Stat

The ISI's Journal for the Rapid Dissemination of Statistics Research

(wileyonlinelibrary.com) DOI: 10.1002/sta4.46

Propensity score estimation in the presence of length-biased sampling: a non-parametric adjustment approach

Ashkan Ertefaie[a]*, Masoud Asgharian[b] and David Stephens[b]

Received 20 January 2014; Accepted 13 February 2014

The pervasive use of prevalent cohort studies on disease duration increasingly calls for an appropriate methodology to account for the biases that invariably accompany samples formed by such data. It is well known, for example, that subjects with shorter lifetime are less likely to be present in such studies. Moreover, certain covariate values could be preferentially selected into the sample, being linked to the long-term survivors. The existing methodology for estimating the propensity score using data collected on prevalent cases requires the correct conditional survival/hazard function given the treatment and covariates. This requirement can be alleviated if the disease under study has stationary incidence, the so-called stationarity assumption. We propose a non-parametric adjustment technique based on a weighted estimating equation for estimating the propensity score, which does not require modeling the conditional survival/hazard function when the stationarity assumption holds. The estimator's large-sample properties are established, and its small-sample behavior is studied via simulation. The estimated propensity score is utilized to estimate the survival curves. Copyright © 2014 John Wiley & Sons, Ltd.

Keywords: causal inference; length-biased sampling; propensity score; survival curve

1 Introduction

Survival or failure time data typically comprise an initiating event, say onset of a disease, and a terminating event, say death. In an ideal situation, recruited subjects have not experienced the initiating event, the so-called incident cases. These cases are then followed to a terminating event or censoring, say loss to follow-up. In many practical situations, however, recruiting incident cases is infeasible because of logistic or other constraints. In such circumstances, subjects who have experienced the initiating event prior to the start of the study, so-called prevalent cases, are recruited. It is well known that these cases tend to have a longer survival time and hence form a biased sample from the target population. This bias is termed *length bias* when the initiating events are generated by a stationary Poisson process (Cox & Lewis, 1966; Zelen & Feinlein, 1969), the so-called stationarity assumption.

Studies on length-biased sampling can be traced as far back as Wicksell (1925) and his corpuscle problem. The phenomenon was later noticed by Fisher (1934) in his article on methods of ascertainment. Neyman (1955) discussed length-biased sampling further and coined the term *incidence-prevalence* bias. Cox (1969) studied length-biased sampling in industrial applications, while Zelen & Feinlein (1969) observed the same bias in screening tests for

[a]**Department of Statistics, University of Michigan, Ann Arbor, MI, 48108, USA**

[b]**Department of Mathematics and Statistics, McGill University, Montreal, Quebec, H3A 2K6, Canada**

***Email: ertefaie@umich.edu**

Stat

(wileyonlinelibrary.com) DOI: 10.1002/sta4.46

The ISI's Journal for the Rapid Dissemination of Statistics Research

disease prevalence (Asgharian et al., 2002; Asgharian & Wolfson, 2005). More recently, Shen et al. (2009), Qin & Shen (2010), and Ning et al. (2010) have studied the analysis of covariates under biased sampling.

In observational studies, treatment is assigned to the experimental units without randomization. Thus, in each treatment group, the covariate distributions may be imbalanced, which may lead to bias in estimating the treatment effect if the covariate imbalance is not properly taken into account (Cochran & Rubin, 1973; Rubin, 1973). The propensity score is a tool that is widely used in causal inference to adjust for this source of bias (Robins et al., 2000; Hernán et al., 2000). Rosenbaum & Rubin (1983) defined the propensity score for a binary treatment D as $p(D = 1|\mathbf{X})$, where $\mathbf{X}$ is a vector of measured covariates. They show that under some assumptions, treatment is independent of the covariates inside each propensity score stratum (the *balancing property* of the propensity score).

In cases where the sample is not representative of the population, naive propensity score estimation will not in general have the balancing property. Cheng & Wang (2012) developed a method that consistently estimates the parameters of the propensity score from prevalent survival data. They also presented a method that can be used in a special case of length-biased sampling. Their method requires correct specification of the conditional hazard model given the treatment and covariates. We refer to their estimator as CW in the sequel.

Our goal is to develop a method that estimates the propensity score using a weighted logistic regression where weights are estimated non-parametrically. Our estimating equation is designed specifically for length-biased data, i.e., for a disease with stationary incidence. Unlike the method proposed by CW, our method does not require any model specification for the conditional failure time given the exposure and the covariates. We also generalize a non-parametric survival curve estimation method introduced by Huang & Qin (2011) to accommodate confounding as well as length-biased sampling.

2 Length-biased sampling

In this section, we introduce concepts and notations necessary to formulate problems involving length-biased sampling. We adopt the common modeling framework for prevalent cohort studies. We assume that affected individuals in the study population develop the condition of interest (*onset*) according to some stochastic mechanism at random points in calendar time and undergo a terminal event (*failure*) at some subsequent time point that is also determined by a stochastic mechanism. Individuals enter into the study at some census time and are followed up until the terminal or censoring event.

2.1. Notations

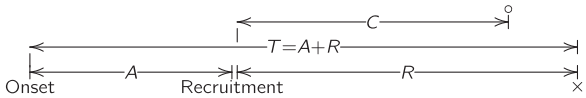
Let T^{pop} be the time measured from the onset to failure time in the target population with an absolutely continuous distribution F and density f. Also, let D^{pop} and $\mathbf{X}^{\text{pop}}$ be the binary treatment variable and the vector of covariates, respectively. Let T be the same measured time for *observed* subjects with distribution F_{LB}. The variables with superscript pop represent the population variables; variables without pop denote the observed truncated variables. It is well known that if the onset times are generated by a stationary Poisson process, then

$$F_{\text{LB}}(t) = \frac{\int_0^t s\,\mathrm{d}F(s)}{\int_0^\infty s\,\mathrm{d}F(s)} = \frac{1}{\mu}\int_0^t s\,\mathrm{d}F(s) \qquad \text{and} \qquad f_{\text{LB}}(t) = \frac{tf(t)}{\mu}, \tag{1}$$

where f_{LB} is the density function of F_{LB} and μ is the mean survival time under F. The observed event time, T, can be written as $A + R$, where A is the time from the onset of the disease to the recruitment time and R, the residual

(wileyonlinelibrary.com) DOI: 10.1002/sta4.46

lifetime, is the time from recruitment to the event, also called backward and forward recurrence times, respectively. When individuals are also subject to right censoring, the observed survival time is $Y = A + \min(R, C)$, where C is a censoring time measured from the recruitment to the loss to follow-up; for all subjects, both A and $\min(R, C)$ are observed. The censoring indicator is denoted by $\delta(\delta = 1$ indicating failure). The sample consists of $(y_i, a_i, \delta_i, d_i, \mathbf{x}_i)$ for n independent subjects. The following diagram illustrates the different random quantities introduced in this section.



Throughout the paper, we assume that the following assumptions hold:

A1. The variable $(T^{\text{pop}}, D^{\text{pop}}, \mathbf{X}^{\text{pop}})$ is independent of the calendar time of the onset of the disease.
A2. The disease has stationary incidence, i.e., the disease incidence occurs at a constant rate.
A3. The censoring time C is independent of $(A, R, D, \mathbf{X})$.

2.2. Potential outcomes

We use the counterfactual or potential outcome framework to define the causal effect of interest. Potential outcome models are introduced by Neyman (1990) and Rubin (1978) for time independent treatment. We define the counterfactual values $(A(d), R(d), Y(d))$ corresponding to the backward, forward recurrence times, and observed survival time, respectively, had—possibly contrary to fact—the treatment taken the value d. Also, let $T^{\text{pop}}(d)$ denote the counterfactual response if contrary to the fact that all the individuals would have received the treatment $D = d$, and let D denote the treatment received. The observed response, T^{pop}, is defined as $DT^{\text{pop}}(1) + (1 - D)T^{\text{pop}}(0)$.

We make the following *identifiability* assumptions to link the counterfactual outcome and the observed data (Robins, 1994, 1997):

A4. *Consistency*: Potential outcome for a treatment corresponds to the actual outcome if assigned to that treatment.
A5. *No unmeasured confounding*: Given the observed covariates $\mathbf{X}$, the counterfactual outcome $Y(d)$ is independent of the assignment of treatment.
A6. *Positivity*: Let $p_{D|\mathbf{X}}(d|\mathbf{x})$ be the conditional probability of receiving treatment d given $\mathbf{X} = \mathbf{x}$. For each treatment d and for each possible value $\mathbf{x}$, $p_{D|\mathbf{X}}(d|\mathbf{x}) > 0$.

3 Propensity score estimation under length-biased sampling

Assuming a logit model for the propensity score in the target population, we have

$$\pi(\mathbf{x}, \alpha) = p(D^{\text{pop}} = 1|\mathbf{X}^{\text{pop}} = \mathbf{x}) = \frac{\exp(\alpha\mathbf{x})}{1 + \exp(\alpha\mathbf{x})}, \tag{2}$$

where α is a $p \times 1$ vector of parameters. The vector of covariates X may include a column of 1s. It can be shown that under assumption $A2$, we have

$$p_{\text{LB}}(D = 1|\mathbf{X} = \mathbf{x}) = \frac{\mu_1(\mathbf{x}, \theta)p(D^{\text{pop}} = 1|\mathbf{X}^{\text{pop}} = \mathbf{x})}{\pi(\mathbf{x}, \alpha)\mu_1(\mathbf{x}, \theta) + (1 - \pi(\mathbf{x}, \alpha))\mu_0(\mathbf{x}, \theta)}, \tag{3}$$

where $\mu_d(\mathbf{x},\theta) = \int_0^\infty p(T^{\text{pop}}(d) \geq a|\mathbf{X}=\mathbf{x},\theta)\,\mathrm{d}a$ for $d = 0, 1$ is the conditional counterfactual mean failure time if treated at $D = d$ (Bergeron et al. (2008)). Note that under assumptions *A*4–*A*6, $p(T^{\text{pop}}(d) \geq a|,\mathbf{X}=\mathbf{x},\theta) = p(T^{\text{pop}} \geq a|D=d,\mathbf{X}=\mathbf{x},\theta)$, where θ parametrizes the conditional density of T^{pop}.

Assuming the proportional hazard model, i.e., $\lambda_{T^{\text{pop}}}(u|D^{\text{pop}}=d,\mathbf{X}^{\text{pop}}=\mathbf{x}) = \lambda_0(u)e^{\gamma d+\beta\mathbf{x}}$, Cheng & Wang (2012) showed that the parameter of the propensity score can be consistently estimated using the logistic regression but adjusted for the "offset" term $\log\left(\hat{\alpha}\left(x;\hat{\Lambda},\hat{\gamma},\hat{\beta}\right)\right)$ as the intercept, where $\hat{\alpha}\left(x;\hat{\Lambda},\hat{\gamma},\hat{\beta}\right) = \frac{\sum_{i=1}^n \exp\left[-\hat{\Lambda}(a_i)\exp\left(\hat{\gamma}+\hat{\beta}\mathbf{x}\right)\right.}{\sum_{i=1}^n \exp\left[-\hat{\Lambda}(a_i)\exp\left(\hat{\beta}\mathbf{x}\right)\right]}$. The cumulative hazard function Λ is estimated using the Breslow estimator. The consistency of the parameters of the propensity score in the CW method relies on the correct specification of the conditional hazard model given the treatment and covariates.

When the initiating event of the duration variable has stationary incidence, it is possible to devise a robust method for estimating the propensity score that does not require knowledge of the conditional hazard model. See among others Wolfson et al. (2001) and De Uña-álvarez (2004) for examples of such duration variables in medical and labor force studies, respectively.

Let $f(t|d,\mathbf{x},\theta)$ be the unbiased conditional density of survival times given the covariates and treatment. Then, under assumptions *A*1 and *A*2, the joint density of (A,T) given $(D,\mathbf{X})$ is $f(t|d,\mathbf{x},\theta)/\int_0^\infty uf(u|d,\mathbf{x},\theta)\,\mathrm{d}u I(t>a>0)$ as shown by Asgharian et al. (2006). Assumption *A*3 is used to show that

$$p(Y\in(t,t+\mathrm{d}t),A\in(a,a+\mathrm{d}a),\delta=1|d,\mathbf{x},\theta) = \frac{f(t|d,\mathbf{x},\theta)S_c(t-a)\,\mathrm{d}t\mathrm{d}a}{\mu_d(\mathbf{x},\theta)},$$

where $S_c(.)$ is the survival function for the residual censoring variable C, respectively. By integrating the preceding equation over $0 < a < t$, we have

$$p(Y\in(t,t+\mathrm{d}t),\delta=1|d,\mathbf{x},\theta) = \frac{f(t|d,\mathbf{x},\theta)w(t)\,\mathrm{d}t}{\mu_d(\mathbf{x},\theta)}, \tag{4}$$

where $w(t) = \int_0^t S_C(s)\,\mathrm{d}s$ (Shen et al., 2009; Qin & Shen, 2010).

We construct an unbiased estimating equation for estimating the parameters of the propensity score using the weighted logistic regression where weights are estimated non-parametrically. Let $F(d|\mathbf{x})$ be the unbiased conditional distribution of the treatment given the covariates. Then

$$\begin{aligned}\mathbb{E}\left[\delta\frac{(D-\pi(\mathbf{x},\alpha))}{w(Y)}\,\middle|\,\mathbf{X}=\mathbf{x}\right] &= \mathbb{E}\left[\mathbb{E}\left\{\delta\frac{(d-\pi(\mathbf{x},\alpha))}{w(Y)}\,\middle|\,D=d,\mathbf{X}=\mathbf{x}\right\}\right]\\ &= \int(d-\pi(\mathbf{x},\alpha))\int\frac{f(y|\mathbf{x},d,\theta)w(y)}{w(y)\mu_d(\mathbf{x},\theta)}\,\mathrm{d}y\times\frac{\mu_d(\mathbf{x},\theta)\,\mathrm{d}F(d|\mathbf{x})}{\mu(\mathbf{x},\alpha,\theta)}\\ &= \frac{1}{\mu(\mathbf{x},\alpha,\theta)}\int(d-\pi(\mathbf{x},\alpha))\,\mathrm{d}F(d|\mathbf{x}) = 0.\end{aligned}$$

The second equality follows from Equations (4) and (3). The last equality holds because $f(y|\mathbf{x},d)$ is a proper density and (2). An unbiased estimating equation for α is therefore

$$U(\alpha) = \sum_{i=1}^n U_i(\alpha) = \sum_{i=1}^n \delta_i\mathbf{x}_i^\top\frac{(d_i-\pi(\mathbf{x}_i,\alpha))}{\widehat{w}(y_i)} = 0, \tag{5}$$

where $\hat{w}(y) = \int_0^y \hat{S}_C(s)\,\mathrm{d}s$ and $\hat{S}_C$ is the Kaplan–Meier estimator of the survivor function of the residual censoring variable C.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.46

The following theorem presents the asymptotic properties of the estimators obtained by (5) in the presence of length-biased sampling when $w(.)$ is replaced by its estimated value.

Theorem 1

Let $\hat{\alpha}$ be an estimator obtained by (5). Then under conditions C1–C6 and assumptions A1–A6, $\hat{\alpha} \to \alpha_0$ in probability as $n \to \infty$. Moreover,

$$\sqrt{n}\,(\hat{\alpha} - \alpha) \xrightarrow{d} \mathcal{N}(0, \eta(\alpha)),$$

where $\eta(\alpha)$ is given in the Appendix.

Proof

See the Appendix. □

A consistent plug-in estimator of $\eta(\alpha)$ is presented in the Appendix.

Note that the censored individuals contribute to this estimating equation through $\hat{S}_C$ as well as the uncensored ones. Let $M_{iC}(t) = \mathbb{1}(Y_i - A_i < t, \delta_i = 0) - \int_0^t \mathbb{1}(\min(Y_i - A_i, C_i) > u)\, \mathrm{d}\Lambda_C(u)$, with $\Lambda_C(.)$ be the cumulative hazard function of the censoring variable. As a part of the proof of Theorem 1, we show that as $n \to \infty$,

$$\frac{1}{n}\sum_{i=1}^{n} U_i(\alpha) = \frac{1}{n}\sum_{i=1}^{n}\left[\delta_i \mathbf{x}_i^{\top}\frac{d_i - \pi(\mathbf{x}_i, \alpha)}{w(y_i)} + \int_0^s \frac{v(t)\,\mathrm{d}M_{iC}(t)}{S_C(t)S_R(t)}\right] + o_p(n^{-1/2}), \quad s = \sup[\,t : p(C > t) > 0\,], \tag{6}$$

where $S_R(u)$ is the survival function of the residual lifetime, and the function $v(t)$ is defined in the Appendix. The second part of the summation in the right-hand side of (6) is often referred to as the augmentation element (Rotnitzky & Robins, 2005).

4 Survival curve estimation

Various methods have been proposed to adjust for length-biased sampling, including the following: the truncation product-limit estimator (Wang et al., 1986) and the maximum pseudo-partial likelihood estimator (Luo & Tsai, 2009). Here, we generalize the method introduced by Huang & Qin (2011), which incorporates the information in the marginal distribution of the truncation time from disease onset to recruitment time. The bias induced by confounding can be adjusted by creating a pseudo-population using the inverse probability of being in the group that the individuals actually belong to (Nieto & Coresh, 1996; Xie & Liu, 2005; Cole & Hernán, 2004).

Our goal is to estimate the counterfactual survival function $S_d(y)$, where $S_d(y) = \mathbb{E}[\,I(T^{\mathrm{pop}}(d) > t)]$. Under assumptions *A4–A6*, the function $S_d(y)$ can be identified using the observed data as follows:

$$S_d(t) = \mathbb{E}[\,I(T^{\mathrm{pop}}(d) > t)] = \mathbb{E}\left[\frac{I(D = d)}{p(D = d|\mathbf{X})} I(T^{\mathrm{pop}} > t)\right].$$

Following Huang & Qin (2011), $S_d(t)$ can be estimated by

$$\tilde{\mathrm{S}}_d(t) = \prod_{u \in [0,t]}\left[1 - d\hat{\Lambda}_d(u)\right],$$

where

$$d\hat{\Lambda}_{id} = \int_0^t \frac{\mathrm{d}\tilde{N}_d(t)}{\tilde{R}_d(t)}$$

Stat

(wileyonlinelibrary.com) DOI: 10.1002/sta4.46

The ISI's Journal for the Rapid Dissemination of Statistics Research

with

$$\tilde{N}_d(t) = 1/n \sum_{i=1}^{n} \frac{I(d_i = d)}{p(d_i = d|\mathbf{x}_i)} \delta_i I(y_i \leq t)$$

$$\tilde{R}_d(t) = 1/n \sum_{i=1}^{n} \frac{I(d_i = d)}{p(d_i = d|\mathbf{x}_i)} I(y_i \geq t) + \tilde{S}_{dA}(t).$$

Also, the product-limit estimator $\tilde{S}_{dA}(t)$ is

$$\tilde{S}_{dA}(t) = \prod_{u \in [0,t]} \left[1 - \frac{d\tilde{Q}_d(u)}{\tilde{K}_d(u)} \right],$$

where

$$\tilde{Q}_d(u) = 1/n \sum_{i=1}^{n} \frac{I(d_i = d)}{p(d_i = d|\mathbf{x}_i)} [\, I(a_i \leq t) + \delta_i I(y_i - a_i \leq t)]$$

$$\tilde{K}_d(t) = 1/n \sum_{i=1}^{n} \frac{I(d_i = d)}{p(d_i = d|\mathbf{x}_i)} [\, I(a_i \geq t) + I(y_i - a_i \geq t)]\,.$$

We utilize our proposed estimating equation (5) to estimate the propensity score and replace $p(d_i = d|\mathbf{x}_i)$ with $p\,(d_i = d|\mathbf{x}_i, \hat{\alpha})$.

5 Simulation studies

In this section, we describe a simulation study to examine the performance of the proposed propensity score estimator. Our simulation consists of 500 datasets of sizes 500 and 5000. The censoring variable C is generated from a uniform

Table I. Simulation: propensity score parameter estimation.

Method	Bias	SD	Bias	SD
10% Cens.	$n = 500$		$n = 5000$	
$\hat{\alpha}$	(0.01, 0.04, 0.03)	(0.21, 0.43, 0.45)	(0.00, 0.01, 0.01)	(0.09, 0.18, 0.19)
$\hat{\alpha}_w$	(0.08, 0.02, 0.01)	(0.17, 0.28, 0.28)	(0.01, 0.00, 0.01)	(0.06, 0.10, 0.10)
$\hat{\alpha}_w^m$	(0.03, 0.02, 0.49)	(0.17, 0.29, 0.23)	(0.08, 0.03, 0.48)	(0.06, 0.09, 0.08)
$\hat{\alpha}_{\text{Un}}$	(0.10, 0.50, 0.50)	(0.11, 0.21, 0.22)	(0.10, 0.51, 0.50)	(0.04, 0.07, 0.08)
20% Cens.	$n = 500$		$n = 5000$	
$\hat{\alpha}$	(0.01, 0.05, 0.05)	(0.22, 0.42, 0.43)	(0.02, 0.03, 0.03)	(0.09, 0.18, 0.20)
$\hat{\alpha}_w$	(0.02, 0.01, 0.01)	(0.17, 0.29, 0.27)	(0.02, 0.01, 0.01)	(0.06, 0.10, 0.10)
$\hat{\alpha}_w^m$	(0.01, 0.02, 0.47)	(0.16, 0.29, 0.21)	(0.02, 0.05, 0.46)	(0.06, 0.10, 0.08)
$\hat{\alpha}_{\text{Un}}$	(0.11, 0.49, 0.50)	(0.10, 0.22, 0.20)	(0.10, 0.51, 0.50)	(0.04, 0.07, 0.08)
30% Cens.	$n = 500$		$n = 5000$	
$\hat{\alpha}$	(0.04, 0.08, 0.09)	(0.22, 0.44, 0.44)	(0.03, 0.06, 0.07)	(0.10, 0.19, 0.20)
$\hat{\alpha}_w$	(0.07, 0.00, 0.02)	(0.17, 0.29, 0.29)	(0.08, 0.03, 0.02)	(0.06, 0.10, 0.11)
$\hat{\alpha}_w^m$	(0.07, 0.02, 0.45)	(0.17, 0.29, 0.21)	(0.03, 0.06, 0.45)	(0.06, 0.10, 0.08)
$\hat{\alpha}_{\text{Un}}$	(0.11, 0.49, 0.50)	(0.10, 0.22, 0.20)	(0.10, 0.51, 0.50)	(0.04, 0.07, 0.08)

$\hat{\alpha}$, estimated parameters using the proposed method; $\hat{\alpha}_w$, estimated parameters using the CW method; $\hat{\alpha}_w^m$, estimated parameters using the CW method when the hazard model is misspecified; $\hat{\alpha}_{\text{Un}}$, estimated parameters when unadjusted for the length-biased sampling. $\alpha = (-0.1, 1, -1)$.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.46

distribution in the interval $(0, \tau)$ where the parameter τ is set such that it results in a desired censoring proportion. To create length-biased samples, we generate a variable A from a uniform distribution $(0, \rho)$ and ignore those whose generated unbiased failure time is less than A.

We generated the unbiased failure times from the following hazard model $h(t|d, \mathbf{x}) = 0.2 \exp\{d - 0.5x_1 + 0.5x_2 + 0.5dx_1 - 0.5dx_2\}$, where $D \sim$ Bernoulli $\left(\frac{\exp\{-0.1+1x_1-1x_2\}}{1+\exp\{-0.1+1x_1-1x_2\}}\right)$ with X_1 and X_2 distributed according to $N(0, \sigma = 0.5)$. We estimate the parameters of the propensity score using CW and the proposed method and compare the results with the true values. We assume three different censoring proportions 10%, 20%, and 30%.

We estimate the parameters of the propensity score using four different estimators: $\hat{\alpha}$ is the estimator obtained by the proposed method; $\hat{\alpha}_w$ and $\hat{\alpha}_w^m$ are the estimator obtained by the CW method when the hazard model is correctly and incorrectly specified, respectively; and $\hat{\alpha}_{\text{Un}}$ is obtained by a naive method that does not adjust for the length-biased sampling. In $\hat{\alpha}_w^m$, we assume that the interaction between the treatment D and the covariate X_2 has been ignored in the fitted hazard model.

Table I summarizes the estimated propensity score parameters and their standard errors. Our simulation results confirm that the proposed estimating equation (5) adjusts the length-biased sampling. The standard errors, however, are larger

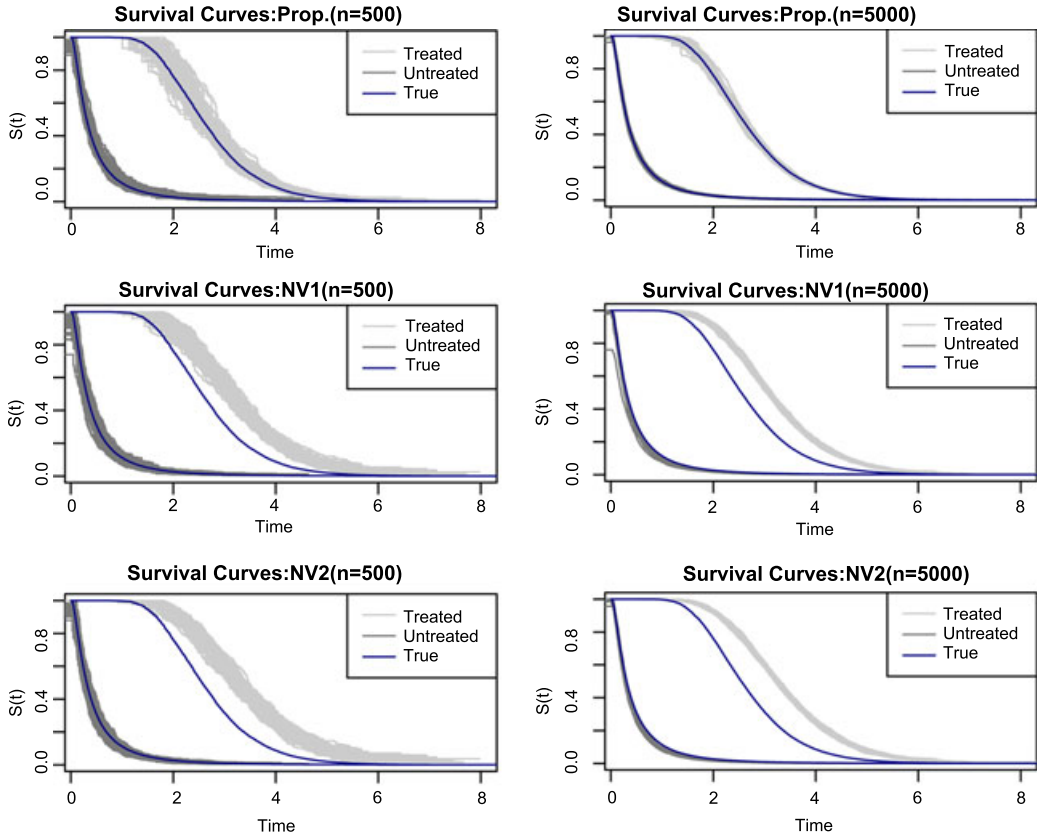


Figure 1. Simulation: the estimated survival curves using the proposed (Prop.) and two naive estimators. The light and dark shaded areas are the treated and untreated survival curves, respectively. The solid line represents the true curve. NV1 is when the propensity score is naively estimated without considering the length-biased sampling. NV2 is when we ignore both the confounding and the length-biased sampling.

than the one obtained by the CW method, which is the price we pay for relaxing the modeling assumption of the hazard model. As we expected, CW estimator is highly sensitive to model misspecification even when just one of the interaction terms is ignored. Specifically, when the interaction term between the treatment and variable X_2 is omitted in the fitted hazard model, the estimated coefficient corresponding to X_2 in the propensity score model is biased. In general, if variables in the study are correlated, then missing one variable in the hazard model may cause bias in the estimation of other variables in the propensity score model as well.

5.1. Survival curve estimation

Here, we compare our proposed method of estimating the survival curves if treated and untreated with two naive approaches. The naive method (NV1) estimates the propensity score without considering the length-biased sampling, and the second naive method (NV2) ignores both the confounding and the length-biased sampling. We generated the unbiased failure times from a Weibull distribution with shape parameter 10 and scale parameter $h_2(t|d,\mathbf{x}) = 0.2\exp\{3d+2.5x_1-2x_2-2dx_1+1.1dx_2\}$, where $D \sim$ Bernoulli $\left(\frac{\exp\{2x_1-2x_2-2x_1x_2\}}{1+\exp\{2x_1-2x_2-2x_1x_2\}}\right)$ with X_1 and X_2 are $N(1,\sigma=0.3)$. We report the result based on the censoring proportion 30%.

The survival curves for the treated and untreated groups are presented in Figure 1. The light and dark gray shaded areas are the survival curves based on the 500 datasets for treated and untreated individuals, respectively. The dark solid line is the true survival curve. As the plots show, the true survival curve lies entirely in the shaded area provided by our proposed estimator while ignoring either or both of the length-biased sampling and confounding result in a biased estimator.

6 Discussion

We present a weighted estimating equation to estimate the parameters of the propensity score from right-censored length-biased samples. In many cases, recruiting prevalent cases is more efficient. However, it is well known that in these cases subjects with longer survival time have a greater chance to be selected. This may affect the distribution of the observed covariates (Bergeron et al., 2008). For example, if treated subjects tend to live longer, then these subjects will be over-represented in the observed sample. As such, if the propensity score is fitted without adjusting for this source of bias, it will be skewed to the left. Recently, Cheng & Wang (2012) proposed a method to adjust for the length-biased sampling, which requires the correctly specified conditional survival function given the treatment and covariates. This modeling assumption may limit the application of this approach, particularly, when an investigator is interested in estimating the marginal causal effect.

In our proposed method, we estimate the weights non-parametrically. Thus, unlike the existing methods, it does not require any modeling assumptions for the conditional hazard function given the treatment and covariates. Our method produces an unbiased estimator, but the standard errors are larger than the method proposed by Cheng & Wang (2012).

Generalizing a non-parametric survival curve estimation method introduced by Huang & Qin (2011), we derive a method for estimating the counterfactual survival curves in the presence of length-biased sampling. The bias induced by confounding is adjusted by creating a pseudo-population using the inverse probability of being in the group to which the individuals actually belong. The treatment assignment probabilities are estimated using the proposed estimating equation.

We confined our attention to the stationary case; the methodology presented in this manuscript can, however, be extended to any other left-truncation cases as long as the left-truncation distribution is known (Luo & Tsai, 2009).

(wileyonlinelibrary.com) DOI: 10.1002/sta4.46

Appendix

In this section, we present the assumptions and proofs of the main result. The following conditions are required for establishing Theorem 1:

C.1 $\mathbf{X}$ is a p vector of bounded covariates, not contained in a $(p-1)$ dimensional hyperplane.
C.2 $\sup[t : p(R > t) > 0] \geq \sup[t : p(C > t) > 0] = s$ and $p(\delta = 1) > 0$.
C.3 $\int_0^s \left[\left(\int_t^s S_C(v)\,\mathrm{d}v\right)^2 / \left(S_C^2(t)S_R(t)\right)\right] \mathrm{d}S_C(t) < \infty$.
C.4 $\det \mathbb{E}\left[\left\{\delta \mathbf{X}^\top \frac{D-\pi(\mathbf{X},\alpha)}{w(Y)}\right\}^{\otimes 2}\right] < \infty$.
C.5 $\Lambda = \mathbb{E}\left[\delta \mathbf{X}^\top \frac{\partial \pi(\mathbf{X},\alpha)}{\partial \alpha}\right]$ is non-singular.
C.6 $\det\left[\int_0^s v^2(t)/\left(S_C^2(t)S_R(t)\right)\,\mathrm{d}S_C(t)\right] < \infty$, where $v(t) = \mathbb{E}\left[\frac{\delta \mathbb{1}(Y>t)\mathbf{X}^\top[D-\pi(\mathbf{X},\alpha))]\int_t^Y S_C(v)\,\mathrm{d}v}{w^2(Y)}\right]$.

C.2 is an identifiability condition (Wang, 1991), and C.3–C.6 are required to obtain an estimator with a finite variance.

Proof of Theorem 1

The stochastic process $M_C(s)$ has mean zero,

$$\begin{aligned}\mathbb{E}[M_C(s)] &= \mathbb{E}[\mathbb{1}(C < Y - A < s)] - \int_0^s \mathbb{E}[\mathbb{1}(Y - A > u).\mathbb{1}(C > u)]\,\mathrm{d}\Lambda_C(u)\\ &= \int_0^s S_C(u)\lambda_C(u)S_R(u)\,\mathrm{d}u - \int_0^s S_C(u)S_R(u)\,\mathrm{d}\Lambda_C(u) = 0.\end{aligned}$$

Using the strong consistency of $\hat{w}(y)$ to $w(y)$ (Pepe & Fleming, 1991), we have

$$\frac{1}{\hat{w}(Y)} = \frac{1}{w(Y)}\left[1 + \frac{w(Y) - \hat{w}(Y)}{w(Y)}\right] + o_p(1).$$

Thus

$$\begin{aligned}\tilde{U}(\alpha) &= \frac{1}{\sqrt{n}}\sum_{i=1}^n U_i(\alpha)\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \delta_i \mathbf{x}_i^\top \frac{d_i - \pi(\mathbf{x}_i,\alpha)}{\hat{w}(y_i)} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \delta_i \mathbf{x}_i^\top \frac{d_i - \pi(\mathbf{x}_i,\alpha)}{w(y_i)}\left[1 + \frac{w(y_i) - \hat{w}(y_i)}{w(y_i)}\right] + o_p(1), \qquad (A.1)\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \left[\delta_i \mathbf{X}_i \frac{D_i - \pi(\mathbf{X}_i,\alpha)}{w(Y_i)} + \int_0^s \frac{\hat{v}(t)\,\mathrm{d}M_C(t)}{S_C(t)S_R(t)}\right],\end{aligned}$$

where

$$\hat{v}(t) = \frac{1}{n}\sum_{i=1}^n \left[\frac{\delta_i I(Y_i > t)\mathbf{X}_i[D_i - \pi(\mathbf{X}_i,\alpha))]\int_t^{Y_i} S_C(v)\,\mathrm{d}v}{w^2(Y_i)}\right].$$

The last equality (A.1) follows from the martingale integral representation $\sqrt{n}\left(\hat{w}(Y) - w(Y)\right)$ (Shen et al., 2009; Qin & Shen, 2010).

Using the standard Taylor expansion, we can derive the asymptotic variance of the estimator as follows:

$$\eta(\alpha) = \Lambda' \Sigma^{-1} \Lambda,$$

where

$$\Sigma = \mathbb{E}\left[\tilde{U}(\alpha)\tilde{U}(\alpha)\right] = \mathbb{E}\left[\left\{\delta \mathbf{X}^\top \frac{D-\pi(\mathbf{X},\alpha)}{w(Y)}\right\}^{\otimes 2}\left\{1+\frac{w(Y)-\hat{w}(Y)}{w(Y)}\right\}^{2}\right]$$
$$= \mathbb{E}\left[\left\{\delta \mathbf{X}^\top \frac{D-\pi(\mathbf{X},\alpha)}{w(Y)} + \int_0^s \frac{v(t)\,\mathrm{d}M_C(t)}{S_C(t)S_R(t)}\right\}^{\otimes 2}\right]$$
$$\Lambda = \mathbb{E}\left[\frac{\partial U_i(\alpha)}{\partial \alpha}\right] = \mathbb{E}\left[\frac{\delta}{w(Y)}\mathbf{X}^\top \frac{\partial \pi(\mathbf{X},\alpha)}{\partial \alpha}\right],$$

where $v(t) = \mathbb{E}\left[\frac{\delta \mathbb{1}(Y>t)\mathbf{X}^\top [D-\pi(\mathbf{X},\alpha))]\int_t^Y S_C(v)\,\mathrm{d}v}{w^2(Y)}\right]$. *Let* $\mathbb{P}_n$ *be the empirical average. The components of the variance–covariance matrix* $\eta(\alpha)$ *can be consistently estimated by*

$$\hat{\Sigma} = \mathbb{P}_n\left[\left\{\delta \mathbf{X}^\top \frac{D-\pi(\mathbf{X},\alpha)}{\hat{w}(Y)} + \int_0^s \frac{\hat{v}(t)\,\mathrm{d}\hat{M}_C(t)}{\hat{S}_C(t)\hat{S}_R(t)}\right\}^{\otimes 2}\right],$$
$$\hat{\Lambda} = \mathbb{P}_n\left[\frac{\partial U_i(\alpha)}{\partial \alpha}\right] = \mathbb{P}_n\left[\frac{\delta}{\hat{w}(Y)}\mathbf{X}^\top \frac{\partial \pi(\mathbf{X},\alpha)}{\partial \alpha}\right].$$

Also, the stochastic process $M_C(\mathrm{s})$ *can be estimated by replacing the* $\Lambda_C(.)$ *by its estimate* $\hat{\Lambda}_C(.)$. □

Acknowledgements

This research was supported in part by NIDA grant P50 DA010075. The second and third authors acknowledge the support of Discovery Grants from the Natural Sciences and Engineering Research Council (NSERC) of Canada.

References

Asgharian, M, M'Lan, CE & Wolfson, DB (2002), 'Length-biased sampling with right censoring', *Journal of the American Statistical Association*, **97**(457), 201–209.

Asgharian, M & Wolfson, DB (2005), 'Asymptotic behavior of the unconditional NPMLE of the length-biased survivor function from right censored prevalent cohort data', *The Annals of Statistics*, **33**(5), 2109–2131.

Asgharian, M, Wolfson, DB & Zhang, X (2006), 'Checking stationarity of the incidence rate using prevalent cohort survival data', *Statistics in Medicine*, **25**(10), 1751–1767.

Bergeron, PJ, Asgharian, M & Wolfson, DB (2008), 'Covariate bias induced by length-biased sampling of failure times', *Journal of the American Statistical Association*, **103**(482), 737–742.

Cheng, Y & Wang, M (2012), 'Estimating propensity scores and causal survival functions using prevalent survival data', *Biometrics*, **68**, 707–716.

Cochran, W & Rubin, D (1973), 'Controlling bias in observational studies: a review', *Sankhyā: The Indian Journal of Statistics, Series A*, **35**, 417–446.

Cole, SR & Hernán, MA (2004), 'Adjusted survival curves with inverse probability weights', *Computer Methods and Programs in Biomedicine*, **75**(1), 45–49.

(wileyonlinelibrary.com) DOI: 10.1002/sta4.46

Cox, DR (1969), 'Some sampling problems in technology', in New Developments in Survey Sampling, Wiley Interscience, New York, 506–527.

Cox, DR & Lewis, P (1966), *The Statistical Analysis of Series of Events*, *John Wiley and Sons*, London.

De Uña-álvarez, J (2004), 'Nonparametric estimation under length-biased sampling and type I censoring: a moment based approach', *Annals of the Institute of Statistical Mathematics*, **56**(4), 667–681.

Fisher, RA (1934), 'The effect of methods of ascertainment upon the estimation of frequencies', *Annals of Human Genetics*, **6**(1), 13–25.

Hernán, M, Brumback, B & Robins, J (2000), 'Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men', *Epidemiology*, **11**(5), 561–570.

Huang, CY & Qin, J (2011), 'Nonparametric estimation for length-biased and right-censored data', *Biometrika*, **98**(1), 177–186.

Luo, X & Tsai, W (2009), 'Nonparametric estimation for right-censored length-biased data: a pseudo-partial likelihood approach', *Biometrika*, **96**(4), 873–886.

Neyman, J (1955), 'Statistics—servant of all science', *Science*, **122**(3166), 401–406.

Neyman, J (1990), 'On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Translation of excerpts by D. Dabrowska and T. Speed', *Statistical Science*, **6**, 462–47.

Nieto, FJ & Coresh, J (1996), 'Adjusting survival curves for confounders: a review and a new method', *American Journal of Epidemiology*, **143**(10), 1059–1068.

Ning, J, Qin, J & Shen, Y (2010), 'Non-parametric tests for right-censored data with biased sampling', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72**(5), 609–630.

Pepe, MS & Fleming, TR (1991), 'Weighted Kaplan–Meier statistics: large sample and optimality considerations', *Journal of the Royal Statistical Society. Series B (Methodological)*, **53**(2), 341–352.

Qin, J & Shen, Y (2010), 'Statistical methods for analyzing right-censored length-biased data under Cox model', *Biometrics*, **66**(2), 382–392.

Robins, J (1997), 'Causal inference from complex longitudinal data', *Latent Variable Modeling and Applications to Causality*, **120**, 69–117.

Robins, J, Hernán, M & Brumback, B (2000), 'Marginal structural models and causal inference in epidemiology', *Epidemiology*, **11**(5), 550–560.

Robins, JM (1994), 'Correcting for non-compliance in randomized trials using structural nested mean models', *Communications in Statistics-Theory and Methods*, **23**(8), 2379–2412.

Rosenbaum, PR & Rubin, DB (1983), 'The central role of the propensity score in observational studies for causal effects', *Biometrika*, **70**(1), 41–55.

Rotnitzky, A & Robins, JM (2005), 'Inverse probability weighting in survival analysis', in *Encyclopedia of Biostatistics*, 2nd edn, Vol. 4, Wiley, New York, 2619–2625.

Rubin, D (1973), 'The use of matched sampling and regression adjustment to remove bias in observational studies', *Biometrics*, **29**, 185–203.

Rubin, DB (1978), 'Bayesian inference for causal effects: the role of randomization', *The Annals of Statistics*, **6**(1), 34–58.

Shen, Y, Ning, J & Qin, J (2009), 'Analyzing length-biased data with semiparametric transformation and accelerated failure time models', *Journal of the American Statistical Association*, **104**(487), 1192–1202.

Wang, MC (1991), 'Nonparametric estimation from cross-sectional survival data', *Journal of the American Statistical Association*, **86**(413), 130–143.

Wang, MC, Jewell, NP & Tsai, WY (1986), 'Asymptotic properties of the product limit estimate under random truncation', *The Annals of Statistics*, **14**(4), 1597–1605.

Wicksell, SD (1925), 'The corpuscle problem: a mathematical study of a biometric problem', *Biometrika*, **17**(1/2), 84–99.

Wolfson, C, Wolfson, DB, Asgharian, M, M'Lan, CE, Østbye, T, Rockwood, K & Hogan, DB (2001), 'A reevaluation of the duration of survival after the onset of dementia', *New England Journal of Medicine*, **344**(15), 1111–1116.

Xie, J & Liu, C (2005), 'Adjusted Kaplan–Meier estimator and log-rank test with inverse probability of treatment weighting for survival data', *Statistics in Medicine*, **24**(20), 3089–3110.

Zelen, M & Feinlein, M (1969), 'On the theory of screening for chronic diseases', *Biometrika*, **56**(3), 601–614.