The Canadian Journal of Statistics
Vol. 00, No. 00, 2020, Pages 000–000
La revue canadienne de statistique

A semiparametric regression model under biased sampling and random censoring: A local pseudo-likelihood approach

Yassir RABHI[1*] and Masoud ASGHARIAN[2]

[1]*Department of Mathematical Sciences, University of Essex, Colchester, U.K.*
[2]*Department of Mathematics and Statistics, McGill University, Montreal, Quebec, Canada*

Key words and phrases: Bandwidth selection; biased sampling; covariate bias; informative censoring; left-truncated and right-censored data; local pseudo-likelihood; semiparametric regression model.

MSC 2010: Primary 62N01; secondary 62G08.

Abstract: Methodologies developed for left-truncated right-censored failure time data can mostly be categorized according to the assumption imposed on the truncation distribution, i.e., being completely unknown or completely known. While the former approach enjoys robustness, the latter is more efficient when the assumed form of the truncation distribution can be supported by the data. Motivated by data from an HIV/AIDS study, we consider the middle ground and develop methodologies for estimation of a regression function in a semiparametric setting where the truncation distribution is parametrically specified while the failure time, censoring and covariate distribution are left completely unknown. We devise an estimator for the regression function based on a local pseudo-likelihood approach that properly accounts for the bias induced on the response variable and covariate(s) by the sampling design. One important spin-off from these results is that they yield the adjustment for length-biased sampling and right-censoring; the so-called stationary case. We study the small and large sample behaviour of our estimators. The proposed method is then applied to analyze a set of HIV/AIDS data. *The Canadian Journal of Statistics* 00: 000–000; 2020 © 2020 Statistical Society of Canada

Résumé: La méthodologie développée pour les données tronquées à gauche et censurées à droite peut généralement être classée en deux groupes selon les hypothèses à propos de la distribution de troncature, à savoir si elle est complètement inconnue ou complètement connue. Alors que la première catégorie fait preuve de robustesse, la seconde est plus efficace lorsque les données peuvent supporter la forme de la distribution supposée pour la troncature. Motivés par des données sur le SIDA, les auteurs proposent un compromis en introduisant un estimateur semi-paramétrique de la fonction de régression pour les données tronquées à gauche et censurées à droite. Ils supposent un modèle paramétrique pour la troncature, tandis que les distributions de la durée de survie et les covariables sont considérées comme inconnues. Ils adoptent une approche de pseudo-vraisemblance locale pour développer leur estimateur en tenant compte du biais induit par le plan d'échantillonnage. Un cas d'intérêt porte sur l'estimateur de la fonction de régression pour les durées de survie biaisées par la longueur, le cas dit stationnaire. Les auteurs étudient les propriétés théoriques et les performances numériques des estimateurs pour des échantillons petits et grands. Ils illustrent finalement leur méthode d'estimation avec des données sur le SIDA. *La revue canadienne de statistique* 00: 000–000; 2020 © 2020 Société statistique du Canada

* *Author to whom correspondence may be addressed.*
E-mail: yassir.rabhi@essex.ac.uk

1. INTRODUCTION

Prospective cohort studies on incident cases, in which disease-free individuals are recruited and followed until death, loss to follow-up, or study termination, is the gold standard when studying the time between an initiating event, say the onset of a disease, and a terminating event, say death. Logistic or other constraints may, however, preclude the possibility of recruiting incident cases. A convenient alternative in such circumstances is to conduct a prevalent cohort study, whereby groups of diseased individuals are recruited at some fixed time, their onset time is ascertained retrospectively and followed until death or loss to follow-up. Under such design a response time, say time from the disease onset to death, for an individual in the population of interest is observed if this response time exceeds an associated truncation time. Despite its feasibility and ease, this design generally leads to selecting subjects with longer survival times. Consequently, the survival data collected on prevalent cases form a biased sample from the population of interest. In other words, the survival times of subjects so recruited are left-truncated. When identified cases are further followed-up, their lifetimes may also be subject to right-censoring. Such censoring is informative. Another aspect of the truncation mechanism is the induced bias on the covariate(s) (Bergeron, Asgharian & Wolfson, 2008).

In prevalent cohort studies, the truncation distribution reflects and explains the patterns of disease incidence over a time period, and in some cases, this distribution can be recovered and parameterized. Dementia incidence rate, for instance, is known to reasonably remain constant over time, i.e., the truncation variable follows a uniform distribution (see Rabhi & Asgharian, 2017). As another example, the HIV-infection onset process was known to have an exponential growth in the 1980's because of its epidemic nature during this period (Wang, 1989, 1992). In this case, the truncation variable can reasonably be modelled by one of the exponential distributions, e.g., Weibull, Gamma, or Exponential cdf [see Figure 1(b)]. When the available information about the truncation cdf is incorporated into a statistical model, the outcome is usually an efficient estimation method as compared with a fully-nonparametric one (Wang, 1989; Asgharian, M'lan & Wolfson, 2002; de Uña-Àlvarez & Iglesias-Pérez, 2010). This is what motivates our study and this article.

In this article, we consider the semiparametric regression model

$$Y^* = \mu(X^*) + \varepsilon, \tag{1}$$

where $\mu(x) = E[Y^*|X^* = x]$ is an unknown smooth function and ε is a random error ($E[\varepsilon|X^* = x] = 0$ and $E[\varepsilon^2|X^* = x] < \infty$), to study the relation between a positive response Y^* (e.g., lifetime) and a covariate X^*. The bivariate (Y^*, X^*) is subject to a random left-truncation T^*, meaning that only subjects satisfying $Y^* \geq T^*$ are observed and followed. In addition, the response Y^*, given $Y^* \geq T^*$, may be subject to right-censoring. Here, we consider a semiparametric model that leaves the cdf of variables of interest Y^* and X^* unspecified, but parameterizes the cdf of T^*. The objective is to nonparametrically estimate μ when the truncation distribution belongs to a known parametric family $\left\{\mathcal{F}_\theta, \theta \in \Theta\right\}$, with unknown parameter θ. This will involve the estimation of θ, which makes the methodology presented in this article semiparametric in some sense. It should be noted that accounting for biased sampling is crucial, and that failing to do so leads to biased results.

Nonparametric estimation of regression function under biased sampling and right-censoring has been mainly studied in two different settings. The first is general truncation, where the truncation cdf is unknown and left unspecified (see Iglesias-Pérez & González-Manteiga, 1999; Liang, de Uña-Àlvarez & Iglesias-Pérez, 2015; Gürler, 1996). The second is the length-biased case, in which the truncation-times are assumed to be uniform, called the stationary assumption (see Wu, 2000; de Uña-Àlvarez, 2003; Cristóbal, Ojeda & Alcalá, 2004), in the uncensored case, and de Uña-Àlvarez & Iglesias-Pérez (2010) for fixed right-censoring]. While the first approach is more robust to the form of the truncation cdf, the second which incorporates available information in the truncation variable is more efficient (see de Uña-Àlvarez & Iglesias-Pérez, 2010).

To the best of our knowledge, however, no methodology has been proposed in the literature for nonparametric estimation of a regression function that combines the advantages of both approaches. In this article, we introduce a semiparametric regression method that retains the efficiency of the "known truncations" approach (see Theorem 1 and Corollary 1) and maintains some robustness to the form of the truncation distribution H, that is, the method allows for the possibility that H belongs to a class of parametric distributions $\mathcal{F}_\theta$. Wang (1989) presented several real examples that assume a parametric form for the truncation distribution is plausible. We devise an estimator for μ based on a local pseudo-likelihood approach, and derive a conditional MLE for the parameter θ. Also, we develop a pseudo-likelihood estimator for μ when the sampling is subject to uniform truncations and random censoring, showing its efficiency as compared with a classical estimator for unspecified H (see Proposition 1(ii)). In addition, we propose a bandwidth selection method to choose the smoothing parameter for the regression function estimator. The rest of this article is organized as follows. In Section 2 we introduce the estimator of the semiparametric regression function μ and study its asymptotic properties. The estimation of μ under length-biased sampling and random censoring is discussed in Section 3, while in Section 4 we introduce a bandwidth selection procedure to choose the smoothing parameter. In Section 5 we study the finite sample behaviour of the semiparametric regression estimator, by means of simulations, and apply our methodology to analyze a set of HIV/AIDS data. The proofs of the main results are given in the Appendix.

2. SEMIPARAMETRIC REGRESSION MODEL

2.1. Data Setting and Notations

We start by defining the variables that represent the general population and the data obtained from the cross-sectional (c-s) sampling, with follow-up. Let (Y^*, X^*) and T^* be two independent random vectors representing, respectively, the bivariate response-covariate (variables of interest) and the truncation-time from the population (T^* is a univariate vector). At the c-s sampling time, one only observes the data (Y^*, X^*, T^*) given that $Y^* \geq T^*$. The resulting sample is biased, as such, we denote by (Y, X, T) the triplet of response-covariate-truncation associated with the observed subjects, which arises from the conditional distribution of (Y^*, X^*, T^*) given that $Y^* \geq T^*$. When the n selected subjects to the study are further followed-up, their residual lifetime $R = Y - T$ is subject to random right-censoring R_c.

The observed data, obtained from such sampling, are of the form $\{(T_i, X_i, Z_i, \delta_i),\ i = 1, \ldots, n\}$, where $Z = T + \gamma$, $\gamma = \min(R, R_c)$ and $\delta = \mathrm{I}(R \leq R_c)$ is the censoring indicator. In this work, we only consider the case of uncensored covariate X (e.g., age of individual at disease onset). However, as noted above, X suffers from a bias induced by the c-s sampling design. In the sequel, we assume that R_c is independent of (Y, X, T). This assumption is common in right-censored left-truncated data settings, and is reasonable in most practical situations (see Bergeron, Asgharian & Wolfson, 2008).

Let H, with density h, and G denote the respective distributions of the truncation-time and the residual censoring. In this section, H is assumed to belong to a known parametric family $\mathcal{F}_\theta$ with unknown θ. The case of uniform truncations is discussed in §3. Let K denote a kernel density function, b_n a bandwidth sequence tending to 0, and $K_b(t) = K(t/b_n)$. In the sequel, $\mathcal{A}$ denotes the compact support set of the density of X^* and $\mathcal{B}$ denotes the interval $[l_L, u_L)$, where l_L and u_L are the lower and upper bounds of the support of $L(y) = P[Z \leq y]$.

2.2. Estimators

In this section, we describe the methodology for estimating the regression function μ and the truncation parameter θ, when $H \in \mathcal{F}_\theta$, under right-censored left-truncated data. First, notice that the population distribution $F_{Y^*|X^*}(y|x) = P[Y^* \leq y|X^* = x]$ is related to the cross-sectional-sample distribution $F_{Z,\delta|X}(y, 1|x) = P[Z \leq y, \delta = 1|X = x]$ through the equation

$$F_{Y^*|X^*}(y|x) = \frac{\int_{v \leq y} w^{-1}(v;\theta)\,\mathrm{d}F_{Z,\delta|X}(v, 1|x)}{\int_{v>0} w^{-1}(v;\theta)\,\mathrm{d}F_{Z,\delta|X}(v, 1|x)}, \tag{2}$$

where w is a weight function, given by

$$w(z;\theta) = \int_0^z \left[1 - G(z-u)\right] \mathrm{d}H(u;\theta).$$

Our approach for estimating the conditional distribution $F_{Y^*|X^*}$, and the regression function $\mu(x) = \int_{y>0} y\,\mathrm{d}F_{Y^*|X^*}(y|x)$, is based on a local version of the pseudo log-likelihood of $\{(Z_i, T_i)|\,(\delta_i = 1, X_i = x),\ i = 1, \ldots, n\}$:

$$\begin{aligned}
\ell_{pl} &= \log\left\{\prod_{i=1}^n \frac{\left[1 - G(z_i - t_i)\right] \mathrm{d}F_{Y,T,X}(z_i, t_i, x)}{\int_{r\geq 0}\int_{u-t\leq r} \mathrm{d}F_{Y,T,X}(u, t, x)\,\mathrm{d}G(r)}\right\}\\
&= \log\left\{\prod_{i=1}^n \frac{\left[1 - G(z_i - t_i)\right] \mathrm{d}F_{T^*}(t_i)\,\mathrm{d}F_{Y^*|X^*}(z_i|x)}{\int_{u>0} w(u;\theta)\,\mathrm{d}F_{Y^*|X^*}(u|x)}\right\}\\
&= \sum_{i=1}^n \log\left\{\frac{\mathrm{d}F_{Y^*|X^*}(z_i|x)}{\int_{u>0} w(u;\theta)\,\mathrm{d}F_{Y^*|X^*}(u|x)}\right\} + Ct = \ell^*_{pl} + Ct,
\end{aligned}$$

(Ct does not depend on $F_{Y^*|X^*}$) given by

$$\ell^*_{lpl} = \sum_{i=1}^n K_b(x - X_i)\,\delta_i\,\log\left\{\frac{\mathrm{d}F_{Y^*|X^*}(z_i|x)}{\int_{u>0} w(u;\theta)\,\mathrm{d}F_{Y^*|X^*}(u|x)}\right\}. \tag{3}$$

The local pseudo log-likelihood in (3) is defined over observations in the covariate neighbourhood of the target point x, and is essentially a weighted version of ℓ^*_{lp}. Covariate observations X_i $(i = 1, \ldots, n)$ that are associated with uncensored responses and closer to x are given more weight than the further ones. The nonparametric maximum pseudo-likelihood estimator of $p_i = \mathrm{d}F_{Y^*|X^*}\left(z_i|x\right)$ (the mass attached to z_i given $X = x$) derived from ℓ^*_{lpl} is

$$\widehat{p}_i = \frac{K_b(x - X_i)\,w^{-1}(Z_i;\theta)\,\delta_i}{\sum_{j=1}^n K_b(x - X_j)\,w^{-1}(Z_j;\theta)\,\delta_j}. \tag{4}$$

It remains to replace the unknown quantities in the expression of w by their estimates. When G is unknown, one may replace G by its Kaplan-Meier (K-M) estimator $\widehat{G}$. Now, to estimate θ, we consider the conditional likelihood of the truncations times T_i given (Z_i, δ_i, X_i), $i = 1, \ldots, n$:

$$\mathcal{L}_c(\theta) = \prod_{i=1}^n \frac{\mathrm{d}H(t_i;\theta)}{w(z_i;\theta)^{\delta_i}\,v(z_i;\theta)^{1-\delta_i}} \left\{\mathrm{d}G(z_i - t_i)^{1-\delta_i}\left[1 - G(z_i - t_i)\right]^{\delta_i}\right\}, \tag{5}$$

where $v(z;\theta) = \int_0^z h(z-u;\theta)\,\mathrm{d}G(u)$. This conditional likelihood approach was considered previously by Wang (1989), in the context of estimating the distribution of uncensored lifetimes under parametric truncation model (see also Andersen, 1970). The terms G and $\mathrm{d}G$ in (5) can be ignored since they do not depend on θ. The proposed estimator for θ is the MLE obtained from the conditional likelihood $\mathcal{L}_c$;

$$\widehat{\theta} = \arg\max_{\theta_0} \mathcal{L}_c^*(\theta_0) = \arg\max_{\theta_0} \left\{ \prod_{i=1}^{n} \frac{\mathrm{d}H(t_i;\theta_0)}{\widehat{w}\big(z_i;\theta_0\big)^{\delta_i}\,\widehat{v}\big(z_i;\theta_0\big)^{1-\delta_i}} \right\}, \tag{6}$$

where $\widehat{w}\big(z;\theta\big) = \int_0^z \big[1-\widehat{G}(z-u)\big]\mathrm{d}H(u;\theta)$ and $\widehat{v}\big(z;\theta\big) = \int_0^z h(z-u;\theta)\,\mathrm{d}\widehat{G}(u)$. Now, replacing the weight function w in (4) by its estimator

$$\widehat{w}\big(y;\widehat{\theta}\big) = \int_0^y \Big[1-\widehat{G}(y-u)\Big]\,\mathrm{d}H(u;\widehat{\theta}),$$

the estimator of $F_{Y^*|X^*}(.|x)$ is

$$\widehat{F}_{Y^*|X^*}\big(y;\widehat{\theta}|x\big) = \sum_{i=1}^{n} \frac{K_b(x-X_i)\,\widehat{w}^{-1}(Z_i;\widehat{\theta})\,\delta_i}{\sum_{j=1}^{n} K_b(x-X_j)\,\widehat{w}^{-1}(Z_j;\widehat{\theta})\,\delta_j}\,\mathrm{I}(Z_i \le y).$$

Hence, the semiparametric regression estimator of μ is

$$\widehat{\mu}\big(x;\widehat{\theta}\big) = \sum_{i=1}^{n} \frac{K_b(x-X_i)\,\widehat{w}^{-1}(Z_i;\widehat{\theta})\,\delta_i}{\sum_{j=1}^{n} K_b(x-X_j)\,\widehat{w}^{-1}(Z_j;\widehat{\theta})\,\delta_j}\,Z_i. \tag{7}$$

This local pseudo-likelihood approach is equivalent to the case where we replace w and $F_{Z,\delta|X}(.,1|x)$, in the R.H.S. of Equation (2), by their respective estimates $\widehat{w}\big(.;\widehat{\theta}\big)$ and

$$\widehat{F}_{Z,\delta|X}(y,1|x) = \sum_{i=1}^{n} \frac{K_b(x-X_i)}{\sum_{j=1}^{n} K_b(x-X_j)}\,\mathrm{I}(Z_i \le y, \delta_i = 1).$$

We note that the regression function estimator $\widehat{\mu}$ is defined on the set $\mathcal{A}$, and that $\widehat{w}$ and $\widehat{v}$ are defined on $\mathcal{B}$, and the conditional functions estimators $\widehat{F}_{Y^*|X^*}(y|x)$ and $\widehat{F}_{Z,\delta|X}(y,1|x)$ are defined on $\mathcal{A}\times\mathcal{B}$.

2.3. Asymptotic Properties

We begin this section by establishing the limit distribution of $\widehat{\mu}(.;\theta)$, evaluated at a fixed value of θ, defined by

$$\widehat{\mu}\big(x;\theta\big) = \sum_{i=1}^{n} \frac{K_b(x-X_i)\,\widehat{w}^{-1}(Z_i;\theta)\,\delta_i}{\sum_{j=1}^{n} K_b(x-X_j)\,\widehat{w}^{-1}(Z_j;\theta)\,\delta_j}\,Z_i.$$

The assumptions used in the next results are given in the Appendix.

Theorem 1. *Let $m(x) = E\big[H(Y^*;\theta)|X^*=x\big]$ and $\widehat{f}_X(x) = (nh_n)^{-1}\sum_{i=1}^{n} K_b(x-X_i)$ the kernel estimator of the density $f_X(x)$ of X. Suppose B1, B2, K1, and K2 hold, then $\widehat{\mu}(x;\theta)$, for $x\in\mathcal{A}$, admits the representation*

$$\widehat{\mu}(x;\theta) - \mu(x) = \frac{1}{nb_n}\sum_{i=1}^{n} K_b(x-X_i)\frac{\delta_i}{w(Z_i)}\Big[Z_i-\mu(x)\Big]\times\frac{m(x)}{\widehat{f}_X(x)} + o_p\left(\frac{1}{\sqrt{nb_n}}\right). \tag{8}$$

Hence, $\sqrt{nb_n}\left[\widehat{\mu}(x;\theta)-\mu(x)\right] \xrightarrow{d} N\Big(0,\sigma^2(x)\Big)$, *with*

$$\sigma^2(x) = E\left[\left(Y^*-\mu(x)\right)^2 w^{-1}(Y^*)\Big| X^*=x\right]\frac{m(x)\int K^2(u)\,\mathrm{d}u}{f_X(x)}.$$

The proof of Theorem 1 is given in the Appendix. Let $\dot{h}$, $\dot{w}$, and $\dot{v}$ denote the respective derivatives of h, w, and v with respect to θ. Denote $F_0(t)=P[Z\le t,\delta=0]$, $F_1(t)=P[Z\le t,\delta=1]$,

$$\begin{aligned}
\chi_i^{F_1} &= \int_{\ell_L}^{u_L}\left[\frac{1}{w(t;\theta)}\chi_i^{\dot{w}}(t)-\frac{\dot{w}(t;\theta)}{w^2(t;\theta)}\chi_i^{w}(t)\right]\mathrm{d}F_1(t),\\
\chi_i^{F_0} &= \int_{\ell_L}^{u_L}\left[\frac{1}{v(t;\theta)}\chi_i^{\dot{v}}(t)-\frac{\dot{v}(t;\theta)}{v^2(t;\theta)}\chi_i^{v}(t)\right]\mathrm{d}F_0(t),\\
\chi_i^{\ell_c} &= \frac{\dot{h}(T_i;\theta)}{h(T_i;\theta)}-\delta_i\frac{\dot{w}(Z_i;\theta)}{w(Z_i;\theta)}-(1-\delta_i)\frac{\dot{v}(Z_i;\theta)}{v(Z_i;\theta)},
\end{aligned}$$

and $\phi_\theta(T_i,Z_i,\gamma_i,\delta_i)=\chi_i^{F_1}+\chi_i^{F_0}-\chi_i^{\ell_c}$, where χ_i^{w}, $\chi_i^{\dot{w}}$, χ_i^{v} and $\chi_i^{\dot{v}}$ $(i=1,\ldots,n)$ are defined in the Appendix. In the next result we establish the limiting distribution of the truncation parameter estimate $\widehat{\theta}$. This result (Theorem 2 below) along with Theorem 1 will lead to the asymptotic normality of the semiparametric regression estimator $\widehat{\mu}(.;\widehat{\theta})$ in (7).

Theorem 2. *Suppose assumptions B1, B3, and B4 hold, and let* $\gamma = R\wedge R_c$. *We have*

$$\widehat{\theta}-\theta = \Omega_\theta^{-1}\frac{1}{n}\sum_{i=1}^{n}\phi_\theta(T_i,Z_i,\gamma_i,\delta_i)+\Big(o_p\left(n^{-1/2}\right),\ldots,o_p\left(n^{-1/2}\right)\Big)^t,$$

where Ω_θ *is defined in the Appendix. Thus,* $\sqrt{n}\Big(\widehat{\theta}-\theta\Big)\xrightarrow{d} N\big(0,\Sigma_1\big)$, *with*

$$\Sigma_1 = \Omega_\theta^{-1}\,E\Big[\phi_\theta(T,Z,\gamma,\delta)\,\phi_\theta^t(T,Z,\gamma,\delta)\Big]\Omega_\theta^{-1}.$$

The proof is detailed in the Appendix. The asymptotic distributions of $\widehat{\mu}(.;\widehat{\theta})$ is stated in Corollary 1.

Corollary 1. *Under the assumptions of Theorems 1 and 2, we have for* $x\in\mathcal{A}$

$$\sqrt{nb_n}\left[\widehat{\mu}(x;\widehat{\theta})-\mu(x)\right]\xrightarrow{d} N\Big(0,\,\sigma^2(x)\Big),$$

where $\sigma^2(x)$ *is defined in Theorem 1.*

The proof of Corollary 1 is given in the Appendix. Notice that the limit distributions of $\widehat{\mu}(.;\theta)$, for a fixed value of θ, and $\widehat{\mu}(.;\widehat{\theta})$, for an estimate of θ, are the same. This is caused by the fact that $\widehat{\theta}$ has a faster convergence rate, i.e., $\mathcal{O}_p(n^{-1/2})$, than the regression function estimator $\widehat{\mu}$, i.e., $\mathcal{O}_p\big((nb)^{-1/2}\big)$. Note that the asymptotic variance of $\widehat{\mu}$

$$\sigma^2(x)=\int\Big(y-\mu(x)\Big)^2 w^{-1}(y)\,\mathrm{d}F_{Y^*|X^*}(y|x)\times\frac{m(x)\int K^2(u)\,\mathrm{d}u}{f_X(x)},$$

can be estimated by replacing θ, w, $F_{Y^*|X^*}$ and μ by their respective estimators $\widehat{\theta}$, $\widehat{w}(.;\widehat{\theta})$, $\widehat{F}_{Y^*|X^*}$ and $\widehat{\mu}(.;\widehat{\theta})$ (defined in Section 2.2), and f_X by $\widehat{f}_X(x)$ (defined in Theorem 1). Note that $m(x)$ can be estimated by $\int_{y>0} H(y;\widehat{\theta})\,\mathrm{d}\widehat{F}_{Y^*|X^*}(y|x)$, K is known, and b can be substituted by $\widehat{b}_{opt}$ defined in (12). The asymptotic variance of $\widehat{\theta}$ in Theorem 2 involves several complicated terms, so much so that it renders a direct estimation of Σ_1 particularly difficult. In this case, we suggest using bootstrap methods to estimate Σ_1.

3. THE LENGTH-BIAS CASE

Analogously to the approach described in Section 2.2, and following similar steps, we introduce an estimator for the regression function μ when the data are length-biased and are subject to random censoring. Under uniform truncations, the local pseudo log-likelihood in (3) becomes

$$\ell^u_{lpl} = \sum_{i=1}^n K_b(x-X_i)\,\delta_i\,\log\left\{\frac{\mathrm{d}F_{Y^*|X^*}(z_i|x)}{\int_{t>0} w_u(t)\,\mathrm{d}F_{Y^*|X^*}(t|x)}\right\},$$

with nonparametric maximizer

$$\widehat{p}^u_i = \frac{K_b(x-X_i)\,w_u^{-1}(Z_i)\,\delta_i}{\sum_{j=1}^n K_b(x-X_j)\,w_u^{-1}(Z_j)\,\delta_j}, \tag{9}$$

where $w_u(z) = \int_0^z \left[1-G(t)\right]\mathrm{d}t$. The weight function w_u in (9), which replaces w of Section 2.2, can be estimated by $\widehat{w}_u(z) = \int_0^z \left[1-\widehat{G}(t)\right]\mathrm{d}t$, with $\widehat{G}$ the K-M estimator of G. Thus, we estimate the conditional distribution $F_{Y^*|X^*}$ by

$$\widehat{F}_{Y^*|X^*}\left(y|x\right) = \sum_{i=1}^n \frac{K_b(x-X_i)\,\widehat{w}_u^{-1}(Z_i)\,\delta_i}{\sum_{j=1}^n K_b(x-X_j)\,\widehat{w}_u^{-1}(Z_j)\,\delta_j}\,\mathrm{I}(Z_i \le y), \tag{10}$$

and the regression function μ by

$$\widehat{\mu}_u\left(x\right) = \sum_{i=1}^n \frac{K_b(x-X_i)\,\widehat{w}_u^{-1}(Z_i)\,\delta_i}{\sum_{j=1}^n K_b(x-X_j)\,\widehat{w}_u^{-1}(Z_j)\,\delta_j}\,Z_i. \tag{11}$$

The estimator in (10) generalizes the estimator of de Uña-Àlvarez & Iglesias-Pérez (2010), from fixed residual-censoring to random censoring. The authors introduced a general estimator in Remark 1 (of their paper); however, they did not study its consistency nor its asymptotic theory. Their estimator is different from ours. Unlike our case, they consider possible dependency between the residual-censoring R_c and the observed covariate X. Next, we establish the limit distribution of $\widehat{\mu}_u$.

Theorem 3. *Under assumptions B1, B2, K1, and K2, $\widehat{\mu}_u(x)$, for $x \in \mathcal{A}$, admits the representation*

$$\widehat{\mu}_u(x) - \mu(x) = \frac{1}{nb_n}\sum_{i=1}^n K_b(x-X_i)\frac{\delta_i}{w_u(Z_i)}\left[Z_i - \mu(x)\right] \times \frac{\mu(x)}{\widehat{f}_X(x)} + o_p\left(\frac{1}{\sqrt{nb_n}}\right).$$

Therefore, $\sqrt{nb_n}\left[\widehat{\mu}_u(x)-\mu(x)\right] \xrightarrow{d} N\left(0,\sigma_u^2(x)\right)$, *with*

$$\sigma_u^2(x) = E\left[\left(Y^*-\mu(x)\right)^2 w_u^{-1}(Y^*)\middle| X^*=x\right] \frac{\mu(x)\int K^2(u)\,\mathrm{d}u}{f_X(x)}.$$

The proof is similar to that of Theorem 1. In the following result, we show the efficiency of $\widehat{F}_{Y^*|X^*}$ in (10) as compared with the product-limit estimator of Iglesias-Pérez & González-Manteiga (1999) for unspecified truncation cdf H. The limit variance of the authors' estimator is

$$S^2(x,y) = \overline{F}^2_{Y^*|X^*}(y|x)\int_0^y C_x^{-2}(t)\,\mathrm{d}F_{Z,\delta|X}(t,1|x)\,\frac{\int K^2(u)\,\mathrm{d}u}{f_X(x)},$$

where $C_x(t)=P[T\le t\le Z|X=x]$; the proportion of recruited subjects who are at risk to fail at time t given $X=x$.

Proposition 1. *Suppose B1, B2, K1 and K2 hold. We have for* $(x,y)\in\mathcal{A}\times\mathcal{B}$,

1. $\sqrt{nb_n}\left[\widehat{F}_{Y^*|X^*}(y|x)-F_{Y^*|X^*}(y|x)\right] \xrightarrow{d} N\left(0,V_u(x,y)\right)$, *with*

$$V_u(x,y) = E\left[\left(\mathrm{I}(Y^*\le y)-F_{Y^*|X^*}(y|x)\right)^2 w_u^{-1}(Y^*)\middle| X^*=x\right] \frac{\mu(x)\int K^2(u)\,\mathrm{d}u}{f_X(x)}.$$

2. $V_u(x,y)\le S^2(x,y)$.

The proof of Proposition 1 is given in the Appendix. Note that the estimation of the variances σ_u^2 in Theorem 3 and V_u in Proposition 1 is similar to that of σ^2, as discussed in Section 2.3.

4. BANDWIDTH SELECTION

One critical issue to the estimation process of the regression function μ is the choice of the smoothing parameter b. For right-censored and left-truncated data, there is a need for a selection method that accounts for both truncation and censoring mechanisms. To choose the bandwidth parameter, we may minimize with respect to b the weighted integrated squared error

$$\mathrm{WISE}(b) = \int_{x>0}\left[\widehat{\mu}^*(x;b)-\mu(x)\right]^2 \mathrm{d}F_{X^*}(x),$$

where $\widehat{\mu}^*(x;b)=\widehat{\mu}(x;\widehat{\theta})$ if $H\in\mathcal{F}_\theta$ (with unknown θ) and $\widehat{\mu}^*(x;b)=\widehat{\mu}_u(x)$ for uniform truncations. This is equivalent to choosing b that minimizes

$$\mathrm{WISE}_*(b) = \int_{x>0}\widehat{\mu}^{*\,2}(x;b)\,\mathrm{d}F_{X^*}(x) - 2\int_{x>0}\int_{y>0} y\,\widehat{\mu}^*(x;b)\,\mathrm{d}F_{X^*,Y^*}(x,y),$$

by noticing that $\mu(x)\,\mathrm{d}F_{X^*}(x)=\int_{y>0} y\,\mathrm{d}F_{X^*,Y^*}(x,y)$. It remains to replace the unknown bivariate distribution F_{X^*,Y^*} and its marginal F_{X^*} in the expression of $\mathrm{WISE}_*(b)$ by their estimates. One can show that F_{X^*,Y^*} and F_{X^*} can consistently be estimated by

$$\widehat{F}_{X^*,Y^*}(x,y) = \sum_{i=1}^n \frac{\widehat{w}_*^{-1}(Z_i)\delta_i}{\sum_{j=1}^n \widehat{w}_*^{-1}(Z_j)\delta_j}\,\mathrm{I}(X_i\le x,\, Z_i\le y),$$

$$\widehat{F}_{X^*}(x) = \sum_{i=1}^{n} \frac{\widehat{w}_*^{-1}(Z_i)\delta_i}{\sum_{j=1}^{n} \widehat{w}_*^{-1}(Z_j)\delta_j} \, \mathrm{I}(X_i \leq x),$$

respectively, where $\widehat{w}_* = \widehat{w}(.;\widehat{\theta})$ if $H \in \mathcal{F}_\theta$ and $\widehat{w}_* = \widehat{w}_u$ if H is uniform. The derivation of $\widehat{F}_{X^*,Y^*}$, under uniform H, have been studied in Rabhi & Bouezmarni (2020). The data driven bandwidth is then

$$\widehat{b}_{opt} = \arg\min_b \left\{ \sum_{i=1}^{n} \frac{\widehat{w}_*^{-1}(Z_i)\delta_i}{\sum_{j=1}^{n} \widehat{w}_*^{-1}(Z_j)\delta_j} \left[\widehat{\mu}^*_{-i}(X_i; b) - Z_i\right]^2 \right\}, \tag{12}$$

where $\widehat{\mu}^*_{-i}$ is a leave-one-out estimate of μ at x, given by

$$\widehat{\mu}^*_{-i}(x; b) = \sum_{\substack{k=1\\k\neq i}}^{n} \frac{K_b(x - X_k)\, \widehat{w}_*^{-1}(Z_k)\, \delta_k}{\sum_{\substack{j=1\\j\neq i}}^{n} K_b(x - X_j)\, \widehat{w}_*^{-1}(Z_j)\, \delta_j} Z_k.$$

Note that the weights $\widehat{w}_*^{-1}(Z_i)\delta_i / \sum_{j=1}^{n} \widehat{w}_*^{-1}(Z_j)\delta_j$ in (12) account for the truncation and censoring mechanisms, via the weight function $\widehat{w}_*$, and replace the uniform weight $1/n$ in the empirical version of $\widehat{h}_{opt}$ for complete data (see Härdle & Marron, 1985).

5. DATA ANALYSIS

5.1. Simulation Study

The following simulation study is carried out to evaluate the finite-sample efficiency performance of the semiparametric regression estimator $\widehat{\mu}(.;\widehat{\theta})$ in (7) as compared with the nonparametric estimator of Iglesias-Pérez & González-Manteiga (1999) for unspecified H. We study the integrated squared error (ISE) of the latter authors' estimator (denoted P-M) and that of $\widehat{\mu}$, given by

$$\mathrm{ISE}(\widehat{\mu}) = \int_{x>0} \left[\widehat{\mu}(x;\widehat{\theta}) - \mu(x)\right]^2 \mathrm{d}x,$$

for various truncation models, under two levels of censoring 25% and 40%. We consider two regression models for the failure time Y^*. The homoscedastic regression model

$$Y^* = \mu_1(X^*) + \varepsilon, \text{ with } \mu_1(x) = 2.9 - 0.25\,x,$$

where $X^* \sim \mathrm{Unif}[6, 10]$ and $\varepsilon \sim \mathrm{Unif}[-0.3, 0.3]$, and the nonlinear regression model

$$Y^* = \mu_2(X^*) + \varepsilon, \text{ with } \mu_2(x) = x^2 + 0.25\,x,$$

where $X^* \sim \mathrm{Unif}[1, 3]$ and $\varepsilon \sim \mathrm{Unif}[-1, 1]$. In the simulation experiment, we generate independently the lifetime data y_i^* $(i = 1, 2, \dots)$ from the aforementioned models and the truncation data t_i^* $(i = 1, 2, \dots)$ from a distribution $H(.;\theta)$. We truncate (exclude) the triple (y_i^*, t_i^*, x_i^*) if $y_i^* < t_i^*$, otherwise we keep generating the data until n observations are collected. The resulting sample is $\{(x_i, y_i, t_i), i = 1, \dots, n\}$. Define $r_i = y_i - t_i$ $(i = 1, \dots, n)$. We then independently generate n residual censoring data $r_{c,1}, \dots, r_{c,n}$ from a gamma distribution G. This allows us to define the data $\{(x_i, t_i, z_i, \delta_i),\ i = 1, \dots, n\}$, where $z_i = t_i + \min(r_i, r_{c,i})$ and $\delta_i = \mathrm{I}\left(r_i \leq r_{c,i}\right)$.

Tables 1 and 2 show that the new semiparametric estimator $\widehat{\mu}(.;\widehat{\theta})$, which incorporates the available information of the truncation distribution H, outperforms the full nonparametric

TABLE 1: Comparative efficiency between the semiparametric estimator $\widehat{\mu}$ and the P-M estimator (based on ISE), for various truncation models in the homoscedastic regression ($\times 10^{-2}$).

		$n = 100$		$n = 200$		$n = 400$	
$F_{T^*}(.;\theta)$	Censoring	$\widehat{\mu}(.;\widehat{\theta})$	P-M	$\widehat{\mu}(.;\widehat{\theta})$	P-M	$\widehat{\mu}(.;\widehat{\theta})$	P-M
Gamma(2, 0.5)	25%	1.94	2.86	1.11	1.74	0.607	0.757
	40%	2.11	3.30	1.19	1.82	0.645	0.804
Weibull(2, 5)	25%	2.24	4.40	1.25	1.73	0.724	0.896
	40%	2.39	4.67	1.33	1.84	0.777	0.945
logNorm(1, 1)	25%	2.15	3.94	1.23	1.93	0.727	1.380
	40%	2.27	4.23	1.31	2.08	0.779	1.440
logLogit(4, 2)	25%	2.16	3.84	1.19	1.72	0.690	0.882
	40%	2.40	4.36	1.32	1.89	0.758	0.948
Gompertz(.05, 2)	25%	1.79	2.41	0.97	1.22	0.563	0.654
	40%	1.97	2.75	1.08	1.35	0.613	0.708

TABLE 2: Comparative efficiency between the semiparametric estimator $\widehat{\mu}$ and the P-M estimator (based on ISE), for various truncation models in the nonlinear regression.

		$n = 100$		$n = 200$		$n = 400$	
$F_{T^*}(.;\theta)$	Censoring	$\widehat{\mu}(.;\widehat{\theta})$	P-M	$\widehat{\mu}(.;\widehat{\theta})$	P-M	$\widehat{\mu}(.;\widehat{\theta})$	P-M
Gamma(5.5, 0.5)	25%	4.63	6.90	2.66	4.60	1.46	2.88
	40%	4.79	7.00	2.70	4.66	1.48	3.04
10.Weibull(2, 5)	25%	0.66	1.45	0.22	0.74	0.11	0.28
	40%	0.70	1.78	0.27	0.76	0.13	0.30
logNorm(3, 1)	25%	0.70	1.85	0.27	1.18	0.14	0.62
	40%	0.73	1.90	0.29	1.20	0.15	0.64
logLogit(20, 2)	25%	0.54	1.28	0.19	0.50	0.10	0.25
	40%	0.58	1.34	0.20	0.52	0.12	0.28

estimator (P-M) of Iglesias-Pérez & González-Manteiga (1999), which leaves H unspecified, in all cases considered in our simulation. As expected the average ISE decreases with increasing n, the sample size, and increases with as the censoring rate increases. The efficiencies of the estimators are based on the average of 1,000 simulations of ISE.

In Table 3, the simulated ratio of the mean-integrated-squared-errors of $\widehat{\mu}(.;\widehat{\theta})$, MISE(Misspecified truncation model)/MISE(True truncation model), indicates that we do not lose much efficiency when the truncation-model used is either Weibull, gamma, Gompertz or log-normal, while the true model is one of the first three models. However, when the true truncation-model is log-normal, there is a relative loss in efficiency if the model used is different from the later. Note that the MISE is computed from the average of 1,000 simulations of ISE.

TABLE 3: Ratio of the mean integrated squared errors of $\widehat{\mu}$, MISE(misspecified truncation model)/MISE(true model). Sample size $n = 400$ and censoring level = 25%.

	True truncation model			
Model used	Weibull	Gamma	Gompertz	Log-normal
Weibull	1	1.003	1.011	2.521
Gamma	1.025	1	1.023	3.472
Gompertz	1.030	1.005	1	2.765
log-Normal	1.080	1.084	1.037	1

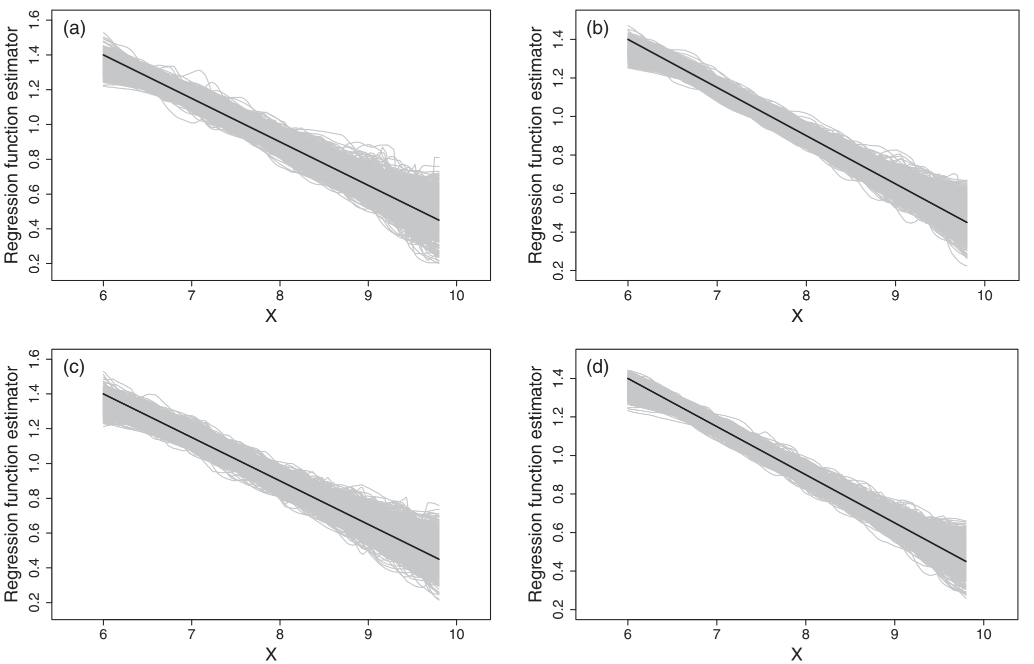

FIGURE 1: One thousand simulated curves of the estimator $\widehat{\mu}(.;\widehat{\theta})$ (grey) and the homoscedastic regression function (black). (a and b) Weibull truncations, (c and d) gamma truncations. (a and c) $n = 200$, (b and d) $n = 400$ (censoring level = 40%).

The plots in Figures 1 and 2 display 1,000 simulated curves of the estimator $\widehat{\mu}(.;\widehat{\theta})$ (grey) and the true regression function (black), in the homoscedastic and nonlinear models, for several truncation distributions. Note that the computational cost of the length-biased estimator (11) is similar to the cost of the estimator of Iglesias-Pérez & González-Manteiga (1999). However, for the estimator given by Equation A7, the computational cost is a bit higher. This is due to the optimization step of estimating the parameter θ. We pay this price, of course, for the gain in efficiency as indicated in Tables 1 and 2. In this article, we used the kernel function $K(x) = 0.75(1 - x^2)\,\mathrm{I}_{[-1,1]}(x)$ and the formula (12) to select the bandwidth b_n for our estimator $\widehat{\mu}(.;\widehat{\theta})$. For the P-M estimator $\widehat{\mu}_{PM}$, we chose the bandwidth that minimizes the ISE of $\widehat{\mu}_{PM}$. Note

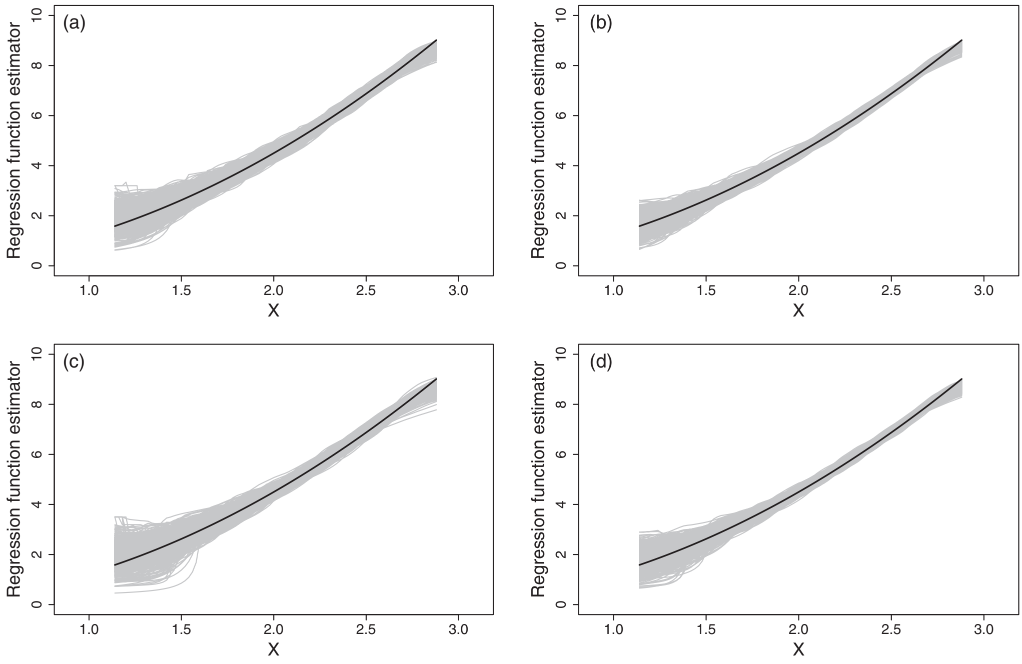

FIGURE 2: One thousand simulated curves of the estimator $\hat{\mu}(.;\hat{\theta})$ (grey) and the nonlinear regression function (black). (a and b) Log-logistic truncations, (c and d) log-normal truncations. (a and c) $n = 200$, (b and d) $n = 400$ (censoring level = 25%).

that we computed the ISE of $\hat{\mu}(.;\hat{\theta})$ and $\hat{\mu}_{PM}$ over the intervals $[6+b, 10-b]$ and $[1+b, 3-b]$ in the homoscedastic and nonlinear models, respectively, and we estimated μ in those regions.

5.2. HIV-Infection and AIDS Data

We illustrate the semiparametric estimation method in Section 2 on a set of prevalent cohort data from the Amsterdam Cohort Study on HIV infection and AIDS (see Geskus, 2000). The data were, initially, collected among men who have sex with men (MSM) who have experienced HIV-infection prior to study recruitment and had not developed AIDS. Those who developed AIDS before the recruitment were not a part of the study, as discussed in Geskus (2000). Ascertaining the date of seroconversion is a source of difficulty in AIDS studies. In most cohort studies on HIV infection, the seroconversion dates can be left-, right- or interval-censored. While for those who are prospectively identified we can consider the midpoint of the date of the last seronegative test and the first seropositive test as the date of seroconversion, for seroprevalent cases the situation is different. Geskus (2000) has thoroughly studied this issue and presented a marker-based approach, using CD4 counts, for imputing missing dates of seroconversion for such cases.

In this analysis, we considered a sample of $n = 204$ individuals who have been infected by HIV before the beginning of the follow-up (prevalent cases). The response Y^* is defined as the time from HIV-infection to AIDS, the covariate X^* is age at HIV-infection and the left-truncation T^* is the time between HIV-infection and study recruitment. Among 204 patients, the responses (Y^*) of 57 individuals were right-censored by the end of study.

The purpose of the present example is to study the impact of the factor age at HIV-infection on the response Y^*. First, we used the nonparametric method of Wang (1991) to estimate

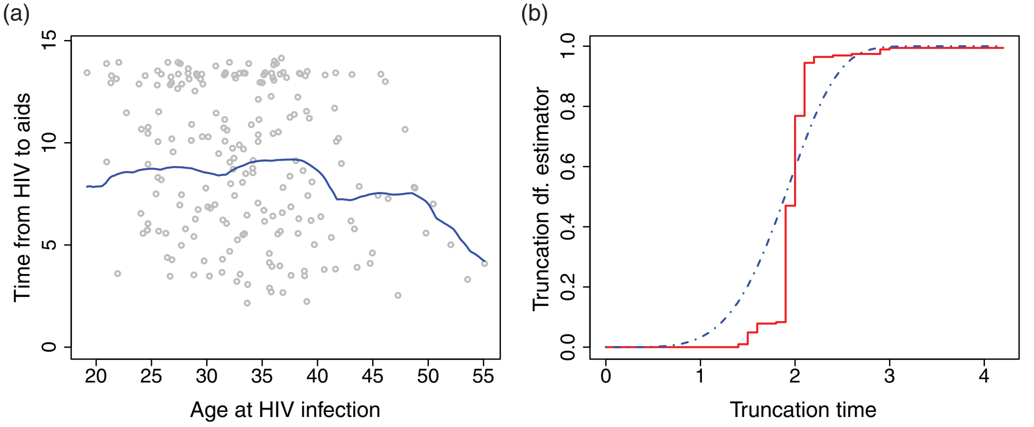

FIGURE 3: (a) Semiparametric regression estimator $\widehat{\mu}(.;\widehat{\theta})$ for 204 patients with HIV-infection under Weibull truncation model (blue line). (b) Nonparametric estimator (red line) and parametric estimator based on Weibull family (blue dashed) of H.

the truncation distribution. Figure 3(b) shows this estimator (red curve) and indicates an exponential trend between 1.5 and 2.5 years. The Weibull distribution Weibull(θ_1, θ_2) (blue dashed) is well supported by the nonparametric estimator (red line), where (θ_1, θ_2) is estimated by the conditional MLE, defined in (6), with value $(\widehat{\theta}_1, \widehat{\theta}_2) = (4.80, 2.042)$. We then used the semiparametric estimator $\widehat{\mu}(., \widehat{\theta})$ based on the Weibull truncation model. Note that we employed the kernel $K(x) = 0.75(1 - x^2)\,\mathrm{I}_{[-1,1]}(x)$ and the bandwidth $\widehat{b} = 3.2$, using formula (12).

Figure 3(a) displays the curve (blue line) of the semiparametric estimator $\widehat{\mu}(., \widehat{\theta})$. The regression function estimator is almost constant in the range $[20, 40]$ years, but after that, it starts decreasing. This indicates that individuals who have been infected by HIV at an older age are more likely to develop AIDS than those infected at a younger age. Some caution is needed for the interpretation of this downtrend, as it may be due to some boundary effect.

6. SUMMARY AND CONCLUSIONS

In this article, we developed a new methodology for estimating the regression function $\mu(x) = E[Y^* | X^* = x]$ under biased sampling and random censoring. The proposed approach is semiparametric. In the setting of the article, we considered a nonparametric model for the response-covariate (Y^*, X^*) (variables of interest) and a parametric model for the truncation distribution. The main contribution of our article is to devise a new estimation procedure that uses the information in the parametric form of the truncation distribution to increase efficiency.

We approach this question by first estimating the conditional cdf $F_{Y^*|X^*=x}(y)$. To this end, we take a likelihood approach by considering the log-likelihood ℓ^*_{pl} of the data $\{(Z_i, T_i) \,|\, (\delta_i = 1, X_i = x),\ i = 1, \ldots, n\}$. More specifically, we employed a weighted version of ℓ^*_{pl}, given by ℓ^*_{lpl} in (3), where the weights are chosen using a kernel approach, which is the reason for using the word "local". The estimator of p_i (the mass attached to z_i given $X = x$) is then obtained by maximizing ℓ^*_{lpl} (see Eq. 4). Given this estimator involves the truncation parameter θ and the censoring distribution G (through w), we estimated G by the Kaplan-Meier estimator $\widehat{G}$ and used the likelihood $\mathcal{L}_c$ in Equation (A5) to estimate θ, and hence, we obtain the estimators of $F_{Y^*|X^*=x}(y)$ and $\mu(x)$.

Several examples are discussed and assessed, on the basis of which we conclude that our semiparametric estimator outperforms classical estimators for unspecified truncation distribution (e.g., Iglesias-Pérez & González-Manteiga, 1999). It should be noted that we employed a standard kernel approach in our estimation procedure. If interest is directed toward boundary regions (which is not the case in our article), we recommend using boundary kernels for regression function estimation. For instance, local-linear kernel smoothing (see Fan & Gijbels, 1996). However, there is a price to pay for using these kernels. Such methods could lead to negative values for the estimation of the regression function even if it's positive. Our asymptotic results are still valid if we use local-linear kernel smoothing. Note that the methodology and theorems can be extended to kernel regression with multiple predictors. The easiest way to achieve this is by using product kernel functions. Effective implementation of the resultant methods, however, requires much larger sample sizes. This is so because the devised estimators converge much more slowly to their target. In such case, parametric and semiparametric approaches are more appropriate when X is a vector, especially if it is of high-dimension.

APPENDIX

Notations

We introduce, and recall, some notations used in the proofs below. For a general distribution F, l_F and u_F represent the respective lower and upper bounds of the support of F, with $\overline{F} = 1 - F$. We denote by $\dot{f}$ and $\ddot{f}$ the first and second derivatives of a function $f(x;\theta)$ with respect to θ, respectively, and by $f^{(i)}$ the i^{th} derivative of $f(x;\theta)$ with respect to x. Let $F_0(t) = P[Z \le t, \delta = 0]$, $F_1(t) = P[Z \le t, \delta = 1]$, $F_{Z|X}(y|x) = P[Z \le y|X = x]$, $L^G(t) = P[\gamma \le t]$, $L_0^G(t) = P[\gamma \le t, \delta = 0]$ and

$$\chi_i^G(t) = \overline{G}(t)\left[\frac{\mathrm{I}(\gamma_i \le t, \delta_i = 0)}{\overline{L^G}(\gamma_i)} - \int_0^{t\wedge\gamma_i} \frac{\mathrm{d}L_0^G(t)}{\overline{L^G}^2(t)}\right] \quad (i = 1, \dots, n),$$

the i.i.d. random term in the representation of the KM-estimator $\widehat{G}$ (see Lo, Mack & Wang, 1989). Denote for $i = 1, \dots, n$

$$\chi_i^w(t) = \int_0^t \chi_i^G(u)\,\mathrm{d}H(t-u;\theta),$$

$$\chi_i^{\dot{w}}(t) = \int_0^t \chi_i^G(u)\,\mathrm{d}H(t-u;\theta),$$

$$\chi_i^v(t) = \chi_i^G(t)\,h(0;\theta) - \int_0^t \chi_i^G(u)\,\mathrm{d}h(t-u;\theta),$$

$$\chi_i^{\dot{v}}(t) = \chi_i^G(t)\,\dot{h}(0;\theta) - \int_0^t \chi_i^G(u)\,\mathrm{d}\dot{h}(t-u;\theta),$$

$$\chi_i^{F_1} = \int_{\ell_L}^{u_L}\left[\frac{1}{w(t;\theta)}\chi_i^{\dot{w}}(t) - \frac{\dot{w}(t;\theta)}{w^2(t;\theta)}\chi_i^w(t)\right]\mathrm{d}F_1(t), \tag{A1a}$$

$$\chi_i^{F_0} = \int_{\ell_L}^{u_L}\left[\frac{1}{v(t;\theta)}\chi_i^{\dot{v}}(t) - \frac{\dot{v}(t;\theta)}{v^2(t;\theta)}\chi_i^v(t)\right]\mathrm{d}F_0(t), \tag{A1b}$$

$$\chi_i^{\ell_c} = \frac{\dot{h}(T_i;\theta)}{h(T_i;\theta)} - \delta_i\frac{\dot{w}(Z_i;\theta)}{w(Z_i;\theta)} - (1-\delta_i)\frac{\dot{v}(Z_i;\theta)}{v(Z_i;\theta)}, \tag{A1c}$$

$$\phi_\theta(T_i, Z_i, \gamma_i, \delta_i) = \chi_i^{F_1} + \chi_i^{F_0} - \chi_i^{\ell_c}. \tag{A1d}$$

$$\Omega_\theta = E\left[\left(\frac{\ddot{h}(T;\theta)}{h(T;\theta)} - \frac{\dot{h}(T;\theta)\,\dot{h}^t(T;\theta)}{h(T;\theta)\,h(T;\theta)}\right) - \delta\left(\frac{\ddot{w}(Z;\theta)}{w(Z;\theta)} - \frac{\dot{w}(Z;\theta)\,\dot{w}^t(Z;\theta)}{w(Z;\theta)\,w(Z;\theta)}\right) - (1-\delta)\right.$$
$$\left.\times\left(\frac{\ddot{v}(Z;\theta)}{v(Z;\theta)} - \frac{\dot{v}(Z;\theta)\,\dot{v}^t(Z;\theta)}{v(Z;\theta)\,v(Z;\theta)}\right)\right]. \tag{A1e}$$

Assumptions

The following regularity conditions are needed to establish the asymptotic results in this article. Let Q be the cdf of the residual lifetime R, and l_F and u_F denote the lower and bounds of the support of a general distribution F.

Assumption B1:

(i) $l_G \wedge l_Q > 0$ and $l_H > 0$.
(ii) $u_G \leq u_Q$ with $G(u_G) < 1$.
(iii) $0 < l_L < u_L < \infty$.

The first part of Assumption (i) essentially means that there is no immediate failure or censoring at the beginning of the study, while the second part of the assumption means that all subjects recruited to the study are prevalent cases. This condition reflects the settings of the HIV-AIDS data and in general is reasonable in most prevalent cohort studies. Assumption (ii) means that the lifetimes of some individuals, who are still alive at the end of the study, will be censored. This is common in the follow-up studies and is due to the limited time of the follow-up. The condition $l_L > 0$ is a direct consequence of $l_H > 0$, while $u_L < \infty$ means that the observed lifetime of individual is finite.

Assumption B2: Let f_{X^*} be the density of F_{X^*}.

(i) f_{X^*} is bounded and bounded away from zero in its compact support $\mathcal{A}$.
(ii) If $[a, c]$ is contained in the support of f_X, then

$$0 < \beta_1 = \inf_{x\in \mathrm{I}_\alpha}\left\{f_X(x)\right\} < \sup_{x\in \mathrm{I}_\alpha}\left\{f_X(x)\right\} = \Gamma < \infty,$$

where $\mathrm{I}_\alpha = [a-\alpha, c+\alpha]$ and $0 < \alpha\Gamma < 1$. For $y \in [l_L, u_L]$, $\inf_{x\in\mathrm{I}_\alpha}\left\{F_{Z|X}(y|x)\right\} > \beta_2 > 0$.
(iii) For $x \in \mathrm{I}_\alpha$, $f_X(x)$, $x \to F_{Z|X}(y|x)$ and $x \to F_{Z,\delta|X}(y, 1|x)$ have bounded continuous second derivatives and $f_X^{(2)}(x)$ is bounded away from zero.

Assumptions B2(ii,iii) (see Dabrowska, 1989) are needed for the uniform convergence rate of $\widehat{F}_{Z,\delta|X}(y, 1|x)$. The first part of B2(ii), which we only need in the proofs, is a consequence of B2(i). The regularity condition B3, below, is required for the asymptotic representation and normality of $\sqrt{n}\left(\widehat{\theta} - \theta\right)$.

Assumption B3:

(i) H (with density h) and G are continuous.
(ii) $h(x;\theta)$ is bounded and continuously differentiable in x and in θ. $h^{(1)}(x;\theta)$ is differentiable in θ and continuous in x. $\ddot{h}(x;\theta)$ exists and is continuous in θ.

Assumption B4: Let $\Psi_n(\theta) = n^{-1}\partial \log(\mathcal{L}_c^*(\theta))/\partial\theta$ and $\Psi(\theta) = E[\Psi_n(\theta)]$.

(i) Ψ and Ψ_n, for sufficiently large n, have only one zero, at θ_0 and $\widehat{\theta}_0$, respectively.
(ii) $\det(\Omega_\theta) > 0$.

Assumption B4(i) is essentially identifiability (see Wang, 1989; Andersen, 1970), while B4(ii) is essentially positive-definiteness of the information matrix. The latter condition rarely fails to hold provided that identifiability and some smoothness conditions hold (see Theorem 3 in Asgharian, 2014). The kernel function K and the bandwidth b_n satisfy the conditions

Assumptions K1, K2:

K1: K is supported on $(-1, 1)$, of bounded total variation and $\int t\,K(t)\,\mathrm{d}t = 0$.
K2: $\log(n)/nb_n = o(1)$ and $nb_n^5/\log(n) = O(1)$, as $n \to \infty$ and $b_n \to 0$.

Proofs

Proof of Theorem 1. Denote $\xi(y,1|x) = F_{Z,\delta|X}(y,1|x)$, $\hat{\xi} = \hat{F}_{Z,\delta|X}(y,1|x)$, $w(x) = w(x;\theta)$ and recall that

$$\hat{\mu}(x;\theta) = \frac{\int_{y>0} y\,\hat{w}^{-1}(y;\theta)\,\mathrm{d}\hat{\xi}(y,1|x)}{\int_{y>0} \hat{w}^{-1}(y;\theta)\,\mathrm{d}\hat{\xi}(y,1|x)}.$$

By using the uniform results $\|\hat{w}(.;\theta) - w(.)\| = \mathcal{O}_{a.s.}\left(\sqrt{n^{-1}\log\log n}\right)$, which follows from $\|\hat{G} - G\| = \mathcal{O}_{a.s.}\left(\sqrt{n^{-1}\log\log n}\right)$, and $\|\hat{\xi} - \xi\| = \mathcal{O}_{a.s.}\left(\sqrt{(nb)^{-1}\log n}\right)$, established by Dabrowska (1989) in Theorem 2.1's proof and Corollary 2.2 (see also Theorem 1 in Iglesias-Pérez & González-Manteiga, 1999), the difference $\hat{\mu} - \mu$ can be written as

$$\begin{aligned}\hat{\mu}(x;\theta) - \mu(x) = {}& \frac{1}{\int_{y>0} w^{-1}(y)\,\mathrm{d}\xi(y,1|x)}\left\{-\int_{y>0}\frac{y}{w^2(y)}\left[\hat{w}(y;\theta) - w(y)\right]\mathrm{d}\xi(y,1|x)\right.\\ &\left.+\int_{y>0}\frac{y}{w(y)}\,\mathrm{d}\left[\hat{\xi}(y,1|x) - \xi(y,1|x)\right]\right\}\\ &+\frac{\int_{y>0} y\,w^{-1}(y)\,\mathrm{d}\xi(y,1|x)}{\left[\int_{y>0} w^{-1}(y)\,\mathrm{d}\xi(y,1|x)\right]^2}\left\{\int_{y>0}\frac{1}{w^2(y)}\left[\hat{w}(y;\theta) - w(y)\right]\mathrm{d}\xi(y,1|x)\right.\\ &\left.-\int_{y>0}\frac{1}{w(y)}\,\mathrm{d}\left[\hat{\xi}(y,1|x) - \xi(y,1|x)\right]\right\} + r_n'(x),\end{aligned}$$

where $\sup_x |r_n'(x)| = o_{a.s.}\left(\left(nb\right)^{-1/2}\right)$. Now, by employing the fact that $\hat{w}(y;\theta) - w(y) = \mathcal{O}_p\left(n^{-1/2}\right)$ and $\mu(x) = \int_{y>0} y\,w^{-1}(y)\,\mathrm{d}\xi(y,1|x)/\int_{y>0} w^{-1}(y)\,\mathrm{d}\xi(y,1|x)$, we find

$$\hat{\mu}(x;\theta) - \mu(x) = m(x)\left\{\int_{y>0}\frac{y}{w(y)}\,\mathrm{d}\hat{\xi}(y,1|x) - \mu(x)\int_{y>0}\frac{1}{w(y)}\,\mathrm{d}\hat{\xi}(y,1|x)\right\} + r_n'(x) + r_n''(x)$$

where $m(x) = \left[\int_{y>0} w^{-1}(y)\,\mathrm{d}\xi(y,1|x)\right]^{-1}$ and $r_n''(x) = \mathcal{O}_p\left(n^{-1/2}\right)$. Hence, we deduce representation (8). The limit distribution of $\sqrt{nb}\left[\hat{\mu}(x;\theta) - \mu(x)\right]$ follows from (8) by using the Lindberg-Feller's CLT Theorem and Slutsky's Theorem. ∎

Proof of Theorem 2. (1) The following proof is given for a uni-dimensional parameter θ, i.e., $d = 1$, the proof is similar for $d \geq 2$. To show the consistency of $\widehat{\theta}$, we adopt similar regularity conditions to the conditional-approach-parameter-estimation of Wang (1989) and Andersen (1970) (assumption B4(i)). Assumption B4(i) is a more relaxed condition than the one used by Wang (1989). Let $\Psi_n(\theta_0) = (\partial/\partial\theta_0)\log(\mathcal{L}_c^*(\theta_0))/n$ and $\Psi(\theta_0) = E[\Psi_n(\theta_0)]$. The function $\theta_0 \to \Psi_n(\theta_0)$ is continuous and $\Psi_n(\theta_0) \xrightarrow{P} \Psi(\theta_0)$ for every θ_0. By Lemma 5.10 in van der Vaart (1998) (p. 47) and assumption B4(i), $\widehat{\theta} \xrightarrow{P} \theta$.

(2) Let $\widehat{F}_0(t)$ and $\widehat{F}_1(t)$ be the respective empirical counterparts of $F_0(t) = P[Z \leq t, \delta = 0]$ and $F_1(t) = P[Z \leq t, \delta = 1]$, and $\widehat{\mathbb{F}}_k(t) = \widehat{F}_k(t) - F_k(t)$ for $k = 0, 1$. Denote $w(x) = w(x;\theta)$, $v(x) = v(x;\theta)$,

$$\ell_c(\theta_0, G) = n^{-1}\log(\mathcal{L}_c(\theta_0)) = \frac{1}{n}\log\left\{\prod_{i=1}^{n}\frac{\mathrm{d}H(t_i;\theta_0)}{w\left(z_i;\theta_0\right)^{\delta_i} v\left(z_i;\theta_0\right)^{1-\delta_i}}\right\},$$

$\ell_c(\theta_0, \widehat{G}) = n^{-1}\log(\mathcal{L}_c^*(\theta_0))$ and, respectively, by $\dot{\ell}_c$ and $\ddot{\ell}_c$ the first and second derivatives of ℓ_c with respect to θ_0. Using a first order Taylor expansion, we have

$$\theta - \widehat{\theta} = \left[\ddot{\ell}_c(\theta^*, \widehat{G})\right]^{-1}\left\{\dot{\ell}_c(\theta, G) - \int_{\ell_L}^{u_L}\left(\frac{\dot{\widehat{w}}(t;\theta)}{\widehat{w}(t;\theta)} - \frac{\dot{w}(t)}{w(t)}\right)\mathrm{d}F_1(t) - \int_{\ell_L}^{u_L}\left(\frac{\dot{\widehat{v}}(t;\theta)}{\widehat{v}(t;\theta)} - \frac{\dot{v}(t)}{v(t)}\right)\mathrm{d}F_0(t)\right.$$
$$\left. - \int_{\ell_L}^{u_L}\left(\frac{\dot{\widehat{w}}(t;\theta)}{\widehat{w}(t;\theta)} - \frac{\dot{w}(t)}{w(t)}\right)\mathrm{d}\widehat{\mathbb{F}}_1(t) - \int_{\ell_L}^{u_L}\left(\frac{\dot{\widehat{v}}(t;\theta)}{\widehat{v}(t;\theta)} - \frac{\dot{v}(t)}{v(t)}\right)\mathrm{d}\widehat{\mathbb{F}}_0(t)\right\},$$

where $\theta^* \in [\theta \wedge \widehat{\theta}, \theta \vee \widehat{\theta}]$. By using the uniform convergence rate of $\widehat{G}$, we find that

$$\theta - \widehat{\theta} = \left[\ddot{\ell}_c(\theta^*, \widehat{G})\right]^{-1}\left\{\dot{\ell}_c(\theta, G) - \int_{\ell_L}^{u_L}\left[\frac{\dot{\widehat{w}}(t;\theta) - \dot{w}(t)}{w(t)} - \frac{\dot{w}(t)}{w^2(t)}\left(\widehat{w}(t;\theta) - w(t)\right)\right]\mathrm{d}F_1(t)\right.$$
$$- \int_{\ell_L}^{u_L}\left[\frac{\dot{\widehat{v}}(t;\theta) - \dot{v}(t)}{v(t)} - \frac{\dot{v}(t)}{v^2(t)}\left(\widehat{v}(t;\theta) - v(t)\right)\right]\mathrm{d}F_0(t)$$
$$- \int_{\ell_L}^{u_L}\left[\frac{\dot{\widehat{w}}(t;\theta) - \dot{w}(t)}{w(t)} - \frac{\dot{w}(t)}{w^2(t)}\left(\widehat{w}(t;\theta) - w(t)\right)\right]\mathrm{d}\widehat{\mathbb{F}}_1(t)$$
$$\left. - \int_{\ell_L}^{u_L}\left[\frac{\dot{\widehat{v}}(t;\theta) - \dot{v}(t)}{v(t)} - \frac{\dot{v}(t)}{v^2(t)}\left(\widehat{v}(t;\theta) - v(t)\right)\right]\mathrm{d}\widehat{\mathbb{F}}_0(t)\right\} + r_n \tag{A2}$$

$$= \left[\ddot{\ell}_c(\theta^*, \widehat{G})\right]^{-1}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\chi_i^{\ell_c} - \chi_i^{F_1} - \chi_i^{F_0}\right] - I_1 - I_2\right\} + r_n, \tag{A3}$$

where $r_n = \mathcal{O}_{a.s.}\left(n^{-1}\log\log n\right)$, and I_1 and I_2 denote the last two terms inside the brackets of Equation (A2), respectively. Note that $\chi_i^{\ell_c}$, $\chi_i^{F_1}$ and $\chi_i^{F_0}$ are defined in the notation section of the Appendix.

(3) The next step is to show that I_1 and I_2 are of order $o_{a.s.}\left(n^{-1/2}\right)$. First, notice that

$$\mathrm{d}[\widehat{w}(t;\theta) - w(t)] = \int_0^t\left[G(t-u) - \widehat{G}(t-u)\right]\mathrm{d}h(u;\theta)\,\mathrm{d}t + \left[G(t) - \widehat{G}(t)\right]h(0;\theta)\,\mathrm{d}t, \tag{A4}$$

where $\mathrm{d}[\hat{\dot{w}}(t;\theta) - \dot{w}(t)]$ has a similar expression by replacing h by $\dot{h}$, and

$$\mathrm{d}[\hat{v}(t;\theta) - v(t)] = \int_0^t \left[\hat{G}(t-u) - G(t-u)\right] \mathrm{d}h^{(1)}(u;\theta)\,\mathrm{d}t + \left[\hat{G}(t) - G(t)\right] h^{(1)}(0;\theta)\,\mathrm{d}t$$
$$+ h(0;\theta)\mathrm{d}\left[\hat{G}(t) - G(t)\right], \tag{A5}$$

where $\mathrm{d}[\hat{\dot{v}}(t;\theta) - \dot{v}(t)]$ has a similar form by replacing h and $h^{(1)}$ by $\dot{h}$ and $\dot{h}^{(1)}$, respectively. Now, using partial integration, we have

$$\begin{aligned} I_1 = & \left[\frac{\hat{\dot{w}}(t;\theta) - \dot{w}(t)}{w(t)} - \frac{\dot{w}(t)}{w^2(t)}\left(\hat{w}(t;\theta) - w(t)\right)\right] \hat{\mathbb{F}}_1(t)\Big|_{t=\ell_L}^{t=u_L} \\ & - \left\{\int_{\ell_L}^{u_L} \left[\hat{\dot{w}}(t;\theta) - \dot{w}(t)\right]\hat{\mathbb{F}}_1(t)\,\mathrm{d}\left(\frac{1}{w(t)}\right) - \int_{\ell_L}^{u_L} \left[\hat{w}(t;\theta) - w(t)\right]\hat{\mathbb{F}}_1(t)\,\mathrm{d}\left(\frac{\dot{w}(t)}{w^2(t)}\right)\right\} \\ & - \left\{\int_{\ell_L}^{u_L} \hat{\mathbb{F}}_1(t)\frac{1}{w(t)}\,\mathrm{d}\left[\hat{\dot{w}}(t;\theta) - \dot{w}(t)\right] - \int_{\ell_L}^{u_L} \hat{\mathbb{F}}_1(t)\frac{\dot{w}(t)}{w^2(t)}\,\mathrm{d}\left[\hat{w}(t;\theta) - w(t)\right]\right\}, \end{aligned}$$

and by employing Equation (A4) and using the uniform convergence rates of $\hat{w}(.;\theta) - w$ and $\hat{\dot{w}}(.;\theta) - \dot{w}$, which follow from the uniform convergence rate of $\hat{G}$, we find that

$$I_1 = \mathcal{O}_{a.s.}\left(n^{-1}\log\log n\right). \tag{A6}$$

Analogously, by using partial integration, Equation (A5) and the uniform convergence rates of $\hat{v}(.;\theta) - v$ and $\hat{\dot{v}}(.;\theta) - \dot{v}$, we find that

$$I_2 = r_{2,n} - \int_{\ell_L}^{u_L} \left[\frac{\dot{h}(0;\theta)}{v(t)} - \frac{\dot{v}(t)\,h(0;\theta)}{v^2(t)}\right] \hat{\mathbb{F}}_0(t)\,\mathrm{d}\left[\hat{G}(t) - G(t)\right] \tag{A7}$$

where $r_{2,n} = \mathcal{O}_{a.s.}\left(n^{-1}\log\log n\right)$. For the second term in the R.H.S. of (A7), divide $[\ell_L, u_L]$ into m sub-intervals $[y_0, y_1], [y_1, y_2], \ldots, [y_{m-1}, y_m]$ $(y_0 = \ell_L, y_m = u_L)$ of equal length $l = a_0 n^{-1/2}(\log n)^q$ ($q \geq 1/2$ and $a_0 > 0$ is some constant), so m is of order $\mathcal{O}\left(n^{-1/2}(\log n)^{-q}\right)$. We have

$$\begin{aligned} & \left|\int_{\ell_L}^{u_L} \left[\frac{\dot{h}(0;\theta)}{v(t)} - \frac{\dot{v}(t)\,h(0;\theta)}{v^2(t)}\right] \hat{\mathbb{F}}_0(t)\,\mathrm{d}\left[\hat{G}(t) - G(t)\right]\right| \\ & \leq \sum_{i=0}^{m-1} \left|\int_{y_i}^{y_{i+1}} \left[\frac{\dot{h}(0;\theta)}{v(t)} - \frac{\dot{v}(t)\,h(0;\theta)}{v^2(t)}\right] \hat{\mathbb{F}}_0(t)\,\mathrm{d}\left[\hat{G}(t) - G(t)\right]\right| \\ & \leq \sum_{i=0}^{m-1} \left\|\frac{\dot{h}(0;\theta)}{v(.)} - \frac{\dot{v}(.)\,h(0;\theta)}{v^2(.)}]\right\| \cdot \left\|\hat{\mathbb{F}}_0\right\| \int_{y_i}^{y_{i+1}} \left|\mathrm{d}\left[\hat{G}(t) - G(t)\right]\right| \\ & \leq \left\|\frac{\dot{h}(0;\theta)}{v(.)} - \frac{\dot{v}(.)\,h(0;\theta)}{v^2(.)}\right\| \cdot \left\|\hat{\mathbb{F}}_0\right\| \sum_{i=0}^{m-1} \sup_{u,v\in[y_i,y_{i+1}]} \left|\left[\hat{G}(v) - G(v)\right] - \left[\hat{G}(u) - G(u)\right]\right|. \end{aligned} \tag{A8}$$

The sup-norm term, inside the summation, on the R.H.S. of (A8) is of order $\mathcal{O}_{a.s.}\left(n^{\frac{-3}{4}}(\log n)^{\frac{1+q}{2}}\right)$, as $n \to \infty$, by the oscillation result in Meng, Bassiakos & Lo (1991) (see proposition 1, page 6). Since $\|\widehat{\mathbb{F}}_0\|$ and m are of order $\mathcal{O}_{a.s.}\left(n^{-1/2}\big(\log\log n\big)^{1/2}\right)$ and $\mathcal{O}\left(n^{-1/2}(\log n)^{-q}\right)$, respectively, the term on the R.H.S. of (A8) is of order $\mathcal{O}_{a.s.}\left(n^{-3/4}(\log n)^{\alpha_1}\right)$ ($\alpha_1 \geq 1$). Hence,

$$I_2 = \mathcal{O}_{a.s.}\left(n^{-3/4}(\log n)^{\alpha_1}\right). \tag{A9}$$

Now, since $\ddot{\ell}_c(\theta^*, \widehat{G})$ converges to Ω_{θ_0} by the consistency of $\widehat{G}$ to G and that of $\widehat{\theta}$ to θ, it follows from (A3), (A6), and (A9) that

$$\widehat{\theta} - \theta = \Omega_\theta^{-1}\frac{1}{n}\sum_{j=1}^{n}\left[-\chi_i^{\ell_c} + \chi_i^{F_1} + \chi_i^{F_0}\right] + o_p\left(n^{-1/2}\right).$$

This completes the proof. ■

Proof of Corollary 1. Recall that $\theta \to \widehat{\mu}(x;\theta)$, defined in (7), is differentiable. We have

$$\begin{aligned}\sqrt{nb_n}\left[\widehat{\mu}(x;\widehat{\theta}) - \mu(x;\theta)\right] &= \sqrt{nb_n}\left[\widehat{\mu}(x;\widehat{\theta}) - \widehat{\mu}(x;\theta)\right] + \sqrt{nb_n}\left[\widehat{\mu}(x;\theta) - \mu(x;\theta)\right] \\ &= \sqrt{nb_n}\,\dot{\widehat{\mu}}(x;\theta^*)\big(\widehat{\theta} - \theta\big) + \sqrt{nb_n}\left[\widehat{\mu}(x;\theta) - \mu(x;\theta)\right],\end{aligned}$$

by using a first order Taylor expansion, where θ_i^* is between θ_i and $\widehat{\theta}_i$, $i = 1, \ldots, d$ (θ_i and $\widehat{\theta}_i$ denote the respective i^{th} elements of θ and $\widehat{\theta}$). Notice that

$$\sqrt{nb_n}\,\dot{\widehat{\mu}}(x;\theta^*)\big(\widehat{\theta} - \theta\big) = \sqrt{b_n}\,\dot{\widehat{\mu}}(x;\theta^*)\left[\sqrt{n}(\widehat{\theta} - \theta)\right],$$

where $\sqrt{n}(\widehat{\theta} - \theta)$ converges to a Gaussian random variable, by Theorems 2, and $\dot{\widehat{\mu}}(x;\theta^*)$ converges to $\dot{\mu}(x)$, which follows from the consistency of $\widehat{G}$ to G and that of $\widehat{\theta}$ to θ. Hence, as $b_n \to 0$ and $n \to \infty$

$$\sqrt{nb_n}\,\dot{\widehat{\mu}}(x;\theta^*)\big(\widehat{\theta} - \theta\big) \xrightarrow{P} 0,$$

and the result follows from Theorem 1, under the assumptions of Theorems 1 and 2. ■

Proof of Proposition 1. (1) Following similar arguments to that of Theorem 1's proof, we can show that $\widehat{F}_{Y^*|X^*}$ in (10) admits the representation

$$\begin{aligned}\widehat{F}_{Y^*|X^*}(y|x) - F_{Y^*|X^*}(y|x) = \frac{1}{nb_n}\sum_{i=1}^{n} K_b(x - X_i)\frac{\delta_i}{w_u(Z_i)}\left[\mathrm{I}(Z_i \leq y) - F_{Y^*|X^*}(y|x)\right] \times \frac{\mu(x)}{\widehat{f}_X(x)} \\ + o_p\left(\frac{1}{\sqrt{nb_n}}\right),\end{aligned}$$

where $\widehat{f}_X(x) = (nb)^{-1}\sum_{i=1}^{n} K_b(x - X_i)$. The limit distribution of $\widehat{F}_{Y^*|X^*}$ follows then from the above representation, by employing the Lindberg-Feller and Slutsky theorems.

(2) Under uniform truncations, one can show that

$$\mathrm{d}F_{Z,\delta|X}(t,1|x) = \frac{w_u(t)\,\mathrm{d}F_{Y^*|X^*}(t|x)}{\mu(x)}$$

and

$$C_x(t) = \frac{w_u(t)\,\overline{F}_{Y^*|X^*}(t|x)}{\mu(x)}.$$

The formula of $S^2(x,y)$ then becomes

$$S^2(x,y) = \overline{F}^2_{Y^*|X^*}(y|x)\int_0^y \overline{F}^{-2}_{Y^*|X^*}(t|x)\,\frac{\mathrm{d}F_{Y^*|X^*}(t|x)}{w_u(t)} \times \frac{\mu(x)\int K^2(u)du}{f_X(x)}.$$

Therefore, one needs to prove that

$$\begin{aligned}
&\overline{F}^2_{Y^*|X^*}(y|x)\int_0^y \overline{F}^{-2}_{Y^*|X^*}(t|x)\,\frac{\mathrm{d}F_{Y^*|X^*}(t|x)}{w_u(t)} \\
&\qquad \geq E\left[\left(\mathrm{I}(Y^* \leq y) - F_{Y^*|X^*}(y|x)\right)^2 w_u^{-1}(Y^*)\middle| X^* = x\right] \\
&\qquad = \overline{F}^2_{Y^*|X^*}(y|x)\int_0^y \frac{\mathrm{d}F_{Y^*|X^*}(t|x)}{w_u(t)} + F^2_{Y^*|X^*}(y|x)\int_y^\infty \frac{\mathrm{d}F_{Y^*|X^*}(t|x)}{w_u(t)},
\end{aligned}$$

which is equivalent to showing that

$$\overline{F}^2_{Y^*|X^*}(y|x)\int_0^y \left[\overline{F}^{-2}_{Y^*|X^*}(t|x) - 1\right]\frac{\mathrm{d}F_{Y^*|X^*}(t|x)}{w_u(t)} \geq F^2_{Y^*|X^*}(y|x)\int_y^\infty \frac{\mathrm{d}F_{Y^*|X^*}(t|x)}{w_u(t)}.$$

By remarking that the weight function w_u is nondecreasing, we find that

$$\begin{aligned}
\overline{F}^2_{Y^*|X^*}(y|x)\int_0^y \left[\overline{F}^{-2}_{Y^*|X^*}(t|x) - 1\right]\frac{\mathrm{d}F_{Y^*|X^*}(t|x)}{w_u(t)} &\geq \frac{\overline{F}^2_{Y^*|X^*}(y|x)}{w_u(y)}\int_0^y \left[\overline{F}^{-2}_{Y^*|X^*}(t|x) - 1\right]\mathrm{d}F_{Y^*|X^*}(t|x) \\
&= \frac{\overline{F}_{Y^*|X^*}(y|x)\,F^2_{Y^*|X^*}(y|x)}{w_u(y)},
\end{aligned}$$

and

$$F^2_{Y^*|X^*}(y|x)\int_y^\infty \frac{\mathrm{d}F_{Y^*|X^*}(t|x)}{w_u(t)} \leq \frac{\overline{F}_{Y^*|X^*}(y|x)\,F^2_{Y^*|X^*}(y|x)}{w_u(y)}.$$

Thus, $S^2(x,y) \geq V_u(x,y)$.

■

Remark 1. *Note that failing to account for the selection bias in the sample leads to biased results. In the length-biased sample case, for instance, it is well known that* $f_{lb}(x,y) = y\,f(x,y)/E[Y^*]$ *, where* f_{lb} *and* f *are the respective bivariate densities from the sampled and* *targeted populations, and Y^* is the selection-variable (e.g., lifetime) (see Rabhi ς Bouezmarni, 2020). This leads to the relationship*

$$f_{lb}(y|x) = \frac{y\, f(y|x)}{\int_0^\infty y\, f(y|x)\mathrm{d}y},$$

with $f_{lb}(y|x)$ and $f(y|x)$ the conditional densities from the sampled and targeted populations, and therefore,

$$E[Y|X=x] = \int_0^\infty y\, f_{lb}(y|x)\mathrm{d}y = \frac{E[Y^{*\,2}|X^*=x]}{E[Y^*|X^*=x]} = E[Y^*|X^*=x] + \frac{\mathrm{Var}[Y|X^*=x]}{E[Y^*|X^*=x]}.$$

If one ignores the fact that the data are left-truncated (i.e., $(X,Y) \neq (X^*,Y^*)$)), then $\widehat{\mu}(x)$ (estimator of $\mu(x)$) will converge to $E[Y|X=x] = \mu(x) + \mathrm{Var}[Y|X^*=x]/E[Y^*|X^*=x] > \mu(x)$, which means that $\widehat{\mu}(x)$ is inconsistent. In the general biased sampling case, note that the conditional cumulative hazard of Y given $X=x$ is $\Lambda_{Y|X}(y|x) = \int_0^y [1-F_{Z|X}(t|x)]^{-1}\mathrm{d}F_{Z,\delta|X}(t,1|x)$ (see Dabrowska, 1989), p. 1157), if left-truncation is ignored. However, notice that $\Lambda_{Y|X}$ can be written as

$$\Lambda_{Y|X}(y|x) = \int_0^y \frac{P[t \le Z|X=x])}{P[T \le t \le Z|X=x]}\,\frac{\mathrm{d}F_{Z,\delta|X}(y,1|x)}{P[T \le t \le Z|X=x]} = \int_0^y \frac{P[t \le Z|X=x])}{P[T \le t \le Z|X=x]}\,\mathrm{d}\Lambda_{Y^*|X^*}(t|x),$$

where $\Lambda_{Y^*|X^*}$ is the cumulative hazard from the targeted population (see Iglesias-Pérez ς González-Manteiga, 1999, p. 217), and therefore

$$E[Y|X=x] = \int_0^\infty [1-F_{Y|X}(y|x)]\,\mathrm{d}y = \int_0^\infty \exp\left[-\int_0^y \frac{P[t \le Z|X=x])}{P[T \le t \le Z|X=x]}\,\mathrm{d}\Lambda_{Y^*|X^*}(t|x)\right]\,\mathrm{d}y.$$

This implies that $\widehat{\mu}(x)$ will converge to $E[Y|X=x] > \int_0^\infty \exp\left[-\int_0^y \mathrm{d}\Lambda_{Y^*|X^*}(y|x)\right]\,\mathrm{d}y = \mu(x)$, if left-truncation is not accounted for, and hence, $\widehat{\mu}(x)$ is inconsistent.

ACKNOWLEDGEMENTS

Parts of this article were developed and written during Y. Rabhi's postdoctoral fellowship at Université Catholique de Louvain, in Belgium. This work has been supported in part by Canssi and Statlab-CRM (Y. Rabhi) and NSERC of Canada (M. Asgharian). The authors thank Prof Grace Yi (the editor), the associate editor and two referees for their comments.

BIBLIOGRAPHY

Andersen, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society, Series B*, 32, 283–301.

Asgharian, M. (2014). On the singularities of the information matrix and multi-path change-point problems. *Theory of Probability and Its Applications*, 58, 546–561.

Asgharian, M., M'lan, C. E., & Wolfson, D. B. (2002). Length-biased sampling with right-censoring: An unconditional approach. *Journal of the American Statistical Association*, 97, 202–209.

Bergeron, P. -J., Asgharian, M., & Wolfson, D. B. (2008). Covariate bias induced by length-biased sampling of failure times. *Journal of the American Statistical Association*, 103, 737–742.

Cristóbal, J. A., Ojeda, J. L., & Alcalá, J. T. (2004). Confidence bands in nonparametric regression with length-biased data. *Annals of the Institute of Statistical Mathematics*, 56, 475–496.

Dabrowska, D. M. (1989). Uniform consistency of the kernel conditional Kaplan–Meier estimate. *Annals of Statistics*, 17, 1157–1167.

de Uña-Àlvarez, J. (2003). Large sample results under biased sampling when covariables are present. *Statistics and Probability Letters*, 63, 287–293.

de Uña-Àlvarez, J. & Iglesias-Pérez, M. C. (2010). Nonparametric estimation of a conditional distribution from length-biased data. *Annals of the Institute of Statistical Mathematics*, 62, 323–341.

Fan, J. & Gijbels, I. (1996). *Local polynomial modelling and its applications, Monographs on statistics and applied probability*, Vol. 66. Chapman & Hall, CRC press.

Geskus, R. B. (2000). On the inclusion of prevalent cases in HIV/AIDS natural history studies through a marker-based estimate of time since seroconversion. *Statistics in Medicine*, 19, 1753–1769.

Gürler, Ü.. (1996). Bivariate estimation with right-truncated data. *Journal of the American Statistical Association*, 91, 1152–1165.

Härdle, W. & Marron, J. S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *Annals of Statistics*, 13, 1465–1481.

Iglesias-Pérez, M. C. & González-Manteiga, W. (1999). Strong representation of a generalized product-limit estimator for truncated and censored data with some applications. *Journal of Nonparametric Statistics*, 10, 213–244.

Liang, H. -Y., de Uña-Àlvarez, J., & Iglesias-Pérez, M. C. (2015). A central limit theorem in non-parametric regression with truncated, censored and dependent data. *Scandinavian Journal of Statistics*, 42, 256–269.

Lo, S. H., Mack, Y. P., & Wang, J. L. (1989). Density and hazard rate estimation for censored data via strong representation of the K-M estimator. *Probability Theory and Related Fields*, 80, 461–473.

Meng, X. -L., Bassiakos, Y., & Lo, S. -H. (1991). Large-sample properties for a general estimator of the treatment effect in the two-sample problem with right censoring. *Annals of Statistics*, 19, 1786–1812.

Rabhi, Y. & Asgharian, M. (2017). Inference under biased sampling and right censoring for a change-point in the hazard function. *Bernoulli*, 23, 2720–2745.

Rabhi, Y. & Bouezmarni, T. (2020). Nonparametric inference for copulas and measures of dependence under length-biased sampling and informative censoring. *Journal of the American Statistical Association*, 115, 1268–1278.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

Wang, M. C. (1989). A semiparametric model for randomly truncated data. *Journal of the American Statistical Association*, 84, 742–748.

Wang, M. C. (1991). Nonparametric estimation from cross-sectional survival data. *Journal of the American Statistical Association*, 86, 130–143.

Wang, M. C. (1992). The analysis of retrospectively ascertained data in the presence of reporting delays. *Journal of the American Statistical Association*, 87, 397–406.

Wu, C. O. (2000). Local polynomial regression with selection biased data. *Statistica Sinica*, 10, 789–817.

Received 28 October 2018
Accepted 21 May 2020