}$ ($i = 1, \dots, n$) with $\varepsilon_{ij} \sim N(0, \sigma_{\varepsilon j}^2)$. The parameters $(\phi_{j0}, \phi_j, \sigma_{\varepsilon j}^2)$ are estimated by maximizing the j -th likelihood given $\mathbf{U}_i$

$$\mathcal{L}_j(\phi_{j0}, \phi_j, \sigma_{\varepsilon j}^2) = \prod_{i=1}^n \left\{ f(z_{ij} - \phi_{j0} - \phi_j' \mathbf{U}_i; 0, \sigma_{\varepsilon j}^2) \right\}^{\eta_{ij}} \left\{ F(z_{ij} - \phi_{j0} - \phi_j' \mathbf{U}_i; 0, \sigma_{\varepsilon j}^2) \right\}^{1 - \eta_{ij}}$$

where f and F are the normal density and distribution function; $(0, \sigma_{\varepsilon j}^2)$ refer to their mean and variance. The resultant maximum likelihood estimators $(\hat{\phi}_{j0}, \hat{\phi}_j, \hat{\sigma}_{\varepsilon j}^2)$ are consistent. Maximization is repeated for each Z_j separately ($j = 1, \dots, p$) and easily performed by

existing software, e.g. the `survreg` function in R (R Development Core Team, 2017). The marginal mean $EZ_j^\circ = E_{\mathbf{U}}\{E(Z_j^\circ|\mathbf{U})\}$ can therefore be estimated by

$$\hat{EZ}_j^\circ = \hat{\phi}_{j0} + \hat{\phi}_{j1}\bar{U}_{\cdot 1} + \dots + \hat{\phi}_{jq}\bar{U}_{\cdot q} \quad (j = 1, \dots, p)$$

where $\bar{U}_{\cdot \ell} = n^{-1}\sum_i U_{i\ell}$. We shall refer to this estimator as the maximum marginal likelihood estimator (MMLE) since it estimates a marginal mean and uses no information from the other Z variates. The marginal variance is

$$\text{Var}(Z_j^\circ) = \text{Var}_{\mathbf{U}}\{E(Z_j^\circ|\mathbf{U})\} + E_{\mathbf{U}}\{\text{Var}(Z_j^\circ|\mathbf{U})\} = \text{Var}_{\mathbf{U}}\{\phi_j'\mathbf{U}\} + \text{Var}(\varepsilon_j) = \phi_j' \Sigma_{\mathbf{U}\mathbf{U}} \phi_j + \sigma_{\varepsilon_j}^2$$

which can be estimated by

$$\widehat{\text{Var}}(Z_j^\circ) = \hat{\phi}_j' \hat{\Sigma}_{\mathbf{U}\mathbf{U}} \hat{\phi}_j + \hat{\sigma}_{\varepsilon_j}^2, \quad (2)$$

where $\hat{\Sigma}_{\mathbf{U}\mathbf{U}}$ is the usual sample covariance matrix of $\mathbf{U}$. As typical in regression settings, marginal means and variances of the outcome depend on representative samples of subjects for covariate patterns, and one should be careful in extrapolating these results beyond the range of the data, such as age ranges not covered by the sample. Two generally useful facts are that the population linear regression coefficients and intercept can be written as $\phi_j = \Sigma_{\mathbf{U}\mathbf{U}}^{-1} \Sigma_{\mathbf{U}Z_j^\circ}$ and $\phi_{j0} = EZ_j^\circ - \phi_j' E\mathbf{U}$, respectively. Hence, we can estimate the covariance $\text{Cov}(\mathbf{U}, Z_j^\circ)$ by

$$\hat{\Sigma}_{\mathbf{U}Z_j^\circ} = \hat{\Sigma}_{\mathbf{U}\mathbf{U}} \hat{\phi}_j. \quad (3)$$

In Stage 2, we propose a pairwise pseudo-likelihood approach for estimating the correlation and covariance matrices of $\mathbf{Z}^\circ$. From model (1) and the fact that any pair of error terms is bivariate normal, i.e. $(\varepsilon_j, \varepsilon_m) \sim N_2(0, 0, \sigma_{\varepsilon_j}^2, \sigma_{\varepsilon_m}^2, \rho_{\varepsilon_{jm}})$, it follows that

$$\begin{aligned} \text{Cov}(\varepsilon_j, \varepsilon_m) &= \text{Cov}\left\{Z_j^\circ - EZ_j^\circ - \Sigma_{Z_j^\circ \mathbf{U}} \Sigma_{\mathbf{U}\mathbf{U}}^{-1} (\mathbf{U} - E\mathbf{U}), \right. \\ &\quad \left. Z_m^\circ - EZ_m^\circ - \Sigma_{Z_m^\circ \mathbf{U}} \Sigma_{\mathbf{U}\mathbf{U}}^{-1} (\mathbf{U} - E\mathbf{U})\right\} \\ &= \text{Cov}(Z_j^\circ, Z_m^\circ) - \Sigma_{Z_j^\circ \mathbf{U}} \Sigma_{\mathbf{U}\mathbf{U}}^{-1} \Sigma_{\mathbf{U}Z_m^\circ} \end{aligned}$$

and hence

$$\text{Cov}(Z_j^\circ, Z_m^\circ) = \Sigma_{Z^\circ U} \Sigma_{UU}^{-1} \Sigma_{UZ_m^\circ} + \text{Cov}(\varepsilon_j, \varepsilon_m). \quad (4)$$

We only need to estimate $\text{Cov}(\varepsilon_j, \varepsilon_m)$, $j \neq m$, since the other covariance terms in (4) are estimated in Stage 1.

In Stage 2, the pairwise error correlations $\rho_{\varepsilon jm}$ are estimated by generalizing the pairwise pseudo-likelihood approach proposed by Jones et al. (2015) in order to handle the setting described by (1). For each pair $(\varepsilon_j, \varepsilon_m)$, we form the pair-specific pseudo-likelihood

$$PPL_{jm}(\rho_{\varepsilon jm}) = \mathcal{L}_{jm}(0, 0, \hat{\sigma}_{\varepsilon j}^2, \hat{\sigma}_{\varepsilon m}^2, \rho_{\varepsilon jm}) \quad (5)$$

where L_{jm} is the left-censored bivariate normal likelihood and $(\hat{\sigma}_{\varepsilon j}^2, \hat{\sigma}_{\varepsilon m}^2)$ are the MMLEs of the variances found in Stage 1. This likelihood is based solely on $\{(r_{ij}, \eta_{ij}, r_{im}, \eta_{im}) : i = 1, \dots, n\}$ where

$$r_{ij} = Z_{ij} - \hat{E}Z_{ij} - \hat{\Sigma}_{Z^\circ U} \hat{\Sigma}_{UU}^{-1} (U_{ij} - \bar{U}_{\cdot j}).$$

Thus, r_{ij} is the estimated (uncensored) residual $\hat{\varepsilon}_{ij}$ if $\eta_{ij} = 1$ and is left censored if $\eta_{ij} = 0$. The pairwise normal likelihood $\mathcal{L}_{jm}(0, 0, \hat{\sigma}_{\varepsilon j}^2, \hat{\sigma}_{\varepsilon m}^2, \rho_{\varepsilon jm})$ for left-censored data is

$$\begin{aligned} & \prod_{i=1}^n \left\{ f_{jm}(r_{ij}, r_{im}) \right\}^{\eta_{ij} \eta_{im}} \left\{ (\partial / \partial a) F_{jm}(a, r_{im}) \Big|_{a=r_{ij}} \right\}^{\eta_{ij} (1 - \eta_{im})} \\ & \times \left\{ (\partial / \partial b) F_{jm}(r_{ij}, b) \Big|_{b=r_{im}} \right\}^{(1 - \eta_{ij}) \eta_{im}} \left\{ F_{jm}(r_{ij}, r_{im}) \right\}^{(1 - \eta_{ij}) (1 - \eta_{im})}, \end{aligned}$$

where f_{jm} and F_{jm} are the bivariate normal density and distribution function with means (0, 0), variances evaluated at $(\hat{\sigma}_{\varepsilon j}^2, \hat{\sigma}_{\varepsilon m}^2)$ and free correlation parameter $\rho_{\varepsilon jm}$. Maximization of $PPL_{jm}(\rho_{\varepsilon jm})$ produces the maximum pairwise pseudo-likelihood estimator (MPPLE) $\hat{\rho}_{\varepsilon jm}$. This maximization is repeated for all $p(p-1)/2$ pairings to compute the MPPLE $\hat{\mathbf{P}}$ of the error correlation matrix $\mathbf{P}$. Computing formulae for the likelihood components can be found in Jones et al. (2015), along with other details on maximization.

The MPPLE of the covariance matrix Σ_ε of error terms in (1) is defined as $\hat{\Sigma}_\varepsilon = \hat{\mathbf{V}}^{1/2} \hat{\mathbf{P}} \hat{\mathbf{V}}^{1/2}$, where $\hat{\mathbf{V}}$ is the Stage 1-derived diagonal matrix whose diagonal contains the MMLEs $(\hat{\sigma}_{\varepsilon 1}^2, \dots, \hat{\sigma}_{\varepsilon p}^2)$. Since the MMLEs $(\hat{\sigma}_{\varepsilon j}^2, \hat{\sigma}_{\varepsilon m}^2)$ are $\sqrt{n}$ -consistent estimators, the maximizer

$\hat{\rho}_{\varepsilon jm}$ of the pseudo-likelihood $PPL_{jm}(\rho_{\varepsilon jm})$ is also consistent by the pseudo-likelihood theory of Gong and Samaniego (1981). By the continuous mapping theorem the MPPL $\hat{\Sigma}_{\varepsilon}$ is therefore consistent. The MPPL $\hat{\Sigma}_{Z^{\circ}Z^{\circ}}$ of the covariance matrix of $\mathbf{Z}^{\circ}$ uses consistent plug-in estimators for the population parameters in equation (4) calculated in Stage 1 and Stage 2 and hence is also consistent. At this point we have consistent estimators of $\mu_{D^{\circ}}$ and $\Sigma_{D^{\circ}}$.

2.2 | Case of no uncensored covariates

In the case of no U covariates ($q = 0$), the assumption in (1) implies, for all (j, m) pairings,

$$(Z_{1j}^{\circ}, Z_{1m}^{\circ}), \dots, (Z_{nj}^{\circ}, Z_{nm}^{\circ}) \sim iid N_2(\mu_j, \mu_m, \sigma_{\varepsilon j}^2, \sigma_{\varepsilon m}^2, \rho_{\varepsilon jm}) \quad (6)$$

where $\mu_{\varepsilon} = \phi_{\varepsilon 0}$ in (1). This is precisely the same setting explored by Jones et al. (2015). Stage 1, as described above, is used to estimate (μ_j, μ_m) and $(\sigma_{\varepsilon j}^2, \sigma_{\varepsilon m}^2)$, the marginal mean and variance of $(Z_j^{\circ}, Z_m^{\circ})$, and Stage 2, as described above, is used to estimate $\rho_{\varepsilon jm}$, the correlation between Z_j° and Z_m° . The MMLEs $\hat{\mu}_j, \hat{\mu}_m, \hat{\sigma}_{\varepsilon j}^2, \hat{\sigma}_{\varepsilon m}^2$ and MPPL $\hat{\rho}_{\varepsilon jm}$ are consistent under the assumption of pairwise bivariate normality (6). This assumption may be reasonable in many settings, but should be seriously scrutinized in each problem. As a cautionary tale, consider two types of workers at a plant, those involved in the manufacturing process with high exposure levels to a toxicant and other workers with far less exposure. Blood toxicant levels, i.e. $(Z^{\circ} | \text{group} = g)$, may follow a normal distribution for each group but with different means and variances. Ignoring the existence of the binary covariate and combining the two groups into one results in a mixture of normals rather than in a normal, violating the assumption in (6). The correct approach involves model (1) with U covariates describing the exposure patterns.

2.3 | Estimation for linear functions of left-censored variates

Under the assumptions of multivariate normal errors and no unmeasured confounding in model (1), we have successfully estimated the marginal means $\hat{\mu}_{D^{\circ}}$ and covariance matrix $\hat{\Sigma}_{D^{\circ}}$ of all variates as described in the previous section. We can achieve substantially greater generality by considering linear combinations of the uncensored and censored variates, for which we will estimate means as well as covariance and correlation matrices. This will extend the methodology beyond typical censored-data methods by allowing us, for example, to compute summary statistics for possibly weighted averages and differences of left-censored variates.

Again, $D^{\circ} = (U', \mathbf{Z}^{0'})'$ consists of the q -vector $\mathbf{U}$, not subject to left censoring, and the p -vector $\mathbf{Z}^{\circ}$, which is the uncensored version of $\mathbf{Z}$. Let $\mathbf{G}$ be a $(k+1) \times (q+p)$ matrix of constants. The mean and covariance matrix of the linear combination $\mathbf{GD}^{\circ}$ are estimated by

$$\hat{E}(\mathbf{GD}^\circ) = \mathbf{G}\hat{\mu}_{D^\circ}, \quad \widehat{\text{Cov}}(\mathbf{GD}^\circ) = \mathbf{G}\hat{\Sigma}_{D^\circ}\mathbf{G}' \quad (7)$$

where $\hat{\mu}_{D^\circ}$ and $\hat{\Sigma}_{D^\circ}$ are calculated in the two-stage process described earlier.

3 | METHODS FOR LINEAR REGRESSION

Using the covariance matrix and mean vector from the previous section, Section 3.1 gives straightforward formulae for estimation of the regression coefficients and population multiple R^2 . The nonparametric bootstrap is proposed for standard errors. Section 3.2 covers the typical setting in which all left-censored Z° variates are generated from the uncensored variates, as in (1), and then one of them is chosen to be the outcome Y° to be regressed on the rest. The regression framework (1) for censored variates may not seem reasonable when some censored covariates are generated from other censored variates. Section 3.3 addresses this problem via a recursive system of regression models which are then shown to reduce to the original equation (1) framework, and hence allow estimation and hypothesis testing.

3.1 | Multiple Linear Regression

Our regression setup will allow both the outcome and covariates to be linear combinations of uncensored and left-censored covariates as defined in Section 2.3. Specifically, our goal is to estimate the regression coefficients of the multiple linear regression model

$$Y_i^\circ = \beta_0 + \beta_1 X_{i1}^\circ + \cdots + \beta_k X_{ik}^\circ + e_i \quad (8)$$

where both Y° and $(X_1^\circ, \dots, X_k^\circ)$ can be linear combinations of uncensored variates and variates that are subject to left censoring; $e_1, \dots, e_n$ have mean zero and constant variance σ_e^2 , and are independent of each other and of the covariates. Let $\mathbf{G}$ be a $(k+1) \times (q+p)$ matrix of constants, where $\mathbf{G}_Y$ is the first row and $\mathbf{G}_X$ is the remaining k rows of $\mathbf{G}$. Construct $\mathbf{G}$ so that

$$Y^\circ = \mathbf{G}_Y \mathbf{D}^\circ, \quad \mathbf{X}^\circ = \mathbf{G}_X \mathbf{D}^\circ.$$

By the previous section, the means and covariances of the Y° and $\mathbf{X}^\circ$ can be estimated by

$$\begin{aligned} \hat{\mu}_{Y^\circ} &= \mathbf{G}_Y \hat{\mu}_{D^\circ}, & \hat{\Sigma}_{Y^\circ Y^\circ} &= \mathbf{G}_Y \hat{\Sigma}_{D^\circ} \mathbf{G}_Y', \\ \hat{\mu}_{X^\circ} &= \mathbf{G}_X \hat{\mu}_{D^\circ}, & \hat{\Sigma}_{X^\circ X^\circ} &= \mathbf{G}_X \hat{\Sigma}_{D^\circ} \mathbf{G}_X', \\ \hat{\Sigma}_{X^\circ Y^\circ} &= \mathbf{G}_X \hat{\Sigma}_{D^\circ} \mathbf{G}_Y'. \end{aligned}$$

Minimization of the mean squared error $E(Y^\circ - \beta_0 - \beta'X^\circ)^2$ for an arbitrary linear function leads to $\Sigma_{X^\circ X^\circ} \beta = \Sigma_{X^\circ Y^\circ}$ and $\beta_0 = EY^\circ - \beta'EX^\circ$. Our proposal is to use the substitution or plug-in principle, so that the MPPLEs of the regression coefficients are

$$\hat{\beta} = \hat{\Sigma}_{X^\circ X^\circ}^{-1} \hat{\Sigma}_{X^\circ Y^\circ}, \quad \hat{\beta}_0 = \hat{\mu}_{Y^\circ} - \hat{\beta}' \hat{\mu}_{X^\circ}. \quad (9)$$

The nonparametric bootstrap is proposed for standard errors and Wald-type confidence intervals. In addition, one can estimate the multiple correlation coefficient

$$R^2 = \Sigma'_{X^\circ Y^\circ} \Sigma_{X^\circ X^\circ}^{-1} \Sigma_{X^\circ Y^\circ} / \Sigma_{Y^\circ Y^\circ} \quad (10)$$

by replacing its population components by their MPPLEs. Note that this R^2 is a population quantity, and care should be given to interpreting $\hat{R}^2$ as the percentage of the variability of a potentially censored Y explained by the regression function of the observed X 's, also subject to censoring, when using MPPLEs of the regression coefficients for a given data set.

3.2 | Typical Linear Regression Example

A typical environmental health problem often involves uncensored variates $(U_1, \dots, U_q)$ and contaminants $(Z_1^\circ, \dots, Z_p^\circ)$ subject to left censoring for which the linear system of equations (1) holds along with the associated assumptions. The $\varepsilon = (\varepsilon_1, \dots, \varepsilon_p)'$ error terms in (1) are allowed to share unmeasured factors that induce correlations among the Z° variates. In this section we shall assume that no Z° directly generates any other Z° (e.g. no metabolites) so that model (1) is directly applicable. We shall relax this assumption in the next section. Suppose that the scientific problem of interest is how one of the Z° 's, say $Y^\circ = Z_p^\circ$, is related to the others by a single regression model

$$Y_i^\circ = \beta_0 + \beta_1' U_i + \beta_2' Z_i^{*\circ} + e_i \quad (11)$$

where $Z_i^{*\circ} = (Z_1^\circ, \dots, Z_{p-1}^\circ)'$. In this section we explore the parametric connection between (1) and (11). Although somewhat terse, matrix algebra is useful for simplified expressions and derivations. Model (1) can be abbreviated to $Z_i^\circ = \phi_0 + \phi U_i + \varepsilon_i$, where $\phi_0 = (\phi_{10}, \dots, \phi_{p0})'$ and $\phi' = (\phi_1, \dots, \phi_p)$, the latter being a $q \times p$ matrix. The major blocks of the $(q+p) \times (q+p)$ covariance matrix Σ_{D° of $D^\circ = (U', Z'^\circ)'$ can easily be shown to be Σ_{UU} , $\Sigma_{UZ^\circ} = \Sigma_{UU} \phi'$ and $\Sigma_{Z^\circ Z^\circ} = \phi \Sigma_{UU} \phi' + \Sigma_\varepsilon$. Denote the vector of predictors in (11) by $X^\circ = (U', Z^{*\circ})'$. Let $\mathbf{G}$ be the identity matrix of dimension $q+p$, $\mathbf{G}_Y$ is the last row of $\mathbf{G}$, and $\mathbf{G}_X$ is the first $q+p-1$ rows of $\mathbf{G}$, so that $Y^\circ = \mathbf{G}_Y D^\circ$ and $X^\circ = \mathbf{G}_X D^\circ$. Then $\Sigma_{X^\circ X^\circ} = \mathbf{G}_X \Sigma_{D^\circ} \mathbf{G}_X'$ and $\Sigma_{X^\circ Y^\circ} = \mathbf{G}_X \Sigma_{D^\circ} \mathbf{G}_Y'$, so that $\beta = \Sigma_{X^\circ X^\circ}^{-1} \Sigma_{X^\circ Y^\circ}$ and

$\beta_0 = EY^\circ - \beta' EX^\circ$. Equivalently, one could write

$\text{Cov}(\mathbf{X}^\circ, e) = \text{Cov}(\mathbf{X}^\circ, Y^\circ - \beta' \mathbf{X}^\circ) = \Sigma_{X^\circ Y^\circ} - \Sigma_{X^\circ X^\circ} \beta = 0$. Hence, the $\hat{\beta}$ estimators are functions of $\hat{\phi}$, $\hat{\Sigma}_{UU}$ and $\hat{\Sigma}_e$, i.e. estimators of the parameters underlying model system (1).

A different approach, that can offer valuable insight, is to assume the first $p - 1$ models of (1) for $(Z_1^\circ, \dots, Z_{p-1}^\circ)$ plus model (11) for Y° and then to derive how they determine the ϕ parameters for $Z_p^\circ = Y^\circ$, i.e. the parameters of the last equation in (1). Since

$Z_{ij}^\circ = \phi_{j0} + \phi_j' \mathbf{U}_i + \varepsilon_{ij}$ by model (1), substitution into (11) gives

$$\begin{aligned} Y_i^\circ &= \beta_0 + \beta_1' \mathbf{U}_i + \sum_{j=1}^{p-1} \beta_{2j} (\phi_{j0} + \phi_j' \mathbf{U}_i + \varepsilon_{ij}) + e_i = \left(\beta_0 + \sum_{j=1}^{p-1} \beta_{2j} \phi_{j0} \right) + \left(\beta_1 + \sum_{j=1}^{p-1} \beta_{2j} \phi_j \right)' \mathbf{U}_i \\ &+ \left(\sum_{j=1}^{p-1} \beta_{2j} \varepsilon_{ij} + e_i \right) = \phi_{Y0} + \phi_Y' \mathbf{U}_i + e_{Yi}, \end{aligned}$$

with the obvious equivalence of (ϕ_{Y0}, ϕ_Y) to the parameters in the last line of (1) and with

$e_{Yi} = \varepsilon_{ip}$. Furthermore, $e_i = \varepsilon_{ip} - \sum_{j=1}^{p-1} \beta_{2j} \varepsilon_{ij}$. Letting

$$\mathbf{c} = (-\beta_{21}, \dots, -\beta_{2,p-1}, 1), e \sim N(0, \mathbf{c} \Sigma_e \mathbf{c}')$$

3.3 | Recursive system of regression models

In some settings the model (1) assumption that the Z° variates are generated solely from a set of uncensored U variates may not seem realistic. For example, in drawing a causal graph for a specific real-world problem, some variates with lower limits of quantification may be generated from other variates having a lower limit of quantification. As a simple example, consider the recursive system of models

$$Z_{i1}^\circ = \gamma_{10} + \gamma_1' \mathbf{U}_i + \varepsilon_{i1}^* \quad (12)$$

$$Z_{i2}^\circ = \gamma_{20} + \gamma_2' \mathbf{U}_i + \psi_{21} Z_{i1}^\circ + \varepsilon_{i2}^* \quad (13)$$

$$Z_{i3}^\circ = \gamma_{30} + \gamma_3' \mathbf{U}_i + \psi_{31} Z_{i1}^\circ + \psi_{32} Z_{i2}^\circ + \varepsilon_{i3}^* \quad (14)$$

where $(\varepsilon_{i1}^*, \varepsilon_{i2}^*, \varepsilon_{i3}^*)$ is multivariate normal with mean $\mathbf{0}$ and covariance Σ_{e^*} . As an example, the Z_{ij}° 's might be the true contaminant levels in years $j = 1, 2, 3$ of a study. We make the usual assumption of no unmeasured confounders, whereby in each model the U covariates and the Z° 's, used as covariates, are uncorrelated with the regression error. This assumption is required for unbiased estimation of the regression parameters when all Z° covariates are

measured exactly and is necessary here for the censored-data setting as well. Since $\text{cov}(Z_1^\circ, \varepsilon_2^*) = \gamma_1' \text{cov}(\mathbf{U}, \varepsilon_2^*) + \text{cov}(\varepsilon_1^*, \varepsilon_2^*)$, the requirement that ε_2^* be uncorrelated with Z° and $\mathbf{U}$ in (13) means ε_1^* and ε_2^* must be uncorrelated. Similarly, one can show that if ε_3^* is uncorrelated with $(\mathbf{U}, Z_1^\circ, Z_2^\circ)$ in model (14), then ε_3^* is uncorrelated with ε_1^* and ε_2^* . Hence, Σ_{ε^*} is a diagonal matrix, as generally assumed for recursive systems (Johnston, 1972) — see exception below.

The recursive system of equations described in (12)–(14) can be rewritten in the form of (1), that only involves $\mathbf{U}$ covariates, by making sequential substitutions of the Z_{ij}° outcome from one model into the Z_{ij}° covariates of succeeding models. This results in the following model (1) parameters: $(\phi_{10}, \phi_1) = (\gamma_{10}, \gamma_1)$ for model (12), $(\phi_{20}, \phi_2) = (\gamma_{20} + \psi_{21}\gamma_{10}, \gamma_2 + \psi_{21}\gamma_1)$ for model (13), and $(\phi_{30}, \phi_3) = (\gamma_{30} + \psi_{31}\gamma_{10} + \psi_{32}\phi_{20}, \gamma_3 + \psi_{31}\gamma_1 + \psi_{32}\phi_2)$ for model (14), with model (1) error terms becoming

$$\begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ \psi_{21} & 1 & 0 \\ \psi_{31} + \psi_{32}\psi_{21} & \psi_{32} & 1 \end{pmatrix} \begin{pmatrix} \varepsilon_{i1}^* \\ \varepsilon_{i2}^* \\ \varepsilon_{i3}^* \end{pmatrix}. \quad (15)$$

Writing (15) succinctly as $\varepsilon = \mathbf{B}\varepsilon^*$, the model (1) error terms are multivariate normal with mean $\mathbf{0}$ and covariance $\mathbf{B}\Sigma_{\varepsilon^*}\mathbf{B}'$. And very importantly, ε is uncorrelated with $\mathbf{U}$. The exception to Σ_{ε^*} being a diagonal matrix occurs as follows. If $\psi_{21} = 0$, then $\varepsilon_{i2}^* = \varepsilon_{i2}$ and $\text{Cov}(\varepsilon_{i1}, \varepsilon_{i2})$ need not be zero. Section 4.3 gives an example of a more general semi-recursive system in which (12) and (13) each contain multiple models rather than just a single model.

The important result here is that the recursive system of models (12)–(14) falls into the framework of model (1) for which the assumption of multivariate normal errors is satisfied. Hence, the means and covariance matrices of $(\mathbf{U}, \mathbf{Z}^\circ)$ can therefore be estimated by the two-stage process described in Section 2. Estimation of regression parameters is as follows. For model (13) let $Y^\circ = Z_{i2}^\circ$ represent the outcome and $\mathbf{X}_i^\circ = (\mathbf{U}_i', Z_{i1}^\circ)'$ be the covariates with regression parameters $\beta_0 = \gamma_{20}$ and $\beta = (\gamma_2', \psi_{21})'$. For model (14) let

$Y_i^\circ = Z_{i3}^\circ$ and $\mathbf{X}_i^\circ = (\mathbf{U}_i', Z_{i1}^\circ, Z_{i2}^\circ)'$ with intercept $\beta_0 = \gamma_{30}$ and regression coefficients

$\beta = (\gamma_3', \psi_{31}, \psi_{32})'$. In either case let $\Sigma_{\mathbf{X}^\circ\mathbf{X}^\circ}$ be the covariance matrix of $\mathbf{X}^\circ$ and $\Sigma_{\mathbf{X}^\circ\mathbf{Y}^\circ}$ be the vector of covariances between $\mathbf{X}^\circ$ and $\mathbf{Y}^\circ$. These quantities are estimable via (2)–(4) since the recursive system of equations falls into the framework of model (1). Hence, under the usual assumption of no unmeasured confounders, the consistent estimators are again given by $\hat{\beta} = \hat{\Sigma}_{\mathbf{X}^\circ\mathbf{X}^\circ}^{-1} \hat{\Sigma}_{\mathbf{X}^\circ\mathbf{Y}^\circ}^{-1}$, $\hat{\beta}_0 = \hat{E}Y^\circ - \hat{\beta}'\hat{E}\mathbf{X}^\circ$. This estimation process is performed successively down the list of recursive equation models.

In some settings the data analyst may be provided a single regression model with multiple left-censored covariates. The obvious solution is to add model equations, one for each left-censored covariate. For example, if presented with just model (14), models (12) and (13) can be added, with or without ψ_{21} depending on subject matter knowledge. The tools of this section are then available.

4 | SPECIFIC DESIGNS AND SIMULATION STUDIES

In this section we describe the details of the proposed methods for specific designs, and then report on simulation experiments, comparing the new estimators to both the true values and the estimators obtained from analysis of the uncensored version of the data. In Section 4.1 we examine first how well the MMLEs of means and variances and the MPPLEs of correlations from Section 2.1 work and then consider a standard regression setting which parallels Section 3.2. For comparison, we consider the LOQ/2 and LOQ/ $\sqrt{2}$ imputation methods, given the lack of other methods that can handle the general settings under study. As an aside we assume the data have been pre-transformed to normal, hence appropriate replacement of a left-censored value by, say, LOQ/2 is done on the lognormal scale. In Section 4.2 we consider a regression model in which the outcome is the difference in two left-censored variates and the covariates consist of uncensored and left-censored variates. Section 4.3 studies a model in which the outcome is the average of two left-censored variates and the covariates consist of uncensored variates and the average of three left-censored variates. Given the poor performance of the imputation methods in Section 4.1, they are not reported in the later sections.

4.1 | Standard regression design

In the first simulation experiment suppose a study measures 7 variates on $n = 100$ subjects. There are 2 uncensored variates: $U_1 \sim \text{Bin}(1, 0.5)$ and U_2 is the normal mixture $U_1 + \delta$, where $\delta \sim N(0, 4)$. There are also 5 left-censored variates ($Z_1^\circ, \dots, Z_5^\circ$) that satisfy model equations (1) with

$$\phi = \begin{pmatrix} 0.0 & 0.5 & 1.0 \\ 0.5 & 0.3 & 1.0 \\ -0.5 & -0.5 & -1.0 \\ 0.3 & 0.0 & -1.0 \\ 0.2 & 0.4 & -0.6 \end{pmatrix} \text{ and } \Sigma_\epsilon = \begin{pmatrix} 1.0 & 0.2 & 0.2 & 0.3 & 0.5 \\ 0.2 & 1.0 & 0.4 & 0.3 & 0.4 \\ 0.2 & 0.4 & 1.0 & 0.5 & 0.3 \\ 0.3 & 0.3 & 0.5 & 1.0 & 0.2 \\ 0.5 & 0.4 & 0.3 & 0.2 & 1.0 \end{pmatrix}.$$

Furthermore, the percentages of left-censoring are 20%, 30%, 40%, 50% and 35%. The first task is to examine how well the proposed MMLEs/MPPLEs estimate the true means, variances and correlations. For comparison we include estimators based on the uncensored version of the data and the imputed versions based on LOQ/2 and LOQ/ $\sqrt{2}$ approaches. Table 1 summarizes the results over 1000 simulated data sets. For the sake of space only the correlations between Z_5° and the other variates are shown. In all cases the MMLEs/MPPLEs closely match the true values and uncensored data estimates of the means, variances and correlations. The imputation methods overestimate the means and under-estimate the variances. Their estimates of the correlations are biased towards zero as one might expect in

a measurement error case. The second task is to examine estimates of the regression coefficients. Here, $Y^\circ = Z_5^\circ$ is regressed on the other variates using the same data sets. The true values are found as described in Section 3.2. The second panel of Table 1 summarizes the averages and standard deviations of the regression coefficients, and their average standard errors. Fifty bootstrap resamples are used to estimate the standard errors and Wald-type confidence intervals for the proposed MPPLs. The MPPL estimates of the regression coefficients are excellent, especially when one considers the heavy censoring of the covariates and outcome. Estimation of the population R^2 is also very good. The bootstrap standard errors tend to overestimate the simulation standard error, but the confidence coverages range from 95.8% to 97.5%. The imputation estimators are uniformly dismal with a substantial underestimate of R^2 .

4.2 | Regression for a post-minus-pre design

In this problem we consider a difference between two variates subject to left censoring as the outcome variable which is regressed on uncensored and potentially censored variates. A typical application is the post-pre outcome regression model

$$Y_i^\circ = Z_{i2}^\circ - Z_{i1}^\circ = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + e_i \quad (16)$$

where $X_{i1} = U_{i1}$, $X_{i2} = U_{i2}$, and $X_{i3} = Z_{i1}^\circ$. Furthermore, we suppose that a causal diagram implies the following set of regression equations for the variates subject to left censoring

$$Z_{i1}^\circ = \gamma_{10} + \gamma_{11} U_{i1} + \gamma_{12} U_{i2} + \varepsilon_{i1}^* \quad (17)$$

$$Z_{i2}^\circ = \gamma_{20} + \gamma_{21} U_{i1} + \gamma_{22} U_{i2} + \psi Z_{i1}^\circ + \varepsilon_{i2}^* \quad (18)$$

where $(\varepsilon_{i1}^*, \varepsilon_{i2}^*)$ are independent normals and uncorrelated with (U_{i1}, U_{i2}) . As an example, Z_{i1}° and Z_{i2}° might be the true blood levels of some toxicant in years 1 and 2, respectively, while U_{i1} and U_{i2} might be the subject's gender and age. Substitution of (17) into (18) yields

$$Z_{i2}^\circ = (\gamma_{20} + \psi\gamma_{10}) + (\gamma_{21} + \psi\gamma_{11})U_{i1} + (\gamma_{22} + \psi\gamma_{12})U_{i2} + (\psi\varepsilon_{i1}^* + \varepsilon_{i2}^*).$$

We can now rewrite (17) and (18) into a form equivalent to (1)

$$Z_{i1}^\circ = \phi_{10} + \phi_{11} U_{i1} + \phi_{12} U_{i2} + \varepsilon_{i1} \quad (19)$$

$$Z_{i2}^{\circ} = \phi_{20} + \phi_{21}U_{i1} + \phi_{22}U_{i2} + \varepsilon_{i2} \quad (20)$$

where $(\phi_{10}, \phi_{11}, \phi_{12}) = (\gamma_{10}, \gamma_{11}, \gamma_{12})$ and $(\phi_{20}, \phi_{21}, \phi_{22}) = (\gamma_{20} + \psi\gamma_{10}, \gamma_{21} + \psi\gamma_{11}, \gamma_{22} + \psi\gamma_{12})$ and

$$\begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \psi & 1 \end{pmatrix} \begin{pmatrix} \varepsilon_{i1}^* \\ \varepsilon_{i2}^* \end{pmatrix}, \quad (21)$$

which can be abbreviated as $\varepsilon_i = \mathbf{Q}\varepsilon_i^*$. To get unbiased estimators, we require ε_{i1}^* to be uncorrelated with (U_{i1}, U_{i2}) and ε_{i2}^* uncorrelated with $(U_{i1}, U_{i2}, Z_{i1}^{\circ}, \varepsilon_{i1}^*)$. Since

$$\text{Cov}(Z_1^{\circ}, \varepsilon_2^*) = \text{Cov}(\gamma_{11}U_1 + \gamma_{12}U_2 + \varepsilon_1^*, \varepsilon_2^*) = (\gamma_{11}, \gamma_{12})\text{Cov}((U_1, U_2)', \varepsilon_2^*) + \text{Cov}(\varepsilon_1^*, \varepsilon_2^*), \text{ if and } \text{Cov}(Z_1^{\circ}, \varepsilon_2^*) = 0$$

only if $(U_1, U_2, \varepsilon_1^*)$ are uncorrelated with ε_2^* . Hence, by equation (21), $\varepsilon \sim N_2(0, \mathbf{Q}\Sigma_{\varepsilon^*}\mathbf{Q}')$.

Assumptions (A1)-(A3) are therefore satisfied, which implies that the means and covariance matrix Σ_{D0} of $(U_1, U_2, Z_1^{\circ}, Z_2^{\circ})$ are estimable by Stages 1 and 2 given in Section 2.1. Next, define

$$\mathbf{G} = \begin{pmatrix} 0 & 0 & -1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

The covariance matrix of $(Z_2^{\circ} - Z_1^{\circ}, U_1, U_2, Z_1^{\circ})$ is $\mathbf{G}\hat{\Sigma}_{D0}\mathbf{G}'$. Estimators of the β 's in (16)

follow from Section 3.1. One can also estimate the population R^2 and the multiple partial correlation coefficients. The nonparametric bootstrap is used for estimation of standard errors and Wald-type confidence intervals.

Simulation Study and Results: We generated 1000 data sets, each of size $n = 100$, with $U_1 \sim \text{Bin}(1, 0.5)$, $U_2 \sim \text{Unif}(0, 1)$ and $(\varepsilon_1^*, \varepsilon_2^*) \sim N_2(\mathbf{0}, \Sigma_{\varepsilon^*})$ where the coefficients in (17)–(18) and error covariance matrix are

$$\begin{pmatrix} \gamma_{10} & \gamma_{11} & \gamma_{12} \\ \gamma_{20} & \gamma_{21} & \gamma_{22} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 0.25 & -0.50 & -1.00 \end{pmatrix} \Sigma_{\varepsilon^*} = \begin{pmatrix} 1 & 0 \\ 0 & 0.7 \end{pmatrix}$$

and where $\psi = 0.75$. The percent left censoring for $(Z_1^{\circ}, Z_2^{\circ})$ is (22.0%, 53.6%). As seen in Table 2, the proposed MPPLEs of the β coefficients in regression model (16) have little bias, their bootstrap standard errors are very close to the sample standard deviations of the 1000 estimated β 's and the Wald-type bootstrap confidence coverages are close to the nominal

95% level. Moreover, the average of the 1000 estimates of the population multiple correlation (0.358) for the regression model is very close to the true value (0.323) and that of regression fits based on knowing the true values of the Z° variates (0.343).

4.3 | Regression design based on averages

In this example, suppose that in the previous year, 3 measurements of a toxicant are taken on unit i with true values $(Z_{i1}^\circ, Z_{i2}^\circ, Z_{i3}^\circ)$, and in the current year, 2 more measurements are taken with true values $(Z_{i4}^\circ, Z_{i5}^\circ)$. To make this setting more realistic, we include a covariate U_{i1} that would affect last year's values as well as this year's Z° values. In addition, suppose that a random sample of the units received an intervention or remediation after last year's measurements, i.e. U_{i2} is 1 for the intervention and is zero otherwise. Suppose that the model of interest is

$$Y_i^\circ = \beta_0 + \beta_1 U_{i1} + \beta_2 U_{i2} + \beta_3 X_i^\circ + e_i \quad (22)$$

where $Y_i^\circ = (Z_{i4}^\circ + Z_{i5}^\circ)/2$ and $X_i^\circ = (Z_{i1}^\circ + Z_{i2}^\circ + Z_{i3}^\circ)/3$, the true average values of this year's and last year's toxicant levels, respectively. All Z° variates making up these averages are subject to left censoring.

To design a strategy for fitting model (22), suppose that a causal graph describing how each of last year's variates were generated results in

$$Z_{i1}^\circ = \phi_{10} + \phi_{11} U_{i1} + \varepsilon_{i1}, \quad (23)$$

$$Z_{i2}^\circ = \phi_{20} + \phi_{21} U_{i1} + \varepsilon_{i2}, \quad (24)$$

$$Z_{i3}^\circ = \phi_{30} + \phi_{31} U_{i1} + \varepsilon_{i3}. \quad (25)$$

Keeping in mind that model (22) is our goal, we will model this year's Z° values conditional on last year's values as

$$Z_{i4}^\circ = \eta_{40} + \eta_{41} U_{i1} + \eta_{42} U_{i2} + \eta_{43} \left(\frac{Z_{i1}^\circ + Z_{i2}^\circ + Z_{i3}^\circ}{3} \right) + \varepsilon_{i4}^*, \quad (26)$$

$$Z_{i5}^\circ = \eta_{50} + \eta_{51} U_{i1} + \eta_{52} U_{i2} + \eta_{53} \left(\frac{Z_{i1}^\circ + Z_{i2}^\circ + Z_{i3}^\circ}{3} \right) + \varepsilon_{i5}^*, \quad (27)$$

where $(\varepsilon_{i4}^*, \varepsilon_{i5}^*)$ are bivariate normal. Note that U_1 is still allowed to have an effect on this year's measurements beyond what is contained in last year's average X° . As generally required for validity of regression models, U_1 is assumed independent of the error terms $(\varepsilon_1, \varepsilon_2, \varepsilon_3)$ in models (23)-(25) and (U_1, U_2, X°) are independent of $(\varepsilon_4^*, \varepsilon_5^*)$ in models (26)-(27). Models (23)-(25) are allowed to contain the same unmeasured covariates, so that the errors $(\varepsilon_1, \varepsilon_2, \varepsilon_3)$ are allowed to be correlated. These unmeasured covariates are contained in $(Z_1^\circ, Z_2^\circ, Z_3^\circ)$ so that with X° in models (26) and (27), we assume that $(\varepsilon_1, \varepsilon_2, \varepsilon_3)$ are uncorrelated with $(\varepsilon_4^*, \varepsilon_5^*)$. This is the analogue of the zero correlation assumption made in recursive models (Section 3.3). Finally, $(\varepsilon_4^*, \varepsilon_5^*)$ are allowed to be correlated. Note that (23)-(27) are an extension of the recursive system of equations of Section 3.3 by allowing a set of covariates to be added to the previous equation(s) rather than only adding one covariate at a time.

Before we can estimate the parameters of (22), we need to reformulate models (26) and (27) into model framework (1). The $(Z_1^\circ, Z_2^\circ, Z_3^\circ)$ terms in models (26) and (27) are replaced by their respective models (23)-(25). For $j = 4, 5$ this results in

$$\begin{aligned} Z_{ij}^\circ &= \left(\eta_{j0} + \eta_{j3}(1/3) \sum_1^3 \phi_{\ell 0} \right) + \left(\eta_{j1} + \eta_{j3}(1/3) \sum_1^3 \phi_{\ell 1} \right) U_{i1} + \eta_{j2} U_{i2} + \left(\eta_{j3}(1/3) \sum_1^3 \varepsilon_{i\ell} + \varepsilon_{ij}^* \right) \\ &= \phi_{j0} + \phi_{j1} U_{i1} + \phi_{j2} U_{i2} + \varepsilon_{ij} \end{aligned}$$

(28)

where $(\phi_{j0}, \phi_{j1}, \phi_{j2}, \varepsilon_{ij})$ are implicitly defined in (28). Models (23)-(25) and (28) with $j = 4, 5$ have now been reconfigured in the form of (1). Next, note that the corresponding errors are

$$\begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \\ \varepsilon_{i4} \\ \varepsilon_{i5} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \eta_{43}/3 & \eta_{43}/3 & \eta_{43}/3 & 1 & 0 \\ \eta_{53}/3 & \eta_{53}/3 & \eta_{53}/3 & 0 & 1 \end{pmatrix} \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \\ \varepsilon_{i4}^* \\ \varepsilon_{i5}^* \end{pmatrix} = \mathbf{A} \varepsilon^*,$$

so that the error vector of equation (1) is multivariate normal with mean $\mathbf{0}$ and covariance matrix $\mathbf{A} \Sigma_{\varepsilon^*} \mathbf{A}'$. The error terms $(\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4, \varepsilon_5)$ can be correlated. The only stipulation is that $(\varepsilon_1, \varepsilon_2, \varepsilon_3)$ be independent of $(\varepsilon_4^*, \varepsilon_5^*)$. This generalizes the diagonal error covariance matrix assumption made in recursive systems which add only one covariate at a time.

At this point models (23)-(25) plus (28) written out for $j = 4$ and $j = 5$ represent models for $Z_1^\circ, \dots, Z_5^\circ$ in the framework of (1), in which the error terms $(\varepsilon_1, \dots, \varepsilon_5)$ follow a multivariate normal. We are now justified in using Stage 1 and Stage 2 estimation methods to estimate the mean $\hat{\mu}_{D^\circ}$ and covariance matrix Σ_{D° of $D^\circ = (U_1, U_2, Z_1^\circ, \dots, Z_5^\circ)$. Next, define the G matrix

$$\mathbf{G} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 1/3 & 1/3 & 0 & 0 \end{pmatrix},$$

noting that the first row of $G(G_Y)$ is used to define the outcome $Y^\circ = G_Y \mathbf{D}^\circ$, the average of Z_4° and Z_5° , and the last 3 rows of $G(G_X)$ to define the covariates $(U_1, U_2, \mathbf{X}^\circ) = G_X \mathbf{D}^\circ$. Equations (9) and (10) give the estimated regression coefficients (β, β_0) of (22) and the estimated populations R^2 , respectively. The nonparametric bootstrap can be used for standard errors and constructing confidence intervals.

Simulation Study and Results.—We generated 1000 data sets, each with $n = 100$, in which $U_1 \sim U(0,1)$ and $U_2 \sim \text{Bin}(1, 0.5)$. The true $(Z_1^\circ, Z_2^\circ, Z_3^\circ)$ were generated from (23)-(25) in which their ϕ coefficients were all equal to 1, the error terms were trivariate normal with variances of 1 and covariances each equal to 0.3. In models (26)-(27), $(\eta_{j0}, \eta_{j1}, \eta_{j2}, \eta_{j3}) = (0.2, -0.5, -1.0, 0.7)$, for $j = 4, 5$, and hence so too were the β 's of the regression model (1). The error terms, $(\varepsilon_4, \varepsilon_5)$, were generated from a bivariate normal with variances (0.7, 0.7) and covariance 0.4. The lower limits of quantification were chosen so that each observed Z underwent 40% left censoring. The simulation results (Table 3) illustrate that the MPPLEs of the β 's in (22) are very close to the true values and to the analytic results for the uncensored version of the simulated data. The MPPLE standard errors were based on 50 nonparametric bootstrap resamples of a data set. As seen in Table 3 they are a close match to the empirical standard deviations taken over the 1000 simulated data sets; in addition, the confidence interval coverages are quite good. The population R^2 is precisely estimated.

5 | REGRESSION ANALYSIS OF PCB DATA

Polychlorinated biphenyls (PCBs) are toxic man-made industrial chemicals that persist in the environment and bioaccumulate up the food chain. Due to their chemical stability, electrical insulating properties and low flammability, they were widely manufactured for use in hydraulic fluids, insulating fluids in transformers, fluorescent light ballasts, and plasticizers in paints, plastics and caulk. Although the intentional manufacture was banned in the United States and European Union, they are still released into the environment through disposal, waste burning and atmospheric recycling of PCB-containing products. They are even transported from industrial nations by air and ocean currents to highly sensitive Arctic ecosystems (Andersen et al., 2001). There is an extensive literature on

serious health effects in humans and animals including carcinogenicity (Lauby-Secretan et al., 2013); reproductive abnormalities and fetal toxicity; adverse effects on the immune, nervous and endocrine systems (Brouwer, Reijnders and Koeman, 1989; Henry and DeVito, 2003 and references therein).

In the process of creating PCBs, chlorine atoms are substituted for hydrogen atoms on the biphenyl molecule, a bonded pair of benzene rings. The number and location of the chlorine atoms (up to a maximum of 10) result in 209 different PCB compounds, called congeners, and also determine the physical properties and toxicity. In general, the higher the number of chlorine atoms, the greater the bioaccumulation and the lower rate of metabolism. In the U.S., Monsanto manufactured complex mixtures of PCBs and marketed them under the name Aroclor followed by 4 digits. The last two digits are the percent chlorine by weight.

As part of an on-going U. S. National Institutes of Health-funded study, the AESOP study (Ampelman et al., 2015), blood samples are taken from subjects in East Chicago (urban site) and Columbus Junction, Iowa (rural site). Concentrations of congener-specific levels are determined by gas chromatography with tandem mass spectrometry. Of the 209 congeners, 159 chromatographic peaks can be measured: some are individual congeners, e.g PCB 118, and others co-eluting, e.g. PCBs 99 and 83. Of the 159 congeners measured, 10 have been chosen for this analytical example. All have at least 35% of their values uncensored. As PCBs are lipophilic, these measurements are standardized as the ratio of nanograms (ng) of congener per milliliter (ml) of blood lipids.

Two issues arise with the measurements. The first issue is that the amount of congener, the ratio numerator, may fall below the laboratory LOQ, whereas the amount of blood lipids, the ratio denominator, is measured exactly. The left-censoring point of the ratio differs therefore from one subject to another based on different lipid levels. The second issue is the normality assumption. The ratio data are highly skewed, and hence the log transform has been conventionally used in this field. For this analytical data set only one subject stands out in the normal probability plots as an outlier and has been removed. She has extremely high concentrations of most congeners relative to all subjects, even though her blood lipid level is typical.

Year 2 data from the AESOP study were previously analyzed by Jones, Perry & Thorne (2015). We shall use a subset of the year 2 data consisting solely of the 92 adult women on study. Table 4 provides the names of the 10 PCB congener groupings, referred to as peaks, and the percentage of samples that are above the LOQ (uncensored). Two uncensored variables will be incorporated into all regression models: site, to measure urban vs. rural difference, and age, since older subjects should have greater cumulative PCB exposure. Table 4 lists the MMLEs of the mean and standard deviation of each peak, found separately by use of regression model (1) with peak as the outcome, and site and age as covariates. There is considerable variability among congeners as to the amounts in blood samples.

To illustrate a left-censored outcome regressed on uncensored and left-censored covariates, PCB 118 (Peak D) is chosen as the outcome variable of interest. This particularly toxic dioxin-like congener displays clear carcinogenic activity in rats, specifically, liver cancers,

as well as non-neoplastic lesions in the liver, lung, adrenal cortex, pancreas, thyroid, nose and kidney (National Toxicology Program Technical Report, 2010). Table 4 lists the MPPL estimates of the regression coefficients, bootstrap standard errors and Wald ratios for 9 separate regressions predicting PCB 118 using site, age and a single peak as the 3 covariates. The coefficient estimates for site and age are not given. In addition, Table 4 lists the MPPLEs of the population R^2 and partial correlation between PCB 118 and each peak given site and age. One of the most striking results is that Peak B has a slightly larger Wald statistic (7.368) than Peak C (7.192), but its partial correlation of 0.665 is much smaller than Peak C's partial correlation of 0.890. This is also evident when comparing the estimates of the population $\hat{R}^2$ values: 0.485 for Peak B and 0.808 for Peak C. This Wald ratio-partial correlation paradox is due in part to the fact that Peak B has 90.3% of its values uncensored as compared to just 37.6% uncensored values for Peak C. We return to this issue later.

As a further illustration of the proposed multiple linear regression technique using both uncensored and left-censored covariates, we consider multivariable variable selection next. There are several model building strategies, the most popular ones based on significance tests, information criteria and penalized likelihoods. Forward stepwise selection will be exhibited here, since the other methods would require new theoretical development, not done here. Step 1 in Table 5 lists the estimated coefficients, standard errors and Wald statistics for the base model containing site and age. Since Peak B has the largest Wald value (Table 4), it is the first variable added to the base model and is summarized in Step 2 of Table 5. In the next step Peak C has the largest Wald ratio of any peak, and the resulting model is summarized in Step 3. Note that the addition of Peak C has caused the Peak B Wald statistic to not only lose significance but to even change sign. Hence, in the backward glance of the forward stepwise procedure, Peak B is dropped resulting in the model given in Step 4 of Table 5. In the next step Peak I with a Wald statistic of 2.618 is added (Step 5 summary). Note that this addition has almost no effect on the coefficient or standard error of Peak C. The partial correlation between Peak D (PCB 118) and Peak I (PCB 187) given site, age and Peak C (PCB 105) is 0.401. Finally, no other covariate significantly improved on the last model; the highest positive Wald statistic is 0.850 for Peak F and the largest negative Wald statistic is -0.861 for Peak B, which had been dropped in Step 4. Had either of these variates been added, the corresponding estimated population $\hat{R}^2$ s would be 0.850 and 0.853, which are only slight increases over 0.839. Hence, the final model is given in the last column of Table 5. The corresponding 95% bootstrap confidence intervals are $(-0.169, 0.212)$ for Site, $(-0.016, 0.015)$ for Age, $(0.508, 0.896)$ for P105, $(0.043, 0.299)$ for P187 and $(0.825, 2.058)$ for the intercept.

6 | DISCUSSION

Left-censored linear regression is a very important area of research that is vital to toxicology and environmental epidemiology. In 2010 the journal *Epidemiology* devoted a special issue to left-censored data (Schisterman and Little, 2010), with a call for further development. Early developmental work focused on the setting in which only the outcome variate is subject to left censoring. For general settings it was common in environmental and toxicological research to replace left-censored covariates and outcome by the LOQ/2 or

LOQ/ $\sqrt{2}$, then to log transform the data and use common regression tools, like least squares. This practice is still in common use today. More recent proposals have been to use multiple imputation and the EM algorithm, oftentimes developed for the simpler case of a single left-censored covariate.

The goal of this paper is to address the far more general framework in which both the outcome and multiple covariates are subject to left censoring or are linear combinations of left-censored variates. The proposed approach involves a pseudo-likelihood framework that uses plug-in estimators of lower level parameters in order to estimate higher level parameters. One would expect this approach to lose some efficiency relative to classical maximum likelihood but to have some practical advantages too. Simultaneous maximization of a joint likelihood in one stage is not as practical due to the many different patterns of left-censoring (maximum of 2^p) and the complicated integrations. This issue with the MLE is described in greater detail in Jones et al. (2015, pp. 87–8). That paper also showed that joint maximization of the means, variances and correlation for the case of $p = 2$ is hampered by suboptimal convergence rates, and may have no real improvement in bias or mean squared error. The MPPLE approach will be pursued in subsequent papers for other settings, e.g. the analysis of multivariate data including repeated measures, profiles, and principal components. As far as I am aware, all methods proposed in this paper are new for the field of left-censored data analysis. Since the focus in this paper is on measurement involving a lower LOQ, the methods are developed for left-censored data. However, the likelihoods $\mathcal{L}_j$ in Stage 1 and $\mathcal{L}_{jm}$ in Stage 2 are easily modified to handle right-censored or interval-censored data which are then maximized. Usable software is under development and will become available. The framework and proposed methods are briefly summarized next.

After transformation, usually logarithmic, measures of environmental contaminants are very often normally distributed for a homogeneous population given a common exposure. However, this may not be true for non-homogeneous populations or those with different exposure histories. For this reason, toxicants, like PCBs which persist and bioaccumulate so that blood levels may depend on personal characteristics, e.g. age and residence, exhibit a distribution that is a mixture of normals. Modeling framework (1) addresses this issue, by allowing one to adjust for uncensored covariates. A marginal and pairwise pseudo-likelihood approach is proposed to estimate the means and covariance matrix of all uncensored and left-censored covariates in this framework. This estimation is extended to linear combinations, such as differences and averages of left-censored variates. Methods for linear regression involving an outcome and covariates, which can all be linear combinations of left-censored and uncensored variates, are proposed and are shown to perform quite well via simulation studies. This work is further extended into recursive and semi-recursive systems of equations that extend the conceptual framework of (1). The development of these methods for three common research designs is given and are studied via simulation studies. The results are quite good. The proposed linear regression methodology is illustrated on data from an on-going PCB study which motivated the proposed methods.

ACKNOWLEDGEMENTS

I would like to thank Dr. Peter Thorne for helpful discussions on polychlorinated biphenyls. This work was funded in part by 2 NIH grants: the Iowa Superfund Research Program on Semi-Volatile PCBs (NIH P42 ES013661) and the Environmental Health Sciences Research Center, University of Iowa (NIH P30 ES005605). Finally, I would like to thank the Associate Editor and the Reviewer for their insightful comments.

REFERENCES

- Ampelman MD, Martinez A, DeWall J, Rawn DFK, Hornbuckle KC & Thorne PS (2015). Inhalation and dietary exposure to PCBs in urban and rural cohorts via congener-specific measurements. *Environmental Science & Technology*, 49, 1156–1164. [PubMed: 25510359]
- Andersen M, Lie E, Derocher AE, Belikov SE, Bernhoft A, Boltunov AN, Garner GW, Skaare JU & Wiig O (2001) Geographic variation of PCB congeners in polar bears (*Ursus maritimus*) from Svalbard east to the Chukchi Sea. *Polar Biology*, 24, 231–238.
- Atem FD, Sampene E & Greene TJ (2017). Improved conditional imputation for linear regression with a randomly censored predictor. *Statistical Methods in Medical Research*, Epub 2017 Jan 1. 10.1177/0962280217727033
- Brouwer A, Reijnders PJH & Koeman JH (1989). Polychlorinated biphenyl (PCB) contamin- ated fish induces vitamin A and thyroid hormone deficiency in the common seal. *Aquatic Toxicology*, 15, 99–106.
- EFSA. (2010). Management of left-censored data in dietary exposure assessment of chemical substances. EFSA Journal, 8,1557 Ninety-six page scientific report of the European Food Safety Authority.
- Frame GM (2001). The current state-of-the-art of comprehensive, quantitative, congener specific PCB analysis, and what we know about the distributions of individual congeners in commercial Aroclor mixtures In PCBs: Recent Advances in Environmental Toxicology and Health Effects, edited by Robertson LW & Hansen LG, The University Press of Kentucky.
- Fridley BL & Dixon P (2007). Data augmentation for a Bayesian spatial model involving censored observations. *Environmetrics*, 18, 107–123.
- Gadda HJ, Thiébaud R, Chêne G & Commenges D (2000). Analysis of left-censored longitudinal data with application to viral load in HIV infection. *Biostatistics*, 1, 355–368. [PubMed: 12933561]
- Gong G & Samaniego FJ (1981). Pseudo maximum likelihood estimation: theory and applications. *The Annals of Statistics*, 9, 861–869.
- Hald A (1949). Maximum likelihood estimation of the parameters of a normal distribution which is truncated at a known point. *Skandinavisk Aktuarietidskrift*, 32, 119–134.
- Henry TR & DeVito MJ (2003). Non-Dioxin-like PCBs: Effects and consideration in ecologic risk assessment. United States Environmental Protection Agency Report. <http://www.epa.gov/oswer/riskassessment/pdf/1340-erasc-003.pdf>
- Hewett, P. (2006). Analysis of censored data. Appendix VIII of *A Strategy for Accessing and Managing Occupational Exposures*, (eds.) Bullock, W.H. & Ignacio, J.S. (third edition).
- Hewett P & Ganser GH (2007). A comparison of several methods for analyzing censored data. *The Annals of Occupational Hygiene*, 51, 611–632. [PubMed: 17940277]
- Hoffman HJ & Johnson RE (2015). Pseudo-likelihood estimation of multivariate normal parameters in the presence of left-censored data. *Journal of Agricultural, Biological and Environmental Statistics*, 20, 156–171.
- Holden LR, Graham JA, Whitmore RW, Alexander WJ, Pratt RW, Liddle SK & Piper LL (1992). Results of the National Alachlor Well Water Survey. *Environmetnal Science & Technology*, 26, 935–943.
- Hornung RW & Reed LD (1990). Estimation of average concentration in the presence of nondetectable values. *Applied Occupational and Environmental Hygiene*, 5, 46–51.
- Hughes JP (1999). Mixed Effects Models with Censored Data with Application to HIV RNA Levels. *Biometrics*, 55, 625–629. [PubMed: 11318225]
- Johnston J (1972). *Econometric Methods*. New York: McGraw-Hill.

- Jones MP, Perry SS & Thorne PS (2015). Maximum Pairwise Pseudo-Likelihood Estimation of the Covariance Matrix from Left-Censored Data. *Journal of Agricultural, Biological and Environmental Statistics*, 20, 83–99.
- Jones RR, Weyer PJ, DellaValle CT, Inoue-Choi M, Anderson KE, Cantor KP, Krasner S, Robien K, Freeman LEB, Silverman DT & Ward MH (2016). Nitrate from drinking water and diet and bladder cancer among postmenopausal women in Iowa. *Environmental Health Perspectives*, 124, 1751–1758. [PubMed: 27258851]
- Kalbfleisch J & Prentice R (2002). *The Statistical Analysis of Failure Time Data*. New York: Wiley.
- Korwel IK, Hornbuckle KC, Peck A, Ludewig G, Robertson LW, Sulkowski WW, Espandiani P, Gairola CG & Lehmler HJ (2005). Congener-specific tissue distribution of Aroclor 1254 and a highly chlorinated environmental PCB mixture in rats. *Environmental Science & Technology*, 39, 3513–3520. [PubMed: 15954224]
- Lauby-Secretan B, Loomis D, Grosse Y, El Ghissassi F, Bouvard V, Benbrahim-Tallaa L, Guha N, Baan R, Mattock H & Straif K (2013). Carcinogenicity of polychlorinated biphenyls and polybrominated biphenyls. *The Lancet Oncology*, 14, 287–288. [PubMed: 23499544]
- Marek RF, Thorne PS, Wang K, DeWall J & Hornbuckle KC (2013). PCBs and OH-PCBs in serum from children and mothers in urban and rural U.S. communities. *Environmental Science & Technology* 47, 3353–3361. Erratum in: *Environmental Science & Technology*, 47, 9555–9556. [PubMed: 23452180]
- May RC, Ibrahim JG & Chu H (2010). Maximum likelihood estimation in generalized linear models with multiple covariates subject to detection limits. *Statistics in Medicine*, 30, 2551–2561.
- National Toxicology Program. (2010). Toxicology and carcinogenesis studies of 2,3',4,-4',5-pentachlorobiphenyl (PCB 118) (CAS No. 31508-00-6) in female Harlan Sprague-Dawley rats (gavage studies). National Toxicology Program Technical Report Series, 11; 559, 1–174. (http://ntp.niehs.nih.gov/ntp/htdocs/lt_rpts/tr559.pdf)
- Nehls GJ & Akland GG (1973). Procedures for handling aerometric data. *Journal of the Air Pollution Control Association*, 23, 180–184.
- 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:July Supplement, S17–S24. [PubMed: 21422965]
- Pesonen M, Pesonen J & Nevalainen J (2015). Covariance matrix estimation for left-censored data. *Computational Statistics & Data Analysis*, 92, 13–25.
- R Development Core Team. (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria <http://www.rproject.org/>.
- Richardson DB & Ciampi A (2003). Effects of exposure measurement error when an exposure variable is constrained by a lower limit. *American Journal of Epidemiology*, 157, 355–363. [PubMed: 12578806]
- Schisterman EF & Little RJ (2010). Opening the black box of biomarker measurement error. *Epidemiology*, 21(July Supplement), S1–S3. [PubMed: 20539119]
- Shumway RH, Azari RS & Kayhanian M (2002). Statistical approaches to estimating mean water quality concentrations with detection limits. *Environmental Science & Technology*, 36, 3345–3353. [PubMed: 12188364]
- The Economist.(2017). Pollution in China: The bad earth. 6 10–16, 2017, 24–26.
- Thorne PS, Perry SS, Saito R, O'Shaughnessy PT, Mehaffy J, Metwali N, Keefe T, Donham KJ & Reynolds SJ (2010). Evaluation of the Limulus amebocyte lysate and recombinant factor C assays for assessment of airborne endotoxin. *Applied and Environmental Microbiology*, 76, 4988–4995. [PubMed: 20525858]
- Tobin J (1958). Estimation of relationships for limited dependent variables. *Econometrica*, 26, 24–36.

Table 1

Simulation results for standard regression design. Estimation of summary measures, regression coefficients and standard errors, R^2 . Percent left-censoring for $(U, Z^\circ) : 0, 0, 20, 30, 40, 50, 35; n = 100$ over 1000 simulated data sets.

Means	U ₁	U ₂	Z ₁ ⁰	Z ₂ ⁰	Z ₃ ⁰	Z ₄ ⁰	Z ₅ ⁰	
True	0.50	0.50	0.75	1.15	−1.25	−0.20	0.10	
Uncensored	0.500	0.507	0.759	1.156	−1.258	−0.210	0.097	
MPPLE	0.500	0.507	0.758	1.156	−1.258	−0.217	0.096	
LOQ/2	0.500	0.507	0.882	1.389	−0.860	0.357	0.224	
LOQ/ $\sqrt{2}$	0.500	0.507	0.952	1.493	−0.721	0.530	0.345	
Variances								
True	0.250	4.250	5.562	5.422	5.562	5.250	2.450	
Uncensored	0.250	4.239	5.544	5.409	5.547	5.226	2.448	
MPPLE	0.250	4.239	5.534	5.409	5.536	5.254	2.452	
LOQ/2	0.250	4.239	4.549	3.859	3.266	2.517	1.812	
LOQ/ $\sqrt{2}$	0.250	4.239	4.180	3.437	2.829	2.112	1.495	
Correlations with Z ₅ ⁰								
True	−0.064	−0.759	−0.535	−0.566	0.752	0.739		
Uncensored	−0.061	−0.758	−0.535	−0.565	0.751	0.737		
MPPLE	−0.061	−0.760	−0.539	−0.570	0.752	0.739		
LOQ/2	−0.057	−0.717	−0.485	−0.495	0.710	0.681		
LOQ/ $\sqrt{2}$	−0.056	−0.710	−0.471	−0.477	0.711	0.682		
Coefficients	β_0	β_1	β_2	β_3	β_4	β_5	β_6	R ²
True	0.158	0.167	−1.267	0.442	0.281	0.144	−0.089	0.738
Uncensored Data								
Ave	0.156	0.168	−1.273	0.440	0.284	0.1406	−0.088	0.750
Std Dev	0.146	0.180	0.198	0.086	0.091	0.100	0.102	
Ave(SE)	0.147	0.186	0.199	0.088	0.092	0.101	0.100	
MPPLE								
Ave	0.152	0.169	−1.288	0.443	0.288	0.142	−0.098	0.760
Std Dev	0.198	0.223	0.299	0.114	0.135	0.142	0.151	
Ave(BSE)	0.221	0.251	0.343	0.132	0.154	0.166	0.179	
CI Coverage	95.8	95.8	96.7	96.3	95.8	96.1	97.5	
LOQ/2								
Ave	0.193	0.230	−0.744	0.300	0.150	0.204	−0.018	0.668
StD Dev	0.177	0.172	0.157	0.089	0.095	0.116	0.128	
Ave(SE)	0.169	0.179	0.145	0.089	0.0924	0.109	0.113	
LOQ/ $\sqrt{2}$								
Ave	0.255	0.215	−0.646	0.270	0.136	0.202	−0.011	0.662

Coefficients	β_0	β_1	β_2	β_3	β_4	β_5	β_6	R^2
Std Dev	0.181	0.158	0.145	0.084	0.090	0.119	0.135	
Ave(SE)	0.169	0.164	0.132	0.085	0.089	0.107	0.114	

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2

Simulation results for Post-Pre design. Estimation of regression coefficients, bootstrap standard errors and 95% confidence interval coverages. $n = 100$. Percent left-censoring for Z_1^o, Z_2^o : 22.0%, 53.6%

True β	Method	Ave($\hat{\beta}$)	SD($\hat{\beta}$)	Ave(SE($\hat{\beta}$))	CI Coverage
$\beta_0 = 0.25$	Uncensored	0.259	0.203	0.207	94.9%
	MPPLE	0.253	0.298	0.298	94.7%
$\beta_1 = -0.50$	Uncensored	-0.497	0.190	0.189	93.9%
	MPPLE	-0.500	0.237	0.233	94.8%
$\beta_2 = -1.00$	Uncensored	-1.007	0.304	0.306	95.6%
	MPPLE	-1.019	0.401	0.386	93.9%
$\beta_3 = -0.25$	Uncensored	-0.253	0.084	0.085	94.8%
	MPPLE	-0.249	0.127	0.123	93.0%

Note: Multiple R^2 : 0.323 for true model, 0.343 for uncensored data analysis, 0.358 for MPPLE analysis. Results summarized over 1000 simulated data sets.

Table 3

Simulation results for Design Based on Averages. Estimation for regression coefficients, bootstrap standard errors and 95% confidence interval coverages. $n = 100$. Percent left-censoring for $Z_1^o, \dots, Z_5^o$: 40% each

True β	Method	Ave($\hat{\beta}$)	SD($\hat{\beta}$)	Ave(SE($\hat{\beta}$))	CI Coverage
$\beta_0 = 0.2$	Uncensored	0.200	0.194	0.196	95.7%
	MPPLE	0.197	0.233	0.236	95.3%
$\beta_1 = -0.5$	Uncensored	-0.494	0.295	0.280	94.2%
	MPPLE	-0.492	0.330	0.316	93.4%
$\beta_2 = -1.0$	Uncensored	-0.996	0.152	0.149	94.3%
	MPPLE	-0.999	0.173	0.172	95.0%
$\beta_3 = 0.7$	Uncensored	0.698	0.107	0.103	94.2%
	MPPLE	0.699	0.131	0.129	95.0%

Note: Multiple R^2 : 0.499 for true model, 0.496 for uncensored data analysis, 0.499 for MPPL analysis. Results summarized over 1000 simulated data sets.

Table 4

Analysis of PCB congeners^{*}. In each line the outcome variable P118 is regressed on 3 covariates: Site, Age and Peak.

Peak	PCB Congener Grouping [†]	MMLE of Mean (SD)	Percent Uncensored	Linear Regression Slope (SE _B) Wald ^{**}	R ²	Partial Cor Y,Peak ^{††}
A	P70+74 ⁺⁶¹⁺⁷⁶	1.832 (0.675)	40.86	0.625 (0.145) 4.322	0.324	0.519
B	P99+83	1.369 (0.613)	90.32	0.724 (0.098) 7.368	0.485	0.665
C	P105	0.339 (0.830)	37.63	0.714 (0.099) 7.192	0.808	0.890
D	P118 (Y)	1.808 (0.677)	83.87			
E	P138+129+163	2.150 (0.670)	95.70	0.568 (0.142) 4.006	0.317	0.511
F	P146	0.166 (0.826)	75.27	0.321 (0.114) 2.811	0.191	0.353
G	P153+ ¹⁶⁸	2.275 (0.682)	98.92	0.506 (0.141) 3.600	0.267	0.454
H	P156+ ¹⁵⁷	0.486 (0.786)	58.06	0.532 (0.135) 3.928	0.337	0.531
I	P187	0.790 (0.823)	84.95	0.238 (0.126) 1.890	0.137	0.256
J	P203	-0.246 (0.956)	46.24	0.051 (0.102) 0.501	0.080	0.063

^{*} Log of congener values used to achieve normality in model (1).

[†] For co-eluting congeners, the major congener by Aroclor production is listed first.

Minor congeners are superscripted.

^{**} Coefficient (SE), Wald Statistic shown for Peak but not Intercept, Site and Age.

^{††} Partial correlation between P118 (Y) and Peak are conditional on Site and Age.

Table 5

Forward stepwise regression for the outcome P118 with congeners^{*} from Table 4 considered as potential covariates. Site and age are forced into each model.

Summary ^{**} of Forward Stepwise Regression Modeling Steps					
Term [†] (Peak)	1	2	3	4	5
Intercept	0.572 (0.534)	0.326 (0.358)	1.121 (0.423)	1.068 (0.347)	1.441 (0.314)
	1.071	0.910	2.652	3.077	4.586
Site	−0.019 (0.147)	−0.095 (0.108)	0.041 (0.118)	0.030 (0.101)	0.022 (0.097)
	−0.132	−0.883	0.345	0.300	0.222
Age	0.031 (0.012)	0.014 (0.008)	0.013 (0.008)	0.012 (0.008)	−0.0005 (0.008)
	2.507	1.800	1.582	1.513	−0.059
P99 ⁺⁸³ (B)		0.724 (0.098)	−0.071 (0.211)		
		7.368	−0.338		
P105(C)			0.755 (0.169)	0.714 (0.099)	0.702 (0.099)
			4.471	7.192	7.084
P187(I)					0.171 (0.065)
					2.618
MPPLE- <i>R</i> ²	0.076	0.485	0.809	0.808	0.839

^{*} Log of congener values used to achieve normality in model (1).

^{**} Tabled values are regression coefficient estimate (standard error) and Wald statistic.

[†] For co-eluting congeners, the major congener by Aroclor production is listed first.

Minor congeners are superscripted.