Biometrika (2016), **103**, 1, *pp.* 161–174
Printed in Great Britain

doi: 10.1093/biomet/asv055
Advance Access publication 22 January 2016

Semiparametric approach to regression with a covariate subject to a detection limit

By SHENGCHUN KONG

Department of Statistics, Purdue University, West Lafayette, Indiana 47907, U.S.A
kongsc@purdue.edu

AND BIN NAN

Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, U.S.A
bnan@umich.edu

Summary

We consider generalized linear regression with a covariate left-censored at a lower detection limit. Complete-case analysis, where observations with values below the limit are eliminated, yields valid estimates for regression coefficients but loses efficiency, ad hoc substitution methods are biased, and parametric maximum likelihood estimation relies on parametric models for the unobservable tail probability distribution and may suffer from model misspecification. To obtain robust and more efficient results, we propose a semiparametric likelihood-based approach using an accelerated failure time model for the covariate subject to the detection limit. A two-stage estimation procedure is developed, where the conditional distribution of this covariate given other variables is estimated prior to maximizing the likelihood function. The proposed method outperforms complete-case analysis and substitution methods in simulation studies. Technical conditions for desirable asymptotic properties are provided.

Some key words: Accelerated failure time model; Censored covariate; Empirical process; Generalized linear model; Pseudolikelihood estimation.

1. Introduction

A detection limit is a threshold below which measured values are considered to be not significantly different from background noise and hence unreliable. The presence of measurements below the detection limit is common in applications. For example, in the National Health and Nutrition Examination Survey, many of the exposure variables have measurements below their detection limits; one such variable is urine arsenobetaine level, of which 27·8% of measured values are below the detection limit of 0·4 μg/l (Caldwell et al., 2009). In the Diabetes Prevention Program, 66 of the 301 eligible participants had testosterone levels below the detection limit of 8·0 ng/dl (Kim et al., 2013). In an analysis for the Michigan Bone Health and Metabolism Study, up to 66% of the 50 study participants had anti-Mullerian hormone below the detection limit of 0·05 ng/ml (Sowers et al., 2008). For illustration below we will use the National Health and Nutrition Examination Survey, which examines the relationship between arsenic exposure and the prevalence of type 2 diabetes (Navas-Acien et al., 2008).

A variable with detection limit can be either a response or an explanatory variable. We focus on the latter in this article. Although many ad hoc methods have been implemented, appropriate statistical methods for regression models with such a variable are yet to be thoroughly studied (Schisterman & Little, 2010). Complete-case analysis, where observations with values below the detection limit are simply eliminated, yields consistent estimates of the regression coefficients (Little & Rubin, 2002; Nie et al., 2010) but loses efficiency. It is not well-accepted by practitioners, who are usually reluctant to delete data, especially a lot of data, as would be the case in the aforementioned examples. Substitution methods, where values of the covariate Z that lie below the detection limit L are replaced by L, $L/2^{1/2}$ or zero, are frequently used in epidemiological studies; see, for example, Moulton et al. (2002). These methods are easy to implement but can yield large biases (Helsel, 2006; Nie et al., 2010). Richardson & Ciampi (2003) proposed to replace the values below L with $E(Z \mid Z < L)$, which is obtained from a distribution of Z that is assumed known; however, the distributional assumption is not verifiable, and even if $E(Z \mid Z < L)$ is correctly specified, the method will still lead to biased estimators when Z is correlated with other covariates.

Another widely used method is maximum likelihood estimation based on a parametric distributional assumption on the unobservable tail of Z. For example, Nie et al. (2010) and Arunajadai & Rauh (2012) considered linear regression based on normal and generalized gamma distributions for Z; Cole et al. (2009), Albert et al. (2010) and Wu et al. (2012) considered parametric maximum likelihood under generalized linear regression; and D'Angelo & Weissfeld (2008) applied this approach to Cox regression. In practice, however, the underlying distribution of Z is unknown, and a test of the parametric assumption is usually unavailable because there is no information below the detection limit. Lynn (2001) and Nie et al. (2010) noted that a parametric assumption can yield large bias if misspecified, and they argue that such an approach should not be attempted. Nie et al. (2010), one of the very few groups of researchers who have recognized the danger of substitution or parametric maximum likelihood estimation, recommended complete-case analysis despite its drawbacks.

To obtain more efficient yet robust results, we propose a semiparametric likelihood-based approach to fitting generalized linear models with a covariate subject to a detection limit. The tail distribution of the covariate is estimated from a semiparametric accelerated failure time model, conditional on the fully observed covariates. Model checking can be done using martingale residuals for semiparametric accelerated failure time models. The proposed estimator is shown to be consistent and asymptotically normal, and it outperforms existing methods in simulations. The proof of the asymptotic properties relies heavily on empirical process theory and is provided in the Supplementary Material.

2. A semiparametric approach

For a single observation, denote the response variable by Y, the covariate with a detection limit by Z, and the p fully observed covariates by $X = (X_1, \ldots, X_p)^{\mathrm{T}}$. For simplicity, we consider only one covariate that is subject to a detection limit. Consider a generalized linear model (McCullagh & Nelder, 1989; Agresti, 2002) with

$$E(Y) = \mu = g^{-1}(D^{\mathrm{T}}\theta), \tag{1}$$

where g is a link function and $D^{\mathrm{T}}\theta$ is a linear predictor with $D = (1, X^{\mathrm{T}}, Z)^{\mathrm{T}}$ and $\theta = (\beta^{\mathrm{T}}, \gamma)^{\mathrm{T}}$; here β is a $(p+1)$-dimensional vector and γ is a scalar. The variance of Y, typically a function of the mean, is denoted by $\mathrm{var}(Y) = W(\mu) = W\{g^{-1}(D^{\mathrm{T}}\theta)\}$. We have

$$f_{\varpi,\phi}(Y \mid Z, X) = \exp\left\{\frac{Y\varpi - b(\varpi)}{a(\phi)} + c(Y, \phi)\right\}, \tag{2}$$

where ϕ is the dispersion parameter and ϖ is the natural parameter. Then $\mu = E(Y) = \dot{b}(\varpi)$ and $\text{var}(Y) = \ddot{b}(\varpi)a(\phi)$, where $\dot{b}$ denotes the first derivative and $\ddot{b}$ the second derivative of b.

The value of Z is unobservable when $Z < L$, where the constant L is the detection limit; this is an example of left censoring. In practice Z often measures the concentration of some substance and hence is nonnegative. Consider a monotone decreasing transformation h such that $Z = h(T)$; for example, $h(T) = \exp(-T)$. Write $D(T) = \{1, X^{\mathrm{T}}, h(T)\}^{\mathrm{T}}$. If $T \leqslant C = h^{-1}(L)$, then T is observed; otherwise T is right censored by C. We denote the observed value by $V = \min(T, C)$ and the censoring indicator by $\Delta = I(T \leqslant C)$.

The proposed method works for a broad family of link functions satisfying the regularity conditions given in the Appendix. For notational simplicity, we present the main material using the canonical link function $g = (\dot{b})^{-1}$. Then, when T is observed, model (2) becomes

$$f_{\theta,\phi}(Y \mid T, X) = \exp\left[\frac{YD^{\mathrm{T}}(T)\theta - b\{D^{\mathrm{T}}(T)\theta\}}{a(\phi)} + c(Y, \phi)\right]. \tag{3}$$

Let $F_1(t \mid X)$ denote the conditional cumulative distribution function of T given X, and let its density be $f_1(t \mid X)$. The likelihood function for the observed data (V, Δ, Y, X) can be factorized as

$$f(V, \Delta, Y, X) = f_2(V, \Delta \mid Y, X)\, f_3(Y \mid X)\, f_4(X),$$

where f is the joint density of (V, Δ, Y, X), f_2 is the conditional density of (V, Δ) given (Y, X), f_3 is the conditional density of Y given X, and f_4 is the marginal density of X. Going through conditional arguments using Bayes' rule and dropping $f_4(X)$, we obtain the likelihood function

$$L(V, \Delta, Y, X) = \{f_{\theta,\phi}(Y \mid T, X)\, f_1(T \mid X)\}^{\Delta} \left\{\int_C^{\infty} f_{\theta,\phi}(Y \mid t, X)\, \mathrm{d}F_1(t \mid X)\right\}^{1-\Delta}, \tag{4}$$

where only $f_{\theta,\phi}$ contains the parameter of interest θ, and f_1 is a nuisance parameter in addition to ϕ.

Expression (4) consists of two parts: $\{f_{\theta,\phi}(Y \mid T, X) f_1(T \mid X)\}^{\Delta}$ for fully observed subjects, and $\{\int_C^{\infty} f_{\theta,\phi}(Y \mid t, X)\, \mathrm{d}F_1(t \mid X)\}^{1-\Delta}$ for subjects with covariates below the detection limit. Complete-case analysis is based only on the first part, and although it yields a consistent estimator of θ, it clearly loses efficiency. We see from the second part of (4) that the efficiency gain depends on how well we can recover the right tail of the conditional distribution $F_1(t \mid X)$ beyond C. Parametric models for $F_1(t \mid X)$ are often considered (Nie et al., 2010) but could suffer from model misspecification. The nonparametric method degenerates to the complete-case analysis because there is no observation beyond C. We consider a semiparametric approach that allows reliable extrapolation beyond C and is robust with respect to parametric assumptions.

Of all the common semiparametric models for right-censored data, only the accelerated failure time model allows extrapolation beyond C, and model checking can be done by visualizing the cumulative sums of the martingale-based residuals (Lin et al., 1993, 1996; Peng & Fine, 2006). We therefore propose a semiparametric accelerated failure time model for the transformed covariate subject to the detection limit, given by

$$T = X^{\mathrm{T}}\alpha + \varsigma \tag{5}$$

where ς follows some unknown distribution, denoted by η, and is independent of X. We only consider a fixed transformation h for T in this article, but more flexible transformations, such as the Box–Cox transformation (Box & Cox, 1964; Foster et al., 2001; Cai et al., 2005), are worthy of further investigation. Note that X appears in both (1) and (5), but it may refer to different forms of covariates in these two models. For example, X_1 is a covariate in (1) whereas X_1^2 is a covariate in (5). For notational simplicity we use the same X to represent all fully observed covariates. The loglikelihood function then becomes

$$\begin{aligned}\log L = {} & \Delta \log f_{\theta,\phi}(Y \mid T, X) + \Delta \log \dot{\eta}(T - X^{\mathrm{T}}\alpha) \\ & + (1-\Delta) \log \left\{ \int_{C - X^{\mathrm{T}}\alpha}^{\tau} f_{\theta,\phi}(Y \mid t + X^{\mathrm{T}}\alpha, X)\, \mathrm{d}\eta(t) \right\}, \qquad (6)\end{aligned}$$

where τ is a truncation time at the residual scale defined in Condition A4 in the Appendix.

The existence of a finite constant truncation time τ is a common assumption for censored survival data mainly for technical convenience, particularly for establishing asymptotic normality. Since in practice the residuals are finite for any given dataset and $X^{\mathrm{T}}\alpha$ is bounded, the value of τ need not be specified. We do not specify τ in the numerical examples in § 5.

3. PSEUDOLIKELIHOOD METHOD

The loglikelihood function (6) involves an unknown distribution function η and its derivative, so semiparametric maximum likelihood estimation can be complicated. We propose a two-stage pseudolikelihood approach in which the nuisance parameters (ϕ, α, η) are estimated in stage 1 and then the parameter of interest θ is estimated by maximizing the data version of (6) in stage 2, with nuisance parameters replaced by their estimators from stage 1. The approach is a direct extension of the parametric pseudolikelihood method of Gong & Samaniego (1981) to semiparametric models. Details are given below.

Stage 1. Nuisance parameter estimation. The dispersion parameter ϕ is estimated by complete-case analysis of the generalized linear model (2); the accelerated failure time model regression coefficient α is estimated either by rank-based methods (Wei et al., 1990; Jin et al., 2003; Nan et al., 2009) or by sieve maximum likelihood (Ding & Nan, 2011); and the accelerated failure time model error distribution η is estimated from the censored residuals by the Kaplan–Meier estimator.

Complete-case analysis can be performed using any standard statistical package for generalized linear models. The rank-based estimates for the accelerated failure time model are usually obtained by linear programming. The R (R Development Core Team, 2016) package rankreg (Zhou, 2006), now archived by CRAN, can be implemented for small to moderate sample sizes because it solves a linear programming problem of size greater than n^2. A unique solution can be obtained when Gehan weights are used (Jin et al., 2003). This is the method we have implemented. An alternative approach is to modify the Newton algorithm for solving the discrete rank-based estimating equation (Yu & Nan, 2006). A standard Newton–Raphson algorithm can be used to obtain the sieve maximum likelihood estimates (Ding & Nan, 2011) for the accelerated failure time model when the sample size is large.

Stage 2. Pseudolikelihood estimation of θ. Replacing (ϕ, α, η) by their stage 1 estimates $(\hat{\phi}_n, \hat{\alpha}_n, \hat{\eta}_{n,\hat{\alpha}_n})$ in the loglikelihood function yields the following log-pseudolikelihood function for a random sample of n observations:

$$\mathrm{pl}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left\{\Delta_i \log f_{\theta,\hat{\phi}_n}(Y_i \mid X_i, T_i) \right.$$
$$\left. + (1-\Delta_i)\log \int_{C-X_i^{\mathrm{T}}\hat{\alpha}_n}^{\tau} f_{\theta,\hat{\phi}_n}(Y_i \mid X_i, t + X_i^{\mathrm{T}}\hat{\alpha}_n)\,\mathrm{d}\hat{\eta}_{n,\hat{\alpha}_n}(t)\right\}, \tag{7}$$

where

$$f_{\theta,\hat{\phi}_n}(Y_i \mid T_i, X_i) = \exp\left[\frac{Y_i D_i^{\mathrm{T}}(T_i)\theta - b\{D_i^{\mathrm{T}}(T_i)\theta\}}{a(\hat{\phi}_n)} + c(Y_i, \hat{\phi}_n)\right].$$

The term $\Delta \log \dot{\eta}(T)$ in (6) is dropped because it does not involve θ. We maximize (7) by setting its derivative to zero and then solving the equation to obtain the pseudolikelihood estimator $\hat{\theta}_n$, using a standard Newton–Raphson algorithm with the initial value obtained from the complete-case analysis in stage 1. Because of the nice properties of the initial value, the Newton–Raphson algorithm converges quickly in our numerical examples.

Since $\hat{\theta}_n$ is obtained by solving an estimating equation, its asymptotic properties can be obtained from Z-estimation theory. It can be shown that all the estimators obtained in stage 1 have desirable statistical properties for stage 2 estimation. In particular, $\hat{\phi}_n$ obtained from the complete-case analysis is $n^{1/2}$-consistent (Little & Rubin, 2002), $\hat{\alpha}_n$ is $n^{1/2}$-consistent (Nan et al., 2009; Ding & Nan, 2011), and $\hat{\eta}_{n,\hat{\alpha}_n}$ is also $n^{1/2}$-consistent in a finite interval, a proof of which is given in the Supplementary Material.

Clearly the efficiency gain of the proposed two-stage method comes from the contribution of the term containing integrals in (7), and so depends on the strength of the association between T and X. If the association is very weak, then this term contributes minimally to the estimation of θ. The existence of a finite τ rules out the trivial situation of zero integral value when no association exists, because under Condition A4 in the Appendix all the observations subject to the detection limit are truncated by some $\tau < C$.

4. Asymptotic properties

Define a random map

$$\Psi_{\theta,n}(\phi,\alpha,\eta) = \frac{1}{n}\sum_{i=1}^{n}\psi_\theta(Y_i, X_i, V_i, \Delta_i; \phi, \alpha, \eta), \tag{8}$$

where $\psi_\theta(Y, X, V, \Delta; \phi, \alpha, \eta)$ equals

$$\Delta[Y - \dot{b}\{D^{\mathrm{T}}(T)\theta\}]D(T) + (1-\Delta)\left\{\int_{C-X^{\mathrm{T}}\alpha}^{\tau} f_{\theta,\phi}(Y \mid t + X^{\mathrm{T}}\alpha, X)\,\mathrm{d}\eta(t)\right\}^{-1}$$
$$\times \int_{C-X^{\mathrm{T}}\alpha}^{\tau} f_{\theta,\phi}(Y \mid t + X^{\mathrm{T}}\alpha, X)\big[Y - \dot{b}\{D^{\mathrm{T}}(t + X^{\mathrm{T}}\alpha)\theta\}\big]D(t + X^{\mathrm{T}}\alpha)\,\mathrm{d}\eta(t),$$

which is the derivative of (6) with respect to θ. Then, with (ϕ, α, η) replaced by $(\hat{\phi}_n, \hat{\alpha}_n, \hat{\eta}_{n,\hat{\alpha}_n})$ in (8), $\Psi_{\theta,n}(\hat{\phi}_n, \hat{\alpha}_n, \hat{\eta}_{n,\hat{\alpha}_n}) = 0$ becomes the pseudolikelihood estimating equation for θ, and its solution $\hat{\theta}_n$ is called the pseudolikelihood estimator.

A set of regularity conditions is given in the Appendix. Some of the conditions are commonly assumed for accelerated failure time models, while others are for the generalized linear model and are easily verifiable for linear, logistic and Poisson regression models. We then have the following asymptotic results for $\hat{\theta}_n$.

Theorem 1. *Consider models* (3) *and* (5)*, and denote the true value of* θ *by* θ_0*. Suppose that the regularity conditions in the Appendix hold. Then the two-stage pseudolikelihood estimator* $\hat{\theta}_n$ *satisfying* $\Psi_{\hat{\theta}_n,n}(\hat{\phi}_n, \hat{\alpha}_n, \hat{\eta}_{n,\hat{\alpha}_n}) = 0$ *converges in outer probability to* θ_0*, and* $n^{1/2}(\hat{\theta}_n - \theta_0)$ *converges weakly to a zero-mean normal random variable with variance* $A^{-1}BA^{-1}$*, where* A *and* B *are given in the Supplementary Material.*

Because the asymptotic variance of $\hat{\theta}_n$ is complicated, we recommend using the bootstrap variance estimator.

The proof of Theorem 1 is based on the Z-estimation theory of Nan & Wellner (2013). Define a deterministic function

$$\Psi_\theta(\phi, \alpha, \eta) = E\{\psi_\theta(Y, X, V, \Delta; \phi, \alpha, \eta)\},$$

and denote the true values of (ϕ, α, η) by $(\phi_0, \alpha_0, \eta_0)$. We can show that $\Psi_{\theta,n}(\hat{\phi}_n, \hat{\alpha}_n, \hat{\eta}_{n,\hat{\alpha}_n})$ converges uniformly to $\Psi_\theta(\phi_0, \alpha_0, \eta_0)$ as $n \rightarrow \infty$. Then consistency is achieved when θ_0 is the unique solution of $\Psi_\theta(\phi_0, \alpha_0, \eta_0) = 0$. Asymptotic normality is derived by finding an asymptotic linear representation of $n^{1/2}(\hat{\theta}_n - \theta_0)$. The detailed proofs rely heavily on empirical process theory and can be found in the Supplementary Material, where we provide the analytical form of the asymptotic variance only for the Gehan-weighted estimate of α. The analytical forms of the asymptotic variance for other rank-based estimates and sieve maximum likelihood estimates can be obtained similarly.

Correct specification of the accelerated failure time model (5) is crucial for the results in Theorem 1, but this is a much weaker requirement than a parametric model assumption. Another advantage of the two-stage method over the parametric method is that model-checking tools are available. We suggest fitting the accelerated failure time model before applying the two-stage procedure. If the data fit the model reasonably well and there are at least some fully observed covariates that are significantly associated with the covariate subject to the detection limit, then the proposed method is recommended. Otherwise, we suggest complete-case analysis.

5. Numerical results

5·1. *Simulations*

Simulated datasets are generated from the generalized linear model

$$g\{E(Y)\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \gamma Z,$$

where $\beta_0 = -1$, $\beta_1 = 0{\cdot}5$, $\beta_2 = -1$, $\gamma = 2$, and g is chosen to be the canonical link function for the normal, Bernoulli or Poisson distribution. The normal error variance is chosen to be 1 for the linear regression model. The three covariates are $X_1 \sim \text{Ber}(0{\cdot}5)$, X_2 normal with mean 1 and standard deviation 1 truncated at ± 3, and $Z = \exp(-T)$ generated from

$$T = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \varsigma,$$

where $\alpha_0 = 0{\cdot}25$, $\alpha_1 = 0{\cdot}25$, $\alpha_2 = -0{\cdot}5$ and $\varsigma \sim 0{\cdot}5N(0, 1/8^2) + 0{\cdot}5N(0{\cdot}5, 1/10^2)$. The detection limit L for covariate Z is chosen to yield 30% right censoring for T.

Table 1. *Summary statistics for simulations to compare full data, two-stage and complete-case analyses in a linear regression setting, with biases for four substitution methods*

	Sample size		$\beta_0 = -1$	$\beta_1 = 0{\cdot}5$	$\beta_2 = -1$	$\gamma = 2$
Full data	200	Bias (×100)	−3·0	1·1	−0·6	2·4
		var (×100)	41·4	5·1	3·6	37·4
Two-stage		Bias (×100)	−2·9	1·0	−0·5	2·2
		var (×100)	43·8	5·3	3·8	39·5
		Bootstrap var (×100)	46·5	5·8	4·3	42·1
		90% coverage rate (%)	89·0	90·7	91·2	89·0
		95% coverage rate (%)	94·9	94·8	95·6	95·3
Complete-case		Bias (×100)	−1·8	0·4	0·0	1·0
		var (×100)	53·1	8·2	6·6	50·7
L		Bias (×100)	39·9	−22·0	22·8	−49·1
$L/2^{1/2}$		Bias (×100)	70·1	−13·6	14·7	−62·0
Zero		Bias (×100)	183·7	−41·7	42·7	−168·4
$E(Z \mid Z < L)$		Bias (×100)	41·5	−13·3	14·3	−41·8
Full data	400	Bias (×100)	−1·9	0·7	−0·3	1·4
		var (×100)	21·2	2·8	1·9	19·2
Two-stage		Bias (×100)	−1·9	0·8	−0·3	1·5
		var (×100)	22·5	2·9	2·0	20·4
		Bootstrap var (×100)	22·6	2·8	2·1	20·5
		90% coverage rate (%)	89·4	89·2	90·4	89·9
		95% coverage rate (%)	95·0	93·8	95·0	95·0
Complete-case		Bias (×100)	−1·9	−0·1	0·4	0·8
		var (×100)	27·3	4·3	3·3	25·5
L		Bias (×100)	40·4	−22·1	23·0	−49·5
$L/2^{1/2}$		Bias (×100)	72·4	−14·4	15·5	−64·2
Zero		Bias (×100)	185·0	−42·6	43·6	−170·0
$E(Z \mid Z < L)$		Bias (×100)	43·3	−13·8	14·8	−43·4

We simulate 1000 replications for each scenario and compare the results obtained from the proposed method, full data analysis, complete-case analysis, and four different substitution methods. The full data analysis represents the situation of no detection limit, which serves as a benchmark. We conduct simulations with sample sizes 200 and 400. The four substitution methods for $Z < L$ are: (i) replacing Z by L; (ii) replacing Z by $L/2^{1/2}$; (iii) replacing Z by zero; and (iv) replacing Z by $E(Z \mid Z < L)$. For the proposed two-stage method, we report the 90% and 95% coverage proportions, for which the variances are obtained from 200 bootstrap samples. Empirical variances obtained from 1000 independent datasets are reported only for the full data analysis and the two valid methods for the detection limit problem. The summary statistics are presented in Tables 1–3.

The results suggest that all the substitution methods yield biased estimates, including the one that replaces Z by the true $E(Z \mid Z < L)$. The biases for the proposed two-stage method are minimal, comparable to those of both the full data analysis and the complete-case analysis. Clearly, the proposed method is much more efficient than complete-case analysis, and the bootstrap method performs well in estimating the variance, which yields reasonable coverages for both sample sizes. Additional simulations with censoring rates varying from 10% to 60%, not shown here, demonstrate that the efficiency gain of the two-stage method relative to complete-case analysis increases, and the biases of all the substitution methods also increase, as the censoring rate increases.

Table 2. *Summary statistics for simulations to compare full data, two-stage and complete-case analyses in a logistic regression setting, with biases for four substitution methods*

	Sample size		$\beta_0 = -1$	$\beta_1 = 0{\cdot}5$	$\beta_2 = -1$	$\gamma = 2$
Full data	200	Bias (×100)	−3·0	1·3	−3·3	6·0
		var (×100)	215·7	26·8	21·6	203·3
Two-stage		Bias (×100)	−4·1	1·6	−3·7	7·1
		var (×100)	231·3	27·8	23·0	219·1
		Bootstrap var (×100)	242·4	29·9	23·5	226·0
		90% coverage rate (%)	91·4	91·7	90·7	91·0
		95% coverage rate (%)	96·2	96·3	95·9	96·2
Complete-case		Bias (×100)	−7·6	2·1	−4·5	10·6
		var (×100)	282·2	41·3	38·1	284·2
L		Bias (×100)	30·9	−18·5	17·1	−36·8
$L/2^{1/2}$		Bias (×100)	69·0	−12·2	11·0	−57·0
Zero		Bias (×100)	188·0	−45·3	44·1	−171·6
$E(Z \mid Z < L)$		Bias (×100)	35·0	−10·0	8·7	−31·3
Full data	400	Bias (×100)	−3·3	0·7	−1·6	4·1
		var (×100)	93·0	12·3	9·6	88·1
Two-stage		Bias (×100)	−4·3	1·1	−2·0	5·2
		var (×100)	101·3	12·9	10·4	96·4
		Bootstrap var (×100)	110·1	13·8	10·7	102·2
		90% coverage rate (%)	90·8	91·2	90·5	90·6
		95% coverage rate (%)	95·8	96·3	95·8	95·2
Complete-case		Bias (×100)	−3·7	0·5	−1·8	4·8
		var (×100)	116·9	19·0	15·9	116·0
L		Bias (×100)	31·9	−19·3	19·0	−39·8
$L/2^{1/2}$		Bias (×100)	65·1	−11·9	11·7	−55·3
Zero		Bias (×100)	184·1	−44·2	44·0	−169·1
$E(Z \mid Z < L)$		Bias (×100)	33·2	−10·3	10·1	−31·7

The consistency of the estimators from the two-stage method depends on correctly specifying model (5). Although the model assumption is much less restrictive than for a parametric model, model checking is needed before one can apply the proposed two-stage method. Additional simulations given in the Supplementary Material show that, though the problem is less severe than with ad hoc substitution methods, misspecification of the accelerated failure time model can yield biased results, and the bias increases as the severity of misspecification of (5) grows.

5·2. *The National Health and Nutrition Examination Survey*

We consider the National Health and Nutrition Examination Survey 2003–2004. In particular, we focus on examining the effect of left-censored urine arsenobetaine levels on the prevalence of type 2 diabetes (Navas-Acien et al., 2008)

The survey, conducted by the U.S. National Center for Health Statistics, used a complex multi-stage sampling design to obtain a representative sample of civilian noninstitutionalized individuals from the U.S. population. The data include a subsample of 1542 study participants for whom arsenic measurements are available. For each subject in this subsample, measurements of the levels of total urine arsenic and various arsenic species, including arsenobetaine, were collected for analysis. The detection limits for total urine arsenic and urine arsenobetaine were 0·6 and

Table 3. *Summary statistics for simulations to compare full data, two-stage and complete-case analyses in a Poisson regression setting, with biases for four substitution methods*

	Sample size		$\beta_0 = -1$	$\beta_1 = 0{\cdot}5$	$\beta_2 = -1$	$\gamma = 2$
Full data	200	Bias (×100)	2·2	−0·9	0·8	−2·6
		var (×100)	22·5	2·4	1·8	19·8
Two-stage		Bias (×100)	3·4	−1·1	1·1	−3·7
		var (×100)	25·0	2·6	2·0	22·1
		Bootstrap var (×100)	24·9	2·7	2·0	21·8
		90% coverage rate (%)	90·9	89·9	90·0	90·6
		95% coverage rate (%)	94·5	94·8	94·7	94·8
Complete-case		Bias (×100)	2·5	−1·1	1·0	−3·1
		var (×100)	35·1	5·3	4·1	32·5
L		Bias (×100)	58·9	−28·8	28·6	−66·0
$L/2^{1/2}$		Bias (×100)	88·5	−20·0	21·0	−80·1
Zero		Bias (×100)	186·7	−38·0	39·6	−169·1
$E(Z \mid Z < L)$		Bias (×100)	63·7	−21·3	21·7	−62·8
Full data	400	Bias (×100)	1·8	−0·3	0·5	−2·0
		var (×100)	10·5	1·2	0·8	9·2
Two-stage		Bias (×100)	1·9	−0·3	0·5	−2·1
		var (×100)	11·9	1·3	0·9	10·4
		Bootstrap var (×100)	12·1	1·3	1·0	10·5
		90% coverage rate (%)	90·1	90·7	90·5	90·7
		95% coverage rate (%)	95·2	95·3	94·8	95·0
Complete-case		Bias (×100)	1·6	−0·4	0·7	−2·2
		var (×100)	17·5	2·7	2·2	16·3
L		Bias (×100)	57·8	−28·1	28·3	−64·9
$L/2^{1/2}$		Bias (×100)	88·6	−19·6	20·8	−80·0
Zero		Bias (×100)	187·0	−37·3	39·1	−168·9
$E(Z \mid Z < L)$		Bias (×100)	63·3	−20·8	21·5	−62·3

0·4 μg/l, respectively. The percentages of study participants with levels below the detection limits were 1·3% for total urine arsenic and 27·8% for urine arsenobetaine (Caldwell et al., 2009). Navas-Acien et al. (2008) found that total urine arsenic was associated with increased prevalence of type 2 diabetes, so it is included as a covariate in our analysis, and 19 participants with total urine arsenic below the detection limit were dropped from the analysis. The urine creatinine level, used to account for urine dilution in spot urine samples, was fully observed and is included as a covariate in the analysis as well. We excluded 23 participants with missing values in other variables of interest.

For illustration, we focus on the male participants; of the 730 subjects, 24·1% had urine arsenobetaine below the detection limit. Age, race/ethnicity, body mass index, and the logarithms of total urine arsenic and urine creatinine level are used as covariates to fit the accelerated failure time model for −log(arsenobetaine). All of these have p-values below 0·0001, except for one dummy variable for race. Figure 1 shows 50 realizations from the distributions of the score processes. The observed score processes are presented as solid lines, which fluctuate randomly about zero. The accelerated failure time model for urine arsenobetaine fits the data reasonably well, with goodness-of-fit p-values of 0·686, 0·104, 0·706, 0·782, 0·68, 0·794, 0·646 and 0·834 for age, log total urine arsenic, log urine creatinine, body mass index, and race with five ethnic categories, respectively, obtained from 500 simulated martingale residual score processes. The response variable of interest is the status of type 2 diabetes, so logistic regression is considered with age, body

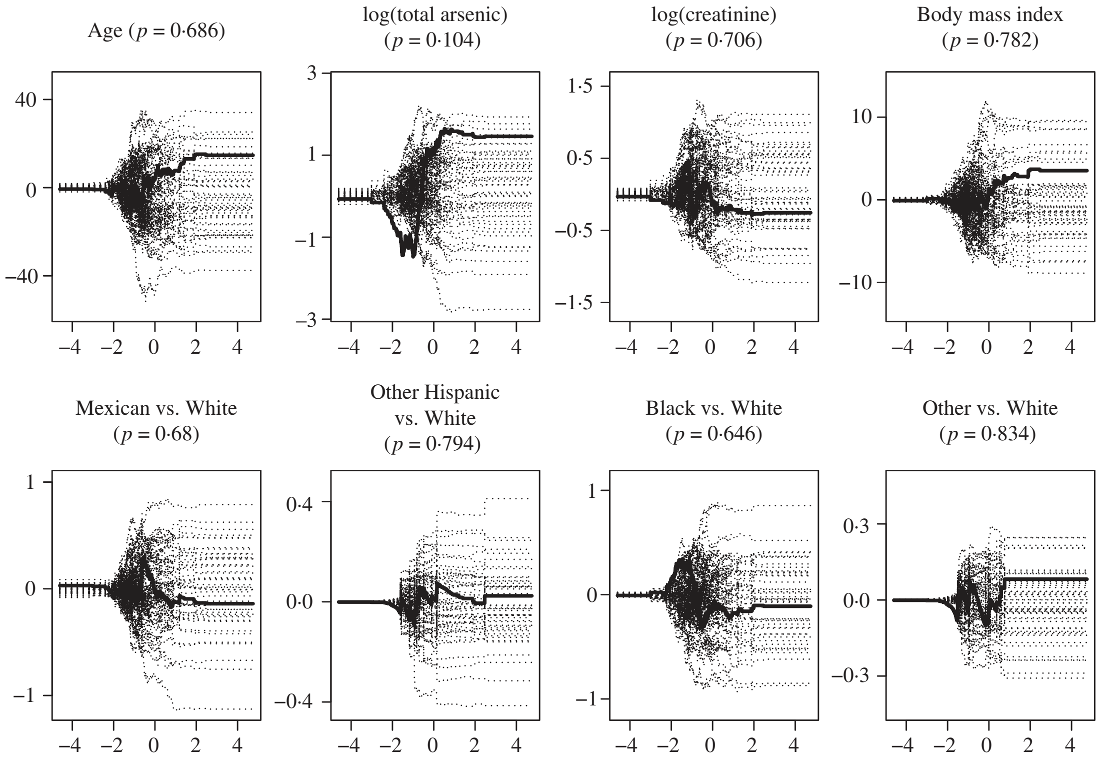

Fig. 1. Goodness of fit of the accelerated failure time model for the data from the National Health and Nutrition Examination Survey 2003–2004. In each panel, the x-axis is the residual of the fitted accelerated failure time model, and the y-axis is the cumulative martingale residual process for the covariate under consideration; the solid line is the observed score process, and the dotted lines are 50 realizations from the distribution of the score processes.

mass index, log total urine arsenic, log urine creatinine and log urine arsenobetaine as covariates. Table 4 shows that the proposed two-stage method yields similar point estimates with smaller variances and p-values than the complete-case analysis, indicating the efficiency gain of the proposed method. Substitution methods where values below the detection limit are replaced by L or $L/2^{1/2}$ yield much smaller point estimates that put the significant results of log(arsenobetaine) into question. The effect of log(arsenobetaine) is clearly biased when urine arsenobetaine values below the detection limit are replaced by 10^{-4}, rather than zero, to avoid $\log 0$.

6. Discussion

The proposed two-stage approach may not be fully efficient, so developing a fully efficient method is of great interest. It may be possible to adopt a semiparametric maximum likelihood approach by discretizing η or a sieve maximum likelihood approach via smoothing η for bundled parameters (Ding & Nan, 2011), but one should anticipate much more challenging numerical implementation and theoretical justification than for the two-stage approach.

The proposed method can be generalized to regression with multiple covariates subject to detection limits. The critical step is to provide a valid nonparametric estimate for the multivariate survival function, for which available methods include those of Dabrowska (1988), Prentice & Cai (1992), van der Laan (1996), Prentice & Moodie (2004) and Prentice (2014); we

Table 4. *Regression analysis results using data from the National Health and Nutrition Examination Survey* 2003–2004 *for the prevalence of type* 2 *diabetes with the covariate urine arsenobetaine subject to a detection limit*

		Age	log(total arsenic)	log(creatinine)	BMI	log(arsenobetaine)
Two-stage	Estimate	0·05	0·50	−0·76	0·10	−0·22
	Bootstrap SD	0·01	0·25	0·27	0·02	0·12
	p-value	< 0·01	0·04	< 0·01	< 0·01	0·07
Complete-case	Estimate	0·06	0·41	−0·81	0·13	−0·20
	SD	0·01	0·32	0·33	0·03	0·20
	p-value	< 0·01	0·20	0·01	< 0·01	0·31
L	Estimate	0·05	0·57	−0·81	0·10	−0·32
	SD	0·01	0·25	0·26	0·02	0·15
	p-value	< 0·01	0·02	< 0·01	< 0·01	0·03
$L/2^1/2$	Estimate	0·05	0·56	−0·80	0·10	−0·30
	SD	0·01	0·24	0·26	0·02	0·14
	p-value	< 0·01	0·02	< 0·01	< 0·01	0·02
Zero	Estimate	0·05	0·25	−0·61	0·10	−0·03
	SD	0·01	0·16	0·24	0·02	0·02
	p-value	< 0·01	0·12	< 0·01	< 0·01	0·11

SD, standard deviation; BMI, body mass index.

recommend the last due to its efficiency and simplicity. We consider a constant detection limit in this article because it is commonly encountered in practice, but this need not constrain the application of our method, which can be used with random detection limits under the same regularity conditions.

When the original laboratory measurements are available, Murphy et al. (2010) and Buck Louis et al. (2012) directly used the machine-read lab values in their analysis to avoid potential biases caused by substitution. More appropriate analyses would treat these data as error-prone (Guo et al., 2010), so methods dealing with measurement errors would apply.

The two-stage method can be generalized to other regression models that have likelihood functions, such as the Cox model for a failure time Y that is also subject to censoring. Extra care will be needed due to the special features of the Cox model, where the nonparametric baseline hazard function may either be estimated from the complete data in the first stage and then plugged into the estimating function for θ in the second stage, or be estimated directly in the second stage after the usual discretization. An anonymous referee pointed out related work by Lee et al. (2003) that would be relevant to further exploration along this line.

Supplementary material

Supplementary Material available at *Biometrika* online contains the proof of Theorem 1 and results of additional simulations.

Acknowledgement

We thank Professor Sung Kyun Park for the National Health and Nutrition Examination Survey arsenic exposure data. We are grateful to the editor, associate editor and two referees for constructive comments. This research was supported by the U.S. National Science Foundation and National Institutes of Health.

Appendix

Regularity conditions

Let $\mathcal{Y}$ denote the sample space of the response variable Y, $\mathcal{X}$ the sample space of the covariate X, Θ the parameter space of θ, $\mathcal{A}$ the parameter space of α, and $\mathcal{H}$ the parameter space of η. In addition to the assumptions of bounded support for (X, Z) and compactness of the parameter spaces Θ and $\mathcal{A}$, we state a set of regularity conditions for Theorem 1:

Condition A1. $\Psi_\theta(\phi_0, \alpha_0, \eta_{0,\alpha_0})$ has a unique root θ_0;

Condition A2. for any constant $U < \infty$, $\sup_{t\in[C,U]} |h(t)| \leqslant c_0 < \infty$, $\sup_{t\in[C,U]} |\dot{h}(t)| \leqslant c_1 < \infty$, and $\sup_{t\in[C,U]} |\ddot{h}(t)| \leqslant c_2 < \infty$, where $\dot{h}$ and $\ddot{h}$ are the first and second derivatives of h, respectively, and c_0, c_1 and c_2 are constants;

Condition A3. ς has bounded density $f = \dot{\eta}_{0,\alpha_0}$ with bounded derivative $\dot{f}$; in other words, $f \leqslant c_3 < \infty$ and $|\dot{f}| \leqslant c_4 < \infty$ for constants c_3 and c_4, and

$$\int_{-\infty}^{\infty} \{\dot{f}(t)/f(t)\}^2 f(t)\,\mathrm{d}t < \infty;$$

Condition A4. there is a constant truncation time $\tau < \infty$ such that $\mathrm{pr}(V - X'\alpha > \tau \mid X = x) \geqslant \xi > 0$ for all $x \in \mathcal{X}$ and $\alpha \in \mathcal{A}$;

Condition A5. $a(\phi)$ is a monotone function satisfying $|1/a(\phi)| \leqslant l < \infty$ for a constant l, and has bounded derivatives $\dot{a}(\cdot)$ and $\ddot{a}(\cdot)$;

Condition A6. $\dot{b}(\cdot)$ is a bounded monotone function;

Condition A7. $\ddot{b}(\cdot)$ is a bounded Lipschitz function;

Condition A8. there exist constants C_i $(i = 1, \ldots, 5)$ such that for any constant $U < \infty$,

$$\sup_{y\in\mathcal{Y},\theta\in\Theta,|1/a(\phi)|\leqslant l,x\in\mathcal{X},t\in[C,U]} \left| f_{\theta,\phi}(y \mid t, x)[y - \dot{b}\{D^{\mathrm{T}}(t)\theta\}]\right| \leqslant C_1 < \infty,$$

$$\sup_{y\in\mathcal{Y},\theta\in\Theta,|1/a(\phi)|\leqslant l,x\in\mathcal{X},t\in[C,U]} \left| \frac{\partial f_{\theta,\phi}(y \mid t, x)}{\partial \phi}[y - \dot{b}\{D^{\mathrm{T}}(t)\theta\}]\right| \leqslant C_2 < \infty,$$

$$\sup_{y\in\mathcal{Y},\theta\in\Theta,|1/a(\phi)|\leqslant l,x\in\mathcal{X},t\in[C,U]} \left| \frac{\partial (f_{\theta,\phi}(y \mid t, x)[y - \dot{b}\{D^{\mathrm{T}}(t)\theta\}])}{\partial t}\right| \leqslant C_3 < \infty,$$

$$\sup_{y\in\mathcal{Y},\theta\in\Theta,|1/a(\phi)|\leqslant l,x\in\mathcal{X},t\in[C,U]} \left| \frac{\partial f_{\theta,\phi}(y \mid t, x)}{\partial \phi}\right| \leqslant C_4 < \infty,$$

$$\sup_{y\in\mathcal{Y},\theta\in\Theta,|1/a(\phi)|\leqslant l,x\in\mathcal{X},t\in[C,U]} \left| \frac{\partial (f_{\theta,\phi}(y \mid t, x)[y - \dot{b}\{D^{\mathrm{T}}(t)\theta\}])}{\partial \theta}\right| \leqslant C_5 < \infty;$$

Condition A9. there exist constants $\delta_1 > 0$ and $\delta_2 > 0$ such that $\int_{C-X^{\mathrm{T}}\alpha}^{\tau} f_{\theta,\phi}(Y \mid t + X^{\mathrm{T}}\alpha, X)\,\mathrm{d}\eta(t) \geqslant \delta_1$ with probability 1 for any $\theta \in \Theta$ and $|\phi - \phi_0| + |\alpha - \alpha_0| + \|\eta - \eta_0\| < \delta_2$;

Condition A1 may be unnecessarily strong for the proposed two-stage method. Direct calculation shows that $\Psi_{\theta_0}(\phi_0, \alpha_0, \eta_0) = 0$ under Condition A4, and yields

$$\dot{\Psi}_{\theta_0} = \frac{\partial \Psi_\theta(\phi_0, \alpha_0, \eta_0)}{\partial \theta}\bigg|_{\theta=\theta_0}$$

$$= E\left[-\Delta \ddot{b}\{D^{\mathrm{T}}(T)\theta_0\}D(T)^{\otimes 2} - (1-\Delta)\left\{\int_{C-X^{\mathrm{T}}\alpha_0}^{\tau} f_{\theta_0,\phi_0}(Y \mid t + X^{\mathrm{T}}\alpha_0, X)\,\mathrm{d}\eta_0(t)\right\}^{-2}\right.$$

$$\left.\times \left\{\int_{C-X^{\mathrm{T}}\alpha_0}^{\tau} f_{\theta_0,\phi_0}(Y \mid t + X^{\mathrm{T}}\alpha_0, X)\big[Y - \dot{b}\{D^{\mathrm{T}}(t + X^{\mathrm{T}}\alpha_0)\theta_0\}\big]D(t + X^{\mathrm{T}}\alpha_0)\,\mathrm{d}\eta_0(t)\right\}^{\otimes 2}\right],$$

which is negative definite. Hence $\dot{\Psi}_\theta$, a continuous matrix with respect to θ, is also negative definite in a neighbourhood of θ_0, which guarantees that θ_0 is the unique solution of $\Psi_\theta(\phi_0, \alpha_0, \eta_0) = 0$ in a neighbourhood of θ_0. The initial value that we use in the Newton–Raphson algorithm for solving $\Psi_{\theta,n}(\hat{\phi}_n, \hat{\alpha}_n, \hat{\eta}_{n,\hat{\alpha}_n}) = 0$ is obtained from the complete-case analysis, which is consistent, so the solution of the proposed two-stage method should also be consistent.

Condition A2 holds for many commonly-used transformations, such as $h(t) = \exp(-t)$ and polynomial functions. Conditions A3 and A4 are standard for accelerated failure time models (Tsiatis, 1990; Nan et al., 2009). It would be of interest to allow a data-dependent τ in Condition A4, such as the soft-truncation of Lai & Ying (1991). However, additional theoretical work is needed to justify such a choice of τ. Conditions A5–A8 hold automatically for common generalized linear models.

Condition A9 is assumed mainly for technical convenience. One way to obtain it is to truncate the response variable Y such that $|Y| \leqslant M < \infty$ for a large constant M and to further truncate the residual in the accelerated failure time model with some constant $\tau' < \tau$. In our simulations, we do not implement such truncations but still obtain satisfactory results.

References

Agresti, A. (2002). *Categorical Data Analysis*. Hoboken, New Jersey: Wiley, 2nd ed.

Albert, P. S., Harel, O., Perkins, N. & Browne, R. (2010). Use of multiple assays subject to detection limits with regression modeling in assessing the relationship between exposure and outcome. *Epidemiology* **21**, S35–43.

Arunajadai, S. G. & Rauh, V. A. (2012). Handling covariates subject to limits of detection in regression. *Envir. Ecol. Statist.* **19**, 369–91.

Box, G. E. P. & Cox, D. R. (1964). An analysis of transformations (with Discussion). *J. R. Statist. Soc.* B **26**, 211–52.

Buck Louis, G. M., Chen, Z., Peterson, C. M., Hediger, M. L., Croughan, M. S., Sundaram, R., Stanford, J. B., Varner, M. W., Fujimoto, V. Y., Giudice, L. C. et al. (2012). Persistent lipophilic environmental chemicals and endometriosis: The ENDO study. *Envir. Health Perspect.* **120**, 811–6.

Cai, T., Tian, L. & Wei, L. J. (2005). Semiparametric Box–Cox power transformation models for censored survival observations. *Biometrika* **92**, 619–32.

Caldwell, K. L., Jones, R. L., Verdon, C. P., Jarrett, J. M., Caudill, S. P. & Osterloh, J. D. (2009). Levels of urinary total and speciated arsenic in the US population: National Health and Nutrition Examination Survey 2003–2004. *J. Expo. Sci. Envir. Epidemiol.* **19**, 59–68.

Cole, S. R., Chu, H., Nie, L. & Schisterman, E. F. (2009). Estimating the odds ratio when exposure has a detection limit. *Int. J. Epidemiol.* **38**, 1674–80.

Dabrowska, D. M. (1988). Kaplan–Meier estimate on the plane. *Ann. Statist.* **16**, 1475–89.

D'Angelo, G. & Weissfeld, L. (2008). An index approach for the Cox model with left censored covariates. *Statist. Med.* **72**, 4502–14.

Ding, Y. & Nan, B. (2011). A sieve M-theorem for bundled parameters in semiparametric models, with application to the efficient estimation in a linear model for censored data. *Ann. Statist.* **39**, 3032–61.

Foster, A. M., Tian, L. & Wei, L. J. (2001). Estimation for the Box–Cox transformation model without assuming parametric error distribution. *J. Am. Statist. Assoc.* **96**, 1097–101.

Gong, G. & Samaniego, F. J. (1981). Pseudo maximum likelihood estimation: Theory and applications. *Ann. Statist.* **9**, 861–9.

Guo, Y., Harel, O. & Little, R. J. (2010). How well quantified is the limit of quantification? *Epidemiology* **21**, S10–6.

Helsel, D. R. (2006). Fabricating data: How substituting values for nondetects can ruin results, and what can be done about it. *Chemosphere* **56**, 2434–9.

Jin, Z., Lin, D. Y., Wei, L. J. & Ying, Z. (2003). Rank-based inference for the accelerated failure time model. *Biometrika* **90**, 341–53.

Kim, C., Kong, S., Laughlin, G. A., Golden, S. H., Mather, K. J., Nan, B., Randolph, J. R., Edelstein, S. L., Labrie, F., Buschur, E. et al. (2013). Reductions in glucose among postmenopausal women who use and do not use estrogen therapy. *Menopause* **20**, 393–400.

Lai, T. L. & Ying, Z. (1991). Large sample theory of a modified Buckley–James estimator for regression analysis with censored data. *Ann. Statist.* **19**, 1370–402.

Lee, S., Park, S. H. & Park, J. (2003). The proportional hazards regression with a censored covariate. *Statist. Prob. Lett.* **61**, 309–19.

Lin, D. Y., Wei, L. J. & Ying, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika* **80**, 557–72.

Lin, D. Y., Robins, J. M. & Wei, L. J. (1996). Comparing two failure time distributions in the presence of dependent censoring. *Biometrika* **83**, 381–93.

Little, R. J. & Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. New York: Wiley, 2nd ed.

Lynn, H. (2001). Maximum likelihood inference for left-censored HIV RNA data. *Statist. Med.* **20**, 33–45.

McCullagh, P. & Nelder, J. (1989). *Generalized Linear Models*. Boca Raton, Florida: Chapman and Hall/CRC, 2nd ed.

Moulton, L. H., Curriero, F. C. & Barroso, P. F. (2002). Mixture models for quantitative HIV RNA data. *Statist. Meth. Med. Res.* **11**, 317–25.

Murphy, L. E., Gollenberg, A. L., Buck Louis, G. M., Kostyniak, P. J. & Sundaram, R. (2010). Maternal serum preconception polychlorinated biphenyl concentrations and infant birth weight. *Envir. Health Perspect.* **118**, 297–302.

Nan, B. & Wellner, J. A. (2013). A general semiparametric Z-estimation approach for case-cohort studies. *Statist. Sinica* **23**, 1155–80.

Nan, B., Kalbfleisch, J. D. & Yu, M. (2009). Asymptotic theory for the semiparametric accelerated failure time model with missing data. *Ann. Statist.* **37**, 2351–76.

Navas-Acien, A., Silbergeld, E. K., Pastor-Barriuso, R. & Guallar, E. (2008). Arsenic exposure and prevalence of type 2 diabetes in US adults. *J. Am. Med. Assoc.* **300**, 814–22.

Nie, L., Chu, H., Liu, C., Cole, S. R., Vexler, A. & Schisterman, E. F. (2010). Linear regression with an independent variable subject to a detection limit. *Epidemiology* **21**, S17–24.

Peng, L. & Fine, J. P. (2006). Rank estimation of accelerated lifetime models with dependent censoring. *J. Am. Statist. Assoc.* **101**, 1085–93.

Prentice, R. L. (2014). Self-consistent nonparametric maximum likelihood estimator of the bivariate survivor function. *Biometrika* **101**, 505–18.

Prentice, R. L. & Cai, J. (1992). Covariance and survivor function estimation using censored multivariate failure time data. *Biometrika* **79**, 495–512.

Prentice, R. L. & Moodie, F. Z. (2004). Hazard-based nonparametric survivor function estimation. *J. R. Statist. Soc.* B **66**, 305–19.

R Development Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.

Richardson, D. B. & Ciampi, A. (2003). Effects of exposure measurement error when an exposure variable is constrained by a lower limit. *Am. J. Epidemiol.* **157**, 355–63.

Schisterman, E. F. & Little, R. J. (2010). Opening the black box of biomarker measurement error. *Epidemiology* **21**, S1–3.

Sowers, M. R., Eyvazzadeh, A. D., McConnell, D., Yosef, M., Jannausch, M. L., Zhang, D., Harlow, S. & Randolph, J. F. Jr. (2008). Anti-mullerian hormone and inhibin B in the definition of ovarian aging and the menopause transition. *J. Clin. Endocrinol. Metab.* **93**, 3478–83.

Tsiatis, A. A. (1990). Estimating regression parameters using linear rank tests for censored data. *Ann. Statist.* **18**, 354–72.

van de Laan, M. J. (1996). Efficient estimation in the bivariate censoring model and repairing NPMLE. *Ann. Statist.* **24**, 596–627.

Wei, L. J., Ying, Z. & Lin, D. Y. (1990). Linear regression analysis of censored survival data based on rank tests. *Biometrika* **77**, 845–51.

Wu, H., Chen, Q., Ware, L. B. & Koyama, T. (2012). A Bayesian approach for generalized linear models with explanatory biomarker measurement variables subject to detection limit: An application to acute lung injury. *J. Appl. Statist.* **39**, 1733–47.

Yu, M. & Nan, B. (2006). A hybrid Newton-type method for censored survival data using double weights in linear models. *Lifetime Data Anal.* **12**, 345–64.

Zhou, M. (2006). *rankreg: Rank regression for censored data AFT model*. R package version 0·2-2, http://ftp.auckland.ac.nz/software/CRAN/src/contrib/Descriptions/rankreg.html.

[*Received September* 2014. *Revised September* 2015]