Published in final edited form as:
Biometrics. 2010 June ; 66(2): 382–392. doi:10.1111/j.1541-0420.2009.01287.x.

Statistical Methods for Analyzing Right-censored Length-biased Data under Cox Model

Jing Qin[1] and **Yu Shen**[2,*]

[1]Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases Bethesda, MD 20892

[2]Department of Biostatistics, M. D. Anderson Cancer Center, Houston, Texas 77030, USA

Summary

Length-biased time-to-event data are commonly encountered in applications ranging from epidemiologic cohort studies or cancer prevention trials to studies of labor economy. A longstanding statistical problem is how to assess the association of risk factors with survival in the target population given the observed length-biased data. In this paper, we demonstrate how to estimate these effects under the semiparametric Cox proportional hazards model. The structure of the Cox model is changed under length-biased sampling in general. Although the existing partial likelihood approach for left-truncated data can be used to estimate covariate effects, it may not be efficient for analyzing length-biased data. We propose two estimating equation approaches for estimating the covariate coefficients under the Cox model. We use the modern stochastic process and martingale theory to develop the asymptotic properties of the estimators. We evaluate the empirical performance and efficiency of the two methods through extensive simulation studies. We use data from a dementia study to illustrate the proposed methodology, and demonstrate the computational algorithms for point estimates, which can be directly linked to the existing functions in S-PLUS or R.

Keywords

Cox model; Dependent censoring; Estimating equation; Length-biased

1. Introduction

In observational studies, such as studies of unemployment duration in the labor economy (Lancaster, 1979; de Una-Alvarez, Otero-Giraldez, and Alvarez-Liorente, 2003), cancer screening trials (Zelen and Feinleib, 1969; Zelen, 2004), and HIV prevalent cohort studies (Lagakos, Barraj, and De Gruttola, 1988), one often encounters right-censored time-to-event data subject to length-biased sampling. Length-biased sampling is a special case of left truncation. Following the terminology in the literature, length-biased data are defined for left-truncated and right-censored data under the stationarity assumption, which assumes that the initiation times follow a stationary Poisson process. As a result, the probability of observing a failure time *t* is proportional to *t* itself.

* yshen@mdanderson.org .

Supplementary Materials

Web Appendices referenced in Section 2.3 and the APPENDIX are available under the Paper Information link at the *Biometrics* website http://www.biometrics.tibs.org.

Length-bias is one of the major biases that are difficult to remove by trial design and may confound the interpretation of disease-specific survival. In a randomized cancer screening trial, the observed survival benefit for individuals whose disease is detected by screening versus by symptoms can be confounded with length bias. This is because cancers with a longer duration of preclinical disease are more likely to be detected by screening examination, and are thus overrepresented among the screen-detected cases. Moreover, cancers with a longer preclinical duration (i.e., with slower growing tumors) are often associated with more favorable prognoses. Another example of length-biased data can be seen in a study of dementia among elderly people. In the Canadian Study of Health and Aging (CSHA), a total of 14,026 subjects who were 65 years or older were randomly selected throughout Canada, and 10,263 agreed to participate in this multicenter epidemiologic study (Wolfson et al., 2001). Among the participants, 1,132 were identified as having dementia and were followed until the end of the study. The investigators noted that patients who had experienced a longer duration of dementia symptoms at the time of recruitment to the CSHA tended to live longer (Wolfson, et al., 2001). That is, the sampled cases were subject to length-bias. It is of great interest to investigate how different types of dementia may impact long-term survival in a regression analysis after adjusting for length-biased sampling.

In the aforementioned example, the observed time from the diagnosis of the disease to the subsequent event (death) is subject to length-biased sampling. The outcome of interest is time from disease diagnosis to death. The data set we considered is a prevalent cohort consisting of subjects with the disease of interest at the examination time who were then followed for a subsequent terminal event (e.g., death). The data on each subject in the cohort include an initiating event (e.g., diagnosis of the disease) and a failure event (disease recurrence or death) for those subjects who have been accrued. Apparently, the failure times will be left truncated if the initiating event is not observed or the failure event occurs before sampling time, and the failure times can be right censored during follow-up. Length-biased sampling occurs in this setting because the “observed” time intervals from initiation to failure in the prevalent cohort tend to be longer than those in the target population.

Methodology development has focused on nonparametric estimation of the length-biased distribution in one-sample problems (Turnbull, 1976; Vardi, 1982, 1985, 1989; Gill, Vardi, and Wellner, 1988; Lagakos et al., 1988; Wang, 1989; Asgharian, MLan, and Wolfson, 2002; Asgharian and Wolfson, 2005). One complication in analyzing such data is the potential dependence between the failure time and the right-censoring time, measured from the initiating event (diagnosis) to the event of interest. The informative censoring induced by the sampling scheme has been avoided by prohibiting right censoring (Vardi, 1985; Wang, 1996) or by simply ignoring it. A second complication occurs when evaluating the covariate effects on the time interval measured from the diagnosis of the disease to the event of failure for the target population. This evaluation proves to be difficult because the model structure assumed for the target population is often different from the one for the observed length-biased data. Recently, Shen, Ning, and Qin (2009) proposed some methods for modeling covariate effects for length-biased data under transformation and accelerated failure time models. Bergeron, Asgharian, and Wolfson (2008) assessed the bias induced for covariate estimates under length-biased sampling using a full-likelihood approach with a parametric model.

Coxs proportional hazards model has been widely used to model the risk factors of a failure time in classical survival analyses (Cox, 1972). On the other hand, it has been noted in the literature that conventional regression methods, such as the standard Cox partial likelihood method, may produce biased estimators if right-censored data is subject to biased sampling. The partial likelihood approach proposed for left-truncated data can be applied to estimate the covariate effects for length-biased data under the Cox model (e.g., Kalbfleisch and Lawless, 1991; Keiding, 1992; Wang, Brookmeyer, and Jewell, 1993). However, the efficiency of the estimators may not be ideal because the important information pertaining to the stationary Poisson process is not utilized. Wang (1996) was the first to use the semiparametric proportional hazards model to estimate covariate effects when the observed failure times were length-biased; Wang used a bias-adjusted risk set for the construction of the pseudo-likelihood. However, a major restriction in her approach was the assumption that the length-biased data are not subject to right censoring. More recently, Ghosh (2008) proposed an estimating equation approach that allows right censoring of the length-biased data under a proportional hazards model. However, Ghosh assumed that the cross-sectional data did not have any follow-up. Therefore, Ghoshs proposed method may not be general enough or valid if there are follow-up data subject to right censoring.

Given the popularity and importance of the Cox regression model for analyzing survival data, the aims of this work are to propose two inference methods to assess the covariate effects under the semiparametric Cox model for length-biased data subject to right censoring, and to compare the proposed methods with the conditional approach for left-truncated data. The proposed methods are based on the generalized estimating equations. One major advantage of the proposed methods is computational simplicity. The estimation algorithms can be directly linked to existing S-PLUS, R, or SAS codes for the Cox model by adding appropriate weights for the linear predictor in the function. The remainder of this paper is organized as follows. In Section 2 we introduce the basic notation and the estimating equations and provide inference procedures and theoretical properties for the proposed estimators. In Section 3 we evaluate the performance of the proposed estimators and compare them with existing methods through simulation studies. We also illustrate the methods through application to the demential data example and demonstrate the computational algorithms for point estimates, which can be directly linked to the existing functions in S-PLUS or R. We provide concluding remarks in Section 4 and proofs of the theorems in the Appendix.

2. Estimation Methods

2.1 Data and Model

Assume $\tilde{T}$ failure, to be the duration from the initiating event (diagnosis or onset of the disease) to A to be the duration from the initiating event to examination, V to be the duration from examination to failure, and C to be the duration from examination to censoring. Under length-biased sampling, one can only observe T among those $\tilde{T} >$ A. Let $T = A + V$ be a positive lifetime random variable, where A is the truncation variable (or backward recurrence time), V is the residual survival time (or forward recurrence time), and $\boldsymbol{X}$ is the baseline covariate vector. It is reasonable to assume that C and (A, V) are independent, and that the censoring distribution is independent of covariate $\boldsymbol{X}$.

For a random sample of n independent subjects, the observed data consist of $\{(A_i, Y_i, \delta_i, \boldsymbol{X}_{\mathrm{i}}),$ $i = 1, \ldots, n\}$, where $Y_i = \min(T_i, A_i + C_i)$, $T_i = A_i + V_i$, and $\delta_i = I(V_i \leqslant C_i)$. Let f represent the unbiased density for $\tilde{T}$, and g represent the length-biased density (conditional on $\tilde{T} > A$). Then, for the observed length-biased data T, its density function g is associated with the unbiased density f, as follows:

$$g(t)=\frac{tf(t)}{\mu}, \quad \mu=\int_0^\infty uf(u)\,du.$$

Given the covariates, $\boldsymbol{X} = \boldsymbol{x}$, the density of T can be expressed as

$$g(t|\boldsymbol{x}) = \frac{tf(t|\boldsymbol{x})}{\int_0^\infty uf(u|\boldsymbol{x})\,du} := \frac{tf(t|\boldsymbol{x})}{\mu(\boldsymbol{x})},$$

where $g(t|\boldsymbol{x})$ and $f(t|\boldsymbol{X})$ denote the covariate-specific length-biased sampling density and the population (unbiased) density. Assume that failure times in the target population (unbiased), $\tilde{T}$, follow the proportional hazards model

$$\lambda(t|\boldsymbol{x}) = \lambda_0(t)\ \exp\left(\beta_0^T \boldsymbol{x}\right), \tag{1}$$

where $\lambda_0(t)$ is an unspecified baseline hazards function and $\boldsymbol{\beta}_0$ is a vector-valued unknown regression coefficient for $\boldsymbol{X}$.

Likelihood principal—In order to better understand the structure of length-biased data, we start with the bivariate observation for A and T. Given the covariate $\boldsymbol{X} = \boldsymbol{x}$, the joint density of (A, T) can be decomposed as a product of the marginal distribution of A and the conditional distribution of T given A. Such a formulation has been utilized in analyzing left-truncated data (e.g., Andersen et al., 1993, pp 166-167; Wang, 1989):

$$f_{A,T}(a, t|\boldsymbol{x}) = f_A(a|\boldsymbol{x})\, f_{T|A}(t|a, \boldsymbol{x}) = \left[\frac{S_U(a|\boldsymbol{x})}{\mu(\boldsymbol{x})}\right]\left[\frac{f(t|\boldsymbol{x})\, I(t>a)}{S_U(a|\boldsymbol{x})}\right],$$

where $S_U(t|\boldsymbol{x})$ is the survival distribution for the unbiased failure time given $\boldsymbol{x}$. Given truncation time $A = a$, the conditional likelihood of Y is proportional to

$$L_C = \prod_{i=1}^{n} \frac{f(y_i|\boldsymbol{X}_i, \beta)^{\delta_i} S_U(y_i|\boldsymbol{X}_i, \beta)^{1-\delta_i}}{S_U(a_i|\boldsymbol{X}_i, \beta)}. \tag{2}$$

As described in detail by Wang et al. (1993), L_C can be further expressed as the product of a partial likelihood and the residual likelihood:

$$L_C = L_P(\beta)\, L_R(\beta, \lambda_0),$$

where

$$L_P(\beta) = \prod_i \left[\frac{\exp\left(\beta^T \boldsymbol{X}_i\right)}{\Sigma_{j \in R_{(y_i)}} \exp\left(\beta^T \boldsymbol{X}_j\right)}\right]^{\delta_i}, \tag{3}$$

and the residual likelihood, $L_R(\boldsymbol{\beta},\lambda_0)$ is referred to as an “ancillary” term by Wang et al. (1993), which includes the baseline hazard function λ_0 and $\boldsymbol{\beta}$. Under the Cox model, L_P has an expression similar to that of the partial likelihood function for traditional survival data (without left truncation) except for the definition of the risk sets $R(y) = \{j : a_j \leqslant y \leqslant y_j\}$. Intuitively, the ignored information for covariates (i.e., $\boldsymbol{\beta}$) contained in the marginal distribution of A and the residual likelihood may lead to a loss of efficiency.

2.2 Estimating Equation Approaches

We start with a special case in modeling covariate effects for length-biased data without right censoring. Let the marginal density function of covariate $\boldsymbol{X}$ be denoted as $h(\boldsymbol{x})$. Then the conditional distribution of $\boldsymbol{X}$ given $T = t$ follows:

$$h(x|T=t)=\frac{g(t|\boldsymbol{x})\,h(\boldsymbol{x})}{\int g(t|\boldsymbol{x})\,h(\boldsymbol{x})\,d\boldsymbol{x}}=\frac{tf(t|\boldsymbol{x})\,h(\boldsymbol{x})\,/\mu(\boldsymbol{x})}{\int tf(t|\boldsymbol{x})\,h(\boldsymbol{x})\,/\mu(\boldsymbol{x})\,d\boldsymbol{x}}.$$

Thus, under the proportional hazards model, the conditional expectation of $\boldsymbol{x}$ is

$$\begin{aligned}E[\boldsymbol{X}|T=t] &= \frac{\int \boldsymbol{x}tf(t|\boldsymbol{x})\,h(\boldsymbol{x})\,/\mu(\boldsymbol{x})\,d\boldsymbol{x}}{\int tf(t|\boldsymbol{x})\,h(\boldsymbol{x})\,/\mu(\boldsymbol{x})\,d\boldsymbol{x}}\\ &= \frac{\int \boldsymbol{x}\exp\left(\beta^T\boldsymbol{x}\right)S_U(t|\boldsymbol{x})\,h(\boldsymbol{x})\,/\mu(\boldsymbol{x})\,d\boldsymbol{x}}{\int \exp\left(\beta^T\boldsymbol{x}\right)S_U(t|\boldsymbol{x})\,h(\boldsymbol{x})\,/\mu(\boldsymbol{x})\,d\boldsymbol{x}}\\ &= \frac{E\left[\boldsymbol{X}\exp(\beta^T\boldsymbol{X})S_U(t|\boldsymbol{X})/\mu(\boldsymbol{X})\right]}{E\left[\exp(\beta^T\boldsymbol{X})S_U(t|\boldsymbol{X})/\mu(\boldsymbol{X})\right]}.\end{aligned}$$

The second equation holds because $\lambda_0(t)$ is canceled out. Using the fact that

$$E\left[T^{-1}I(T>t)\,|\boldsymbol{X}\right]=S_U(t|\boldsymbol{X})\,/\mu(\boldsymbol{X}), \tag{4}$$

we obtain

$$E[\boldsymbol{X}|T=t]=\frac{E\left[\boldsymbol{X}\exp\left(\beta^T\boldsymbol{X}\right)T^{-1}I(T\geqslant t)\right]}{E\left[\exp(\beta^T\boldsymbol{X})\,T^{-1}I(T\geqslant t)\right]}.$$

Therefore, we can construct the following unbiased estimating equation to estimate $\boldsymbol{\beta}$:

$$\sum_{i=1}^{n}\left[\boldsymbol{X}_i-\frac{\Sigma_{j=1}^{n}\boldsymbol{X}_j\exp\left(\beta^T\boldsymbol{X}_j\right)T_j^{-1}I\left(T_j\geqslant T_i\right)}{\Sigma_{j=1}^{n}\exp\left(\beta^T\boldsymbol{X}_j\right)T_j^{-1}I\left(T_j\geqslant T_i\right)}\right]=0.$$

In fact, the above estimating equation is the same as the score equation derived from the pseudo-likelihood function by Wang (1996). By generalizing the above derivations to length-biased data with right censoring, we propose two estimating equation approaches.

Estimating Equation I—When length-biased failure time T is subject to right censoring, a natural extension of the above estimating equation can be proposed as follows. Recall that the joint density distributions of (A, V) and (A, T) given covariate $\boldsymbol{X} = \boldsymbol{x}$ have the same formula without censoring (Asgharian and Wolfson, 2005):

$$f_{A,V}(a,v|\boldsymbol{X}=\boldsymbol{x})=\frac{f(t|\boldsymbol{x})}{\mu(\boldsymbol{x})},\;t=a+v>0.$$

With potential censoring, the probability of observing a pair of uncensored data is

$$\begin{aligned} & \mathrm{pr}\left(A=a, V=y-a, C \geqslant y-a \mid \boldsymbol{X}=\boldsymbol{x}\right) \\ = & \mathrm{pr}\left(A=a, V=y-a \mid \boldsymbol{X}=\boldsymbol{x}\right) \mathrm{pr}\left(C \geqslant y-a\right) \\ = & f\left(y \mid \boldsymbol{x}\right) S_{c}\left(y-a\right) / \mu\left(\boldsymbol{x}\right), \end{aligned} \tag{5}$$

where S_C is the survival distribution for censoring variable C, assuming that the right-censoring variable C is independent of covariate $\boldsymbol{X}$. Using a concept similar to (4), we have the following conditional expectation when there is right censoring:

$$\begin{aligned} & E\left[\frac{\delta I(Y \geqslant y)}{Y S_{C}(Y-A)} \mid \boldsymbol{X}=\boldsymbol{x}\right] \\ = & \int_{y}^{\infty} f\left(t \mid \boldsymbol{x}\right) \int_{0}^{t} S_{c}\left(t-a\right) t^{-1} S_{c}^{-1}\left(t-a\right) da dt / \mu\left(\boldsymbol{x}\right) \\ = & \int_{y}^{\infty} f\left(t \mid \boldsymbol{x}\right) / \mu\left(\boldsymbol{x}\right) dt = S_{U}\left(y \mid \boldsymbol{x}\right) / \mu\left(\boldsymbol{x}\right). \end{aligned} \tag{6}$$

In addition, utilizing equation (6), we can replace $S_U(y|\boldsymbol{X})/\mu(\boldsymbol{X})$ the following conditional expectation:

$$\begin{aligned} & E\left[\boldsymbol{X} \mid Y=y, \delta=1, A=a\right] \\ = & \frac{\int \boldsymbol{x} f\left(y \mid \boldsymbol{x}\right) S_{C}\left(y-a\right) h\left(\boldsymbol{x}\right) / \mu\left(\boldsymbol{x}\right) d\boldsymbol{x}}{\int f\left(y \mid \boldsymbol{x}\right) S_{C}\left(y-a\right) h\left(\boldsymbol{x}\right) / \mu\left(\boldsymbol{x}\right) d\boldsymbol{x}} \\ = & \frac{E\left[\boldsymbol{X} \exp(\beta^{T} \boldsymbol{X}) S_{U}(y \mid \boldsymbol{X}) / \mu(\boldsymbol{X})\right]}{E\left[\exp(\beta^{T} \boldsymbol{X}) S_{U}(y \mid \boldsymbol{X}) / \mu(\boldsymbol{X})\right]}. \end{aligned} \tag{7}$$

Combining (6) and (7) leads to the following estimating equation:

$$\boldsymbol{U}_{1}\left(\beta\right)=\sum_{i=1}^{n} \delta_{i}\left[\boldsymbol{X}_{i}-\frac{\Sigma_{j=1}^{n} I\left(Y_{j} \geqslant Y_{i}\right) \delta_{j} \boldsymbol{X}_{j} \exp\left(\beta^{T} \boldsymbol{X}_{j}\right)\left\{Y_{j} S_{c}\left(Y_{j}-A_{j}\right)\right\}^{-1}}{\Sigma_{j=1}^{n} I\left(Y_{j} \geqslant Y_{i}\right) \delta_{j} \exp\left(\beta^{T} \boldsymbol{X}_{j}\right)\left\{Y_{j} S_{c}\left(Y_{j}-A_{j}\right)\right\}^{-1}}\right]=0. \tag{8}$$

When S_C is unknown, we can replace it with its consistent Kaplan-Meier estimator for residual censoring time, which leads to an asymptotic unbiased estimating equation we call EE-I.

Estimating Equation II—An alternative estimating equation approach with a different weight can be proposed. Given (5), we can express the probability of observing the length-biased failure time at y by integrating out a:

$$\mathrm{pr}\left(Y=y, \delta=1 \mid \boldsymbol{X}=\boldsymbol{x}\right)=\frac{f_{U}\left(y \mid \boldsymbol{x}\right) w_{c}\left(y\right)}{\mu\left(\boldsymbol{x}\right)}, \tag{9}$$

where $w_{c}\left(y\right)=\int_{0}^{y} S_{c}\left(t\right) dt$. Based on (9), we have

$$\begin{aligned} & E\left[I\left(Y>y, \delta=1\right) / w_{c}\left(Y\right) \mid \boldsymbol{X}=\boldsymbol{x}\right] \\ = & \int_{y}^{\infty} f\left(t \mid \boldsymbol{x}\right) dt \int_{0}^{t} S_{c}\left(v\right) dv / w_{c}\left(t\right) / \mu\left(\boldsymbol{x}\right) \\ = & \int_{y}^{\infty} f\left(t \mid \boldsymbol{x}\right) dt / \mu\left(\boldsymbol{x}\right) = S_{U}\left(y \mid \boldsymbol{x}\right) / \mu\left(\boldsymbol{x}\right). \end{aligned} \tag{10}$$

Therefore, we can replace $S_U(y|\boldsymbol{X})/\mu(\boldsymbol{X})$ in equation (7) for the conditional expectation of $\boldsymbol{X}$ with the corresponding observed data to construct the following unbiased estimating equation to estimate $\boldsymbol{\beta}$:

$$\boldsymbol{U}_2(\beta)=\sum_{i=1}^{n}\delta_i\left[\boldsymbol{X}_i-\frac{\Sigma_{j=1}^{n}I\left(Y_j\geqslant Y_i\right)\delta_j\left\{w_c\left(Y_j\right)\right\}^{-1}\boldsymbol{X}_j\exp\left(\beta^T\boldsymbol{X}_j\right)}{\Sigma_{j=1}^{n}I\left(Y_j\geqslant Y_i\right)\delta_j\left\{w_c\left(Y_j\right)\right\}^{-1}\exp\left(\beta^T\boldsymbol{X}_j\right)}\right]=0. \tag{11}$$

By replacing $w_c(t)$ with its consistent estimator, we have an asymptotic unbiased estimating equation, which we call EE-II.

Estimating Equation LT—Clearly the above two estimating equations require the information of the distribution function for censoring variable C. In contrast, an approach proposed for delayed-entry/left-truncated data does not require estimating the survival function for the censoring variable. Conditional on $\boldsymbol{X}$, one has

$$\begin{aligned} & E\left[I(Y\geqslant y, A\leqslant y, \delta=1)\,|\boldsymbol{X}=\boldsymbol{x}\right] \\ = & E\left[I(T\geqslant y, C+A\geqslant y, A\leqslant y)\,|\boldsymbol{X}=\boldsymbol{x}\right] \\ = & \int_y^{\infty}\int_0^{y} f(t|\boldsymbol{x})/\mu(\boldsymbol{x})S_C(t-a)\,dadt \\ = & S_U(y|\boldsymbol{x})\,w_c(y)/\mu(\boldsymbol{x}). \end{aligned}$$

Therefore, when censoring variable C is independent of covariate $\boldsymbol{X}$, $w_c(.)$ is canceled out on the right side of the following equation: Conditional on $\boldsymbol{X}$, one has

$$\begin{aligned} & \frac{E\left[\boldsymbol{X}\exp(\beta^T\boldsymbol{X})I(T\geqslant y, A\leqslant y, \delta=1)\right]}{E\left[\exp(\beta^T\boldsymbol{X})I(Y\geqslant y, A\leqslant y, \delta=1)\right]} \\ = & \frac{E\left[\boldsymbol{X}\exp(\beta^T\boldsymbol{X})(y|\boldsymbol{X})/\mu(\boldsymbol{X})\right]}{E\left[\exp(\beta^T\boldsymbol{X})S_U(y|\boldsymbol{X})/\mu(\boldsymbol{X})\right]}. \end{aligned}$$

Similar to the constructions of the first two estimating equations, one can construct the estimating equation

$$\boldsymbol{U}_L(\beta)=\sum_{i=1}^{n}\delta_i\left[\boldsymbol{X}_i-\frac{\Sigma_{j=1}^{n}I\left\{y_j\geqslant y_i, a_j\leqslant y_i\right\}\boldsymbol{X}_j\exp\left(\beta^T\boldsymbol{X}_{\boldsymbol{j}}\right)}{\Sigma_{j=1}^{n}I\left\{y_j\geqslant y_i, a_j\leqslant y_j\right\}\exp\left(\beta^T\boldsymbol{X}_j\right)}\right], \tag{12}$$

which is the score equation of the partial likelihood of (3) for left-truncated data under the Cox model (Kalbfleisch and Lawless, 1991;Andersen et al., 1993;Wang et al., 1993). We call this equation EE-LT. Unlike EE-I and EE-II, the summations in the fraction terms of EE-LT can include both failure and censored times as long as the pair (a_j, y_j) satisfies the inequality condition. The large sample properties for the above estimating equation have been explored in the literature (Wang, 1989;Wang et al., 1993).

2.3 Asymptotic Properties

The consistency and weak convergence of $\boldsymbol{\beta}$ can be established for estimating equations EE-I and EE-II under the regularity conditions stated in the Appendix. Using the counting process notation of Andersen et al. (1993), for the ith subject, define risk set $R_i(t) = I\{Y_i \geqslant t\}\delta_i$ and $N_i(t) = I\{Y_i \leqslant t, C_i \geqslant Y_i - A_i\}$. Define

$$S_k^{(l)}(\beta,t)=n^{-1}\sum_{i=1}^{n} w_c(t)\,R_i(t)\,W_{ki}\boldsymbol{X}_i^{\otimes l}\exp\left(\beta^T\boldsymbol{X}_i\right),$$

where $k = 1$ for EE-I and $k = 2$ for EE-II, $l = 0, 1, 2$, and

$$W_{1i}=\{Y_i S_c(Y_i-A_i)\}^{-1},\quad W_{2i}=\{w_c(Y_i)\}^{-1}.$$

Also, let

$$\boldsymbol{E}_k(\beta,t)=\frac{S_k^{(1)}(\beta,t)}{S_k^{(0)}(\beta,t)},$$

$\boldsymbol{e}_k(\boldsymbol{\beta}, t)$ be the expectation of $E_k(\boldsymbol{\beta}, t)$ $s_k^{(l)}(\boldsymbol{\beta}, t)$ be the expectation of $S_k^{(l)}(\boldsymbol{\beta}, t)$.

Estimating Equation I for $\hat{\boldsymbol{\beta}}_1$—By generalizing the theoretical formulation of Wang (1996) to the setting with right censoring, we can construct

$$M_{1i}(t)=N_i(t)-\int_0^t w_c(y)\,R_i(y)\,W_{1i}\exp\left(\beta_0^T\boldsymbol{X}_i\right)d\Lambda_0(y)$$

and prove it to be a mean zero stochastic process. Specifically, using equation (6),

$$\begin{aligned} EM_{1i}(t)= &\ E\left[N_i(t)-\int_0^t E\left\{\frac{\delta_i I(Y_i\geqslant y)}{Y_i S_C(Y_i-A_i)}|\boldsymbol{X}_i\right\}w_c(y)\,\lambda(y|\boldsymbol{X}_i)\,dy\right] \\ = &\ E\left[\int_0^t dy\int_0^y\frac{f(y|\boldsymbol{X}_i)}{\mu(\boldsymbol{X}_i)}S_c(y-a)\,da\right]-E\left[\int_0^t\frac{S_U(y|\boldsymbol{X}_i)}{\mu(\boldsymbol{X}_i)}w_c(y)\,\lambda(y|\boldsymbol{X}_i)\,dy\right]=0. \end{aligned}$$

Therefore, if S_C is a known function, estimating equation (8) can be asymptotically represented by the following independent and identically-distributed (i.i.d.) summation of the mean zero process:

$$\boldsymbol{U}_1(\beta_0)=\sum_{i=1}^{n}\int_0^\tau\{\boldsymbol{X}_i-\boldsymbol{e}_1(\beta_0,t)\}\,dM_{1i}(t)+o_p(1). \tag{13}$$

To obtain an estimator for $\boldsymbol{\beta}$, we replace $S_C(t)$ with its consistent Kaplan-Meier estimator $\hat{S}_C(t)$ for the censoring time in (13). We then have the estimating equation

$$\tilde{U}_1(\beta)=\sum_{i=1}^{n}\int_0^\tau\left\{\boldsymbol{X}_i-\tilde{E}_1(\beta,t)\right\}\Bigg\}\,dN_i(t), \tag{14}$$

where

$$\tilde{E}_1(\beta,t)=\frac{\tilde{S}_1^{(1)}(\beta,t)}{\tilde{S}_1^{(0)}(\beta_0,t)},\quad \widehat{W}_{1i}=\left\{Y_i\widehat{S}_c(Y_i-A_i)\right\}^{-1}$$

$$\tilde{S}_1^{(l)}(\beta,t)=\frac{1}{n}\sum_{i=1}^{n}\widehat{w}_c(t)\,R_i(t)\,\widehat{W}_{1i}\exp\left(\beta^T\boldsymbol{X}_i\right)\boldsymbol{X}_i^{\otimes l},\quad \widehat{w}_c(t)=\int_0^t\widehat{S}_c(u)\,du.$$

In the Appendix and Section 1 of the Supplementary Materials, we show that under the regularity conditions there exists a unique solution to the equations $\tilde{\boldsymbol{U}}_1(\boldsymbol{\beta}) = 0$, and $\| \widehat{\beta}_1 - \beta_0 \| \xrightarrow{P} 0$. Moreover, $n^{-1/2}\tilde{\boldsymbol{U}}_1(\boldsymbol{\beta})$ converges weakly to a mean zero Gaussian process with a variance-covariance function $\sum_1$. Let $\hat{\boldsymbol{\beta}}_1$ be the solution to equation (14). By Taylor series expansion,

$$n^{1/2}\left(\widehat{\beta}_1 - \beta_0\right) = \{\Gamma_1(\beta^*)\}^{-1} n^{-1/2} \tilde{U}_1(\beta_0) + o_p(1),$$

where $\boldsymbol{\beta}^*$ is on the line segment between $\hat{\boldsymbol{\beta}}^1$ and $\boldsymbol{\beta}_0$, and $\boldsymbol{\Gamma}_1(\boldsymbol{\beta}) = -n^{-1}\partial\tilde{\boldsymbol{U}}_1(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$.

Given the estimated $\hat{\boldsymbol{\beta}}_1$ for $\boldsymbol{\beta}_0$, a natural estimator for the cumulated baseline hazard function that is similar to Breslows estimator can be proposed:

$$\widehat{\Lambda}_{10}\left(t, \widehat{\beta}_1\right) = \int_0^t \frac{\Sigma_{i=1}^n dN_i(u)}{n\tilde{S}_1^{(0)}\left(\widehat{\beta}_1, u\right)}, \quad t \in [0, \tau].$$

The variance-covariance of $\boldsymbol{\beta}_1$ can be consistently estimated by $\widehat{\boldsymbol{\Gamma}}_1 \widehat{\boldsymbol{\Sigma}}_1^{-1} \widehat{\boldsymbol{\Gamma}}_1$, where

$$\begin{aligned}
\widehat{\boldsymbol{\Gamma}}_1 = n^{-1} \sum_{i=1}^n \int_0^\tau \left\{\boldsymbol{X}_i - \tilde{E}_1\left(\widehat{\beta}_1, t\right)\right\}^{\otimes 2} \widehat{w}_c(t) R_i(t) \widehat{W}_{1i} \exp\left(\widehat{\beta}_1^T \boldsymbol{X}_i\right) d\widehat{\Lambda}_{10}(t), \\
\widehat{\boldsymbol{\Sigma}}_1 = n^{-1} \sum_{i=1}^n \left[\int_0^\tau \left\{\left(\boldsymbol{X}_i - \tilde{E}_1\left(\widehat{\beta}_1, t\right)\right) d\widehat{M}_{1i}(t) + \frac{\boldsymbol{a}\left(\widehat{\beta}_1, t\right)}{\bar{Y}(t)} d\widehat{M}_{Ci}(t)\right\}\right]^{\otimes 2}, \\
\boldsymbol{a}\left(\widehat{\beta}_1, t\right) = -\lim_{n\to\infty} \frac{1}{n^2} \sum_{i=1}^n \sum_{k=1}^n \frac{I(Y_k - A_k \geqslant t)\delta_k \delta_i I(Y_k \geqslant Y_i)\widehat{w}_c(Y_i)\widehat{W}_{1k}\boldsymbol{X_k} \exp\left(\widehat{\beta}^T \boldsymbol{X_k}\right)}{\tilde{S}_1^{(0)}\left(\widehat{\beta}, Y_i\right)} \\
\bar{Y}(t) = \tfrac{1}{n}\Sigma_{i=1}^n I(\min(C_i, V_i) > t), \widehat{M}_{1i}(t) = N_i(t) - \int_0^t \widehat{w}_c(u) R_i(u) \widehat{W}_{1i} \exp\left(\widehat{\beta}_1^T \boldsymbol{X}_i\right) d\widehat{\Lambda}_{10}(u),
\end{aligned}$$

$\widehat{M}_{Ci}(t) = I(\min(V_i, C_i) \leqslant t, \delta_i = 0) - \int_0^t I(\min(V_i, C_i) \geqslant u) d\widehat{\Lambda}_c(u)$, and $\hat{\Lambda}_C(t)$ is the Nelson-Aalen estimator for residual survival time C.

Estimating Equation II for $\hat{\boldsymbol{\beta}}_2$—Similarly, we can prove that the following stochastic process has a mean of zero:

$$M_{2i}(t) = N_i(t) - \int_0^t w_c(y) R_i(y) W_{2i} \exp\left(\beta_0^T \boldsymbol{X}_i\right) d\Lambda_0(y),$$

Using equation (10)

$$\begin{aligned}
EM_{2i}(t) = & \; E\left[N_i(t) - \int_0^t E\left\{\frac{\delta_i}{w_c(Y_i)} I(Y_i \geqslant y) | \boldsymbol{X}_i\right\} w_c(y) \lambda(y|\boldsymbol{X}_i) dy\right] \\
= & \; E\left[\int_0^t dy \int_0^y \frac{f(y|\boldsymbol{X}_i)}{\mu(\boldsymbol{X}_i)} S_c(y-a) da\right] - E\left[\int_0^t w_c(y) \frac{S_U(y|\boldsymbol{X}_i)}{\mu(\boldsymbol{X}_i)} \lambda(y|\boldsymbol{X}_i) dy\right] = 0.
\end{aligned}$$

Therefore, if S_C is a known function, estimating equation (11) can be asymptotically represented by the following i.i.d. summation of the mean zero processes:

$$\boldsymbol{U}_2(\beta_0) = \sum_{i=1}^{n} \int_0^{\infty} \{\boldsymbol{X}_i - \boldsymbol{e}_2(\beta_0, t)\}\, d\boldsymbol{M}_{2i}(t) + o_p(1).$$

By replacing $w_c(t)$ with $\hat{w}_c(t)$, we can solve for $\boldsymbol{\beta}$ using the following estimating equation:

$$\tilde{U}_2(\beta) = \sum_{i=1}^{n} \int_0^{\tau} \left\{ \boldsymbol{X}_i - \tilde{E}_2(\beta, t)) \right\} dN_i(t),$$

where

$$\tilde{E}_2(\beta, t) = \frac{\tilde{S}_2^{(1)}(\beta, t)}{\tilde{S}_2^{(0)}(\beta, t)}, \text{ and } \tilde{S}_2^{(l)}(\beta, t) = \frac{1}{n} \sum_{i=1}^{n} \widehat{w}_c(t)\, R_i(t)\, \widehat{W}_{2i} \exp\left(\beta^T \boldsymbol{X}_i\right) \boldsymbol{X}_i^{\otimes l}.$$

In the Appendix and Section 2 of the Supplementary Materials, we prove that $n^{-1/2}\tilde{\boldsymbol{U}}_2(\boldsymbol{\beta}_0)$ converges weakly to a p-vector mean-zero Gaussian process with a covariance matrix $\sum_2$. In addition, the solution to the equation $\tilde{\boldsymbol{U}}_2(\boldsymbol{\beta}_0) = 0$ is consistent and unique. Using the Taylor series expansion,

$$n^{1/2}\left(\widehat{\beta}_2 - \beta_0\right) = \{\Gamma_2(\beta^*)\}^{-1} n^{-1/2} \tilde{U}_2(\beta_0) + o_p(1),$$

where $\boldsymbol{\beta}^*$ is on the line segment between $\hat{\boldsymbol{\beta}}_2$ and $\boldsymbol{\beta}_0$, $\boldsymbol{\Gamma}_2(\boldsymbol{\beta}) = -n^{-1}\partial\tilde{\boldsymbol{U}}_2(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$, and $\sum_2$ is the covariance matrix of $\lim_{n\to\infty}\tilde{\boldsymbol{U}}_2(\boldsymbol{\beta}_0)$.

The covariance matrix of $\hat{\boldsymbol{\beta}}$ can be consistently estimated by $\widehat{\boldsymbol{\Gamma}_2}^{-1}\widehat{\boldsymbol{\Sigma}}_2\widehat{\boldsymbol{\Gamma}_2}^{-1}$, where

$$\widehat{\boldsymbol{\Gamma}}_2 = n^{-1} \sum_{i=1}^{n} \int_0^{\tau} \left\{ \boldsymbol{X}_i - \tilde{E}_2\left(\widehat{\beta}_2, t\right) \right\}^{\otimes 2} \widehat{w}_c(t)\, R_i \widehat{W}_{2i} d\widehat{\Lambda}_{20}(t),$$

$$\widehat{\Lambda}_{20}(t) = \int_0^t \frac{\sum_{i=1}^n dN_i(u)}{n\tilde{S}_2^{(0)}\left(\widehat{\beta}_2, u\right)}, \quad t \in [0, \tau],$$

$$\widehat{\boldsymbol{\Sigma}}_2 = n^{-1} \sum_{i=1}^{n} \left[\int_0^{\tau} \left\{ \left(\boldsymbol{X}_i - \tilde{E}_2\left(t, \widehat{\beta}_2\right) \right) d\widehat{M}_{2i}(t) + \frac{\boldsymbol{H}\left(\widehat{\beta}_2, t\right)}{\bar{Y}(t)} d\widehat{M}_{Ci}(t) \right\} \right]^{\otimes 2},$$

$$\begin{aligned}
\widehat{M}_{2i}(t) = {} & N_i(t) \\
& - \int_0^t \widehat{w}_c(u)\, R_i(u)\, \widehat{W}_{2i} \exp\left(\widehat{\beta}_2^T \boldsymbol{X}_i\right) d\widehat{\Lambda}_{20}(u), \\
& \widehat{M}_{Ci}(t) \\
& = I(Y_i \leqslant t, \delta_i = 0) \\
& - \int_0^t I(Y_i \geqslant u)\, d\widehat{\Lambda}_c(u), \\
& \widehat{h}_k(t) \\
& = I(t \leqslant Y_k) \int_t^{Y_k} \widehat{S}_c(u)\, du
\end{aligned}$$

, and

$$\boldsymbol{H}\left(\widehat{\beta}_2, t\right)=\lim_{n\to\infty}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{k=1}^{n}\frac{\widehat{w}_c\left(Y_i\right)R_k\left(Y_i\right)\boldsymbol{X_k}\exp\left(\widehat{\beta}_2^T\boldsymbol{X_k}\right)W_{2k}^2\widehat{h}_k\left(t\right)}{\bar{S}_2^{(0)}\left(\widehat{\beta}_2, Y_i\right)}.$$

Remark—The weight functions W_{1i} and W_{2i} in the two estimating equations play a similar role in adjusting for the dependent censoring distribution in length-biased data. Theoretically, both weight functions are valid choices. In the next section, we will investigate the empirical performance of the estimators solved from EE-I and EE-II under various scenarios.

3. Numerical Studies

3.1 Simulations

We carried out a series of simulation studies to assess the efficiencies of the two proposed estimators relative to the estimator from $\boldsymbol{U}_L(\boldsymbol{\beta})$ and the estimator from the estimating equation by Ghosh (2008), which is specified as follows, EE-III

$$\boldsymbol{U}_3\left(\beta\right)=\sum_{i=1}^{n}\delta_i\left[\boldsymbol{X}_i-\frac{\Sigma_{j=1}^{n}I\left(Y_j\geqslant Y_i\right)\delta_j\boldsymbol{X}_j\exp\left(\beta^T\boldsymbol{X}_j\right)\left\{Y_jS_c\left(Y_j\right)\right\}^{-1}}{\Sigma_{j=1}^{n}I\left(Y_j\geqslant Y_i\right)\delta_j\exp\left(\beta^T\boldsymbol{X}_j\right)\left\{Y_jS_c\left(Y_j\right)\right\}^{-1}}\right]=0. \tag{15}$$

Note that (15) seems to have the same components of EE-I in (8), but the term $S_C(Y_i - A_i)$ in (8) is replaced by $S_C(Y_i)$ and information for the left-truncation time A is not used.

We generated unbiased failure times $\tilde{T}_i$ from the proportional hazards model

$$\lambda\left(t|\boldsymbol{X}\right)=\lambda_0\left(t\right)\ \exp\left(\alpha_1X_1+\alpha_2X_2\right),$$

where $\boldsymbol{\beta} = (\alpha_1, \alpha_2) = (1, 1)$ or $(2, 2)$, the binary covariate X_1 ~ Bernoulli(1,0.5), the continuous covariate X_2 ~ uniform(−0.5,0.5), and the baseline hazard function is $2t$. The left-truncation time A_i was independently generated from a uniform distribution $(0, \tau_0)$, and the pairs $(A_i, \tilde{T}_i)$ with $A_i < \tilde{T}_i$ were kept. We chose τ_0 larger than the upper bound of T_i to ensure the stationarity assumption. When τ_0 was large enough, and we let $\tau_0\int_{\tau_0}^{\infty}f_U\left(t\right)dt=\tau_0\left(1-F_U\left(\tau_0\right)\right)$ approximate to zero if $\tau_0 \to \infty$,

$$P\left(T=t|A<T\right)\approx\frac{tf\left(t\right)}{\int_0^{\infty}tf\left(t\right)dt}.$$

The censoring variables measured from the examination time were independently generated from uniform distributions corresponding to various censoring percentages: 20%, 35%, and 50%. The censoring indicator was obtained by $\delta_i = I(T_i \leqslant C_i + A_i)$. For each scenario, we repeated the simulation 1000 times with cohorts of size $n = 200$ or $n = 400$.

The simulation results are summarized in Table 1. When the right-censoring rate varies from low to moderate (20 to 35%), both EE-I and EE-II have unbiased estimators and reasonable coverage probabilities. In contrast, the estimators from (15) led to severe bias and poor coverage probabilities in all the scenarios we investigated. When the censoring percentage is small, which is equivalent to the censoring variable C being subject to heavy censoring, the proposed variance estimators derived from the weak convergence of $\tilde{\boldsymbol{U}}_2$ may slightly overestimate the true variance of $\hat{\boldsymbol{\beta}}$, which leads to an overestimated coverage probability. The reason for this is that the weight $\hat{w}_c(t)$ may be slightly overestimated under the heavy censoring of variable C. The overestimation for the variance estimator disappears when the right-censoring proportion increases. In application, this concern can be eliminated by using the bootstrap variance estimator of $\hat{\boldsymbol{\beta}}$ if the censoring proportion is very small.

When the right-censoring percentage is high (50%), the estimator of $\boldsymbol{\beta}$ from EE-I can be biased due to the instability of weight function W_1 in the denominator. This phenomenon remains when the total sample size increases from 200 to 400. In contrast, the estimators from EE-II are much more robust to various censoring percentages, its biases are small, and its coverage probabilities are close to the nominal level, especially when the sample size increases. The relative ratios for the variance estimators between the estimators from EE-II and from EE-I show a loss of efficiency between 1 and 44% for EE-I. The largest loss of efficiency for EE-I relative to EE-II occurs when there is heavy censoring.

As expected, the estimators obtained from EE-LT have larger variances than the estimators from EE-I and EE-II, especially with small to moderate censoring, because EE-LT ignores part of the information for $\boldsymbol{\beta}$ contained in the residual likelihood. The relative ratios for the variance estimators between the estimators from EE-LT and EE-II show a loss of efficiency of up to 42%. However, EE-LT has the advantage of not requiring an estimate of the censoring distribution of C, which leads to a more robust estimation procedure for different censoring distributions.

As suggested by a referee, we also performed a small simulation study to assess the bias in the proposed estimators when the stationarity assumption is violated. Because the proposed estimating equation approaches are derived for length-biased data, biases are expected in the estimators obtained from EE-I and EE-II when the data do not satisfy the stationarity assumption (shown in Table 2). For both EE-I and EE-II, the bias and the mean square error increase with the percent of censoring. In contrast, the less efficient EE-LT approach proposed for delayed-entry/left-truncated data does not rely on the stationarity assumption; therefore, it leads to unbiased estimators.

3.2 Example: Dementia Study

The Canadian Study of Health and Aging was a multicenter epidemiologic study that has been described in the literature (Wolfson et al., 2001; Asgharian et al., 2002). In the first phase of the study, a total of 14,026 subjects who were 65 years or older were randomly selected from throughout Canada to receive an invitation to participate in a health survey. A total of 10,263 Canadians agreed to participate. The participants were screened for dementia in 1991. From that cohort, 1,132 participants were identified as having dementia. The dates of disease onset were ascertained from the participants medical records, and their dates of death or right censoring were collected prospectively during the second phase until the end of the study in 1996. After excluding participants for whom the date of disease onset or the classification of dementia subtype was missing, there were 818 participants left. Their dementias were classified into the following three categories: probable Alzheimer's disease, n=393; possible Alzheimer's disease, n=252; and vascular dementia, n=173. At the end of the study, 638 participants had died, and the others were right censored. The purpose of this study was to assess whether the subtype of dementia at diagnosis had any effect on overall survival.

We first checked the correlation between the type of dementia and the censoring distribution C and found no statistically significant association. The stationarity assumption for the length-biased data was carefully validated by Asgharian, Wolfson, and Zhang (2006). To perform a semiparametric regression analysis, we used the category of probable Alzheimer's disease as the baseline and defined two indicator variables for possible Alzheimer's disease and vascular dementia under the Cox model. The estimated covariate coefficients for the three methods that adjust for length-biased sampling and the naive analysis that does not adjust for length-biased sampling are summarized in Table 3. The estimated standard errors for the covariate coefficients are obtained from the proposed consistent estimators for EE-I, EE-II. It is clear from the results that the subtype of dementia yields little difference in long-term survival under the methods that adjust for length-biased sampling. This finding is consistent with the nonparametric survival estimators provided by Wolfson et al. (2001). Note that the overestimated coefficient obtained from the naive Cox model suggests a marginally significant prolonged survival for the category of possible Alzheimers disease compared with that for the category of probable Alzheimers disease.

3.3 Computation Algorithms

Estimating covariate effects under the Cox model for right-censored failure time data is easy for the end user with either S-PLUS (R) or SAS, via the function “coxph.” Our aim in this section is to illustrate how to use existing software for traditional right-censored data to analyze length-biased right-censored data under Coxs proportional hazards model. Specifically, we describe slightly modified commands in S-PLUS (R) for the two proposed estimating equations, EE-I and EE-II. For length-biased data, we can use the function “coxph,” but with the “offset” option to add a linear predictor in the Cox model with a known coefficient of one for the weight. For the estimating equation EE-I or EE-II, we use the estimated weight for W_{1i} or W_{2i}, respectively, as the input for “offset” in “coxph,”. To illustrate, using the previously described example, we define vectors $\boldsymbol{X}_V$ and $\boldsymbol{X}_P$ as indicators of Vascular dementia and possible Alzheimer's disease, and then apply

$$>\text{coxph}\left(\text{Surv}(\text{futime, rep}\,(1,\text{m}))\tilde{}\boldsymbol{X}_V+\boldsymbol{X}_P+\text{offset}\left(\log\left(\widehat{W}_1\right)\right)\right),\text{data=fdata}\Big),$$

where “futime” is the observed failure times, m is the total number of the observed failure times, “fdata” is the subset of the whole data matrix among subjects with observed failure times only, and $\hat{\boldsymbol{W}}_1$ is a vector consisting of the estimated weight of each failure time for EE-I. Note that $\hat{\boldsymbol{W}}_1$ is estimated by the Kaplan-Meier estimator of censoring variable C on all $Y_i - A_i$ and $i = 1, \cdots, m$,

$$\widehat{W}_{1i}=\left\{Y_i\widehat{S}_C\,(Y_i-A_i)\right\}^{-1}.$$

Similarly, for EE-II we can use $\hat{W}_{2i} = \{\hat{w}_c(Y_i)\}^{-1}$ to give the input for the “offset” term in “coxph.”

The Newton-Raphson method is used to solve the partial likelihood equation for estimating $\boldsymbol{\beta}$ for conventional right-censored data in both S-PLUS and SAS. For the purpose of comparison, we also used the Newton-Raphson iterative algorithm to solve $\boldsymbol{\beta}$ from EE-I and EE-II. It is not surprising that the numerical solution obtained from the Newton-Raphson method for EE-I (or EE-II) is the same as the one obtained from the command “coxph” with the “offset” option using $\log(\hat{\boldsymbol{W}}_1)$ (or $\log(\hat{\boldsymbol{W}}_2)$), because EE-I (or EE-II) is identical to the estimating equation of the ordinary Cox model treating $\hat{\boldsymbol{W}}_1$ (or $\hat{\boldsymbol{W}}_2$) as a fixed “offset” term and restricting only the failure times. Note that EE-I (8) can also be expressed as

$$\tilde{U}_1(\beta) = \sum_{i=1}^{n} \delta_i \left[\boldsymbol{X}_i - \frac{\Sigma_{j=1}^{n} I\left(Y_j \geqslant Y_i\right) \delta_j \boldsymbol{X}_j \exp\left\{\beta^T \boldsymbol{X}_j + \log\left(W_{1j}\right)\right\}}{\Sigma_{j=1}^{n} I\left(Y_j \geqslant Y_i\right) \delta_j \exp\left\{\beta^T \boldsymbol{X}_j + \log\left(W_{1j}\right)\right\}} \right] = 0,$$

which is the same as the score equation used in the ordinary Cox model with a linear predictor log($\boldsymbol{W}_1$) restricting among the observed failure times. The censoring distribution for C enters into the estimating equations (EE-I or EE-II) only through the estimated weights $\hat{\boldsymbol{W}}_1$ or $\hat{\boldsymbol{W}}_2$.

4. Discussion

The methodology for estimating covariate effects under a proportional hazards model for classical survival data has been implemented in standard software and is widely used in S-PLUS, SAS, and other statistical packages. An advantage for our proposed inference methods for length-biased right-censored data is that we can use the existing functions under the Cox model in S-PLUS or SAS to carry out the point estimation by providing only an estimated weight. Specifically, we only need to estimate weight W_{1i} from the Kaplan-Meier estimator for residual censoring times $S_C(t)$ and W_{2i}, which is an integral of $S_C(t)$. We can then use the “offset” option in S-PLUS (R) to incorporate the weights for the regular “coxph” function. The consistent variance-covariance estimator of $\hat{\boldsymbol{\beta}}_k$ can be obtained by $\widehat{\Gamma}_k \widehat{\Sigma}_k^{-1} \widehat{\Gamma}_k$ for k = , 2, as proposed in Section 2.3 for the estimating equations. Alternatively, one may use the bootstrap approach to obtain the corresponding standard errors using the existing functions “coxph” in S-PLUS (R) or “PROC PHREG” in SAS.

For the two proposed estimating equation approaches, the desired asymptotic properties were derived under mild regularity conditions. For both types of estimating equations, we proposed two estimators for the baseline hazards function for $\Lambda_0(t)$, which can lead to the prediction of covariate-specific survival. However, the establishment of the asymptotic properties of the Breslows estimators for the cumulative baseline hazards function is not a focus of this work. When the censoring distribution of C depends on covariate X, we may use the covariate-dependent weights $\hat{S}_C(t|\boldsymbol{X}_i)$, which may follow a semiparametric model, a parametric model, or a fully nonparametric Kaplan-Meier estimator for discretized covariates.

Of the three estimating equation approaches we investigated, we found that EE-II may be the most promising choice in general because it is robust to different censoring distributions and utilizes all of the available information. The estimators obtained from EE-I are comparable to those obtained from EE-II when the censoring percentages are small to moderate, but can be unstable when censoring is heavy. This is because the impact of an inverse of $S_C(t) \rightarrow 0$ can be large at the tail, whereas the integral of $S_C(t)$ will not go to zero at the tail. The estimating equation from the left-truncation model has the advantages of not requiring an estimate for the censoring distribution and working for general left-truncated data, which include length-biased data as a special case. However, the ignored component from the full likelihood causes a loss of information in the estimating equation EE-LT, which can lead to a loss of efficiency of up to 42% compared with that from EE-II in the scenarios that were investigated. Using the estimating equation by Ghosh (2008) may lead to a severely biased estimator for general right-censored length-biased data, while the method was proposed for length-biased data without follow-up data. In this case, the informative censoring mechanism cannot be accounted adequately. We hope that our results will facilitate applications of the most commonly used semiparametric regression model, Coxs proportional hazards model, in analyzing length-biased failure time data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank Professor David Zucker, the associate editor, and the referee for their constructive comments. This work was supported in part by grants CA79466 and CA016672 from the National Institutes of Health. We thank Professor Masoud Asgharian and the investigators from the Canadian Study of Health and Aging (CSHA) for providing us with the dementia data from the CSHA. The data reported in the example were collected as part of the CSHA. The CSHA study was funded by the Seniors' Independence Research Program through the National Health Research and Development Program (NHRDP) of Health Canada (Project no.6606-3954-MC(S)). Additional funding was provided by Pfizer Canada Incorporated through the Medical Research Council/Pharmaceutical Manufacturers Association of Canada Health Activity Program, NHRDP Project 6603-1417-302(R), Bayer Incorporated, and the British Columbia Health Research Foundation Projects 38 (93-2) and 34 (96-1). The study was coordinated through the University of Ottawa and the Division of Aging and Seniors, Health Canada.

APPENDIX

The regularity conditions for the large sample properties are:

a. $(A_i, Y_i, \delta_i, \boldsymbol{X}_i)$ $(i = 1, \cdots, n)$ are independent and identically distributed;

b. $P(C_i + A_i \geqslant \tau) > 0$, where τ is a predetermined constant;

c. $\boldsymbol{\Gamma}_k = E\left[\int_0^\tau \{\boldsymbol{X} - \boldsymbol{e}_k(\beta_0, t)\}^{\otimes 2} w_c(t) R_1(t) W_{k1}^{-1} \exp\left(\beta_0^T \boldsymbol{X}\right) d\Lambda_0(t)\right]$ is positive definite for $k = 1$ or 2, where $\boldsymbol{\beta}_0$ and $\Lambda_0(t)$ are the true underlying values of $\boldsymbol{\beta}$ and the baseline hazard function;

d. $0 < w_c(\tau) < \infty$ and $\int_0^\tau \left[\left\{\int_t^\tau S_c(u)\, du\right\}^2 / \left\{S_c^2(t) S_v(t)\right\}\right] dS_c(t) < \infty$, where $S_V(t) = \mathrm{pr}(Y - A > t)$.

Consistency and uniqueness of $\hat{\beta}_1$

Consider a likelihood function

$$L_1(\beta) = \frac{1}{n} \sum_{i=1}^n \int_0^\tau \left[(\beta - \beta_0)^T \boldsymbol{X}_i - \log\left\{\frac{\tilde{S}_1^{(0)}(\beta, t)}{\tilde{S}_1^{(0)}(\beta_0, t)}\right\}\right] dN_i(t).$$

Note that $\partial L_\mathrm{i}(\boldsymbol{\beta})/\partial \boldsymbol{\beta} = \mathrm{n}^{-1} \tilde{\boldsymbol{U}}_1(\boldsymbol{\beta})$ and

$$-\frac{\partial^2 L(\beta)}{\partial \beta^2} = \widehat{\boldsymbol{\Gamma}}_1(\beta) = n^{-1} \sum_{i=1}^n \int_0^\tau \left\{\boldsymbol{X}_i - \tilde{E}_1(\beta, t)\right\}^{\otimes 2} \widehat{w}_c(t) R_i(t) \widehat{W}_{1i} \exp\left(\beta^T \boldsymbol{X}_i\right) d\widehat{\Lambda}_{10}(\beta, t).$$

By the strong law of large numbers and under the specified regularity conditions, $L_1(\boldsymbol{\beta})$ converges almost surely to

$$E\left[\int_0^\tau \left[(\beta - \beta_0)^T \boldsymbol{X}_1 - \log\left\{\frac{s_1^{(0)}(\beta, t)}{s_1^{(0)}(\beta_0, t)}\right\}\right] dN_1(t)\right]$$

for any $\boldsymbol{\beta}$, and $\hat{\boldsymbol{\Gamma}}_1(\boldsymbol{\beta}_0)$ converges almost surely to $\boldsymbol{\Gamma}_1$, as n goes to infinity, where

$$\mathbf{\Gamma}_1 = E\left[\int_0^\tau \{\boldsymbol{X}_1 - \boldsymbol{e}_1(\beta_0, t)\}^{\otimes 2} w_c(t) R_1(t) W_{11}^{-1} \exp\left(\beta_0^T \boldsymbol{X}_1\right) d\Lambda_{10}(\beta_0, t)\right]$$

is assumed to be positive definite. Therefore, $L_1(\boldsymbol{\beta})$ is concave for $\boldsymbol{\beta}$, which leads to a unique solution to $\tilde{\boldsymbol{U}}_1(\boldsymbol{\beta})$. The consistency of $\hat{\boldsymbol{\beta}}_1$ also follows.

Weak Convergence of Ũ₁(β)

By decomposing $\tilde{\boldsymbol{U}}_1(\boldsymbol{\beta})$ approximately by the following two components,

$$\begin{aligned}\frac{1}{\sqrt{n}}\tilde{U}_1(\beta) = & \frac{1}{\sqrt{n}}\sum_{i=1}^n \int_0^\tau \{\boldsymbol{X}_i - \boldsymbol{E}_1(\beta, t))\} dN_i(t) + \frac{1}{\sqrt{n}}\sum_{i=1}^n \int_0^\tau \left\{\boldsymbol{E}_i(\beta, t) - \tilde{E}_1(\beta, t))\right\} dN_i(t) \\ = & \frac{1}{\sqrt{n}}\int_0^\tau \{\boldsymbol{X}_i - \boldsymbol{e}_1(\beta, t))\} dM_{1i}(t) + \frac{1}{\sqrt{n}}\sum_{i=1}^n \int_0^\tau \boldsymbol{a}(\beta, t)\frac{dM_{Ci}(t)}{\pi(t)} + o_p(1),\end{aligned} \tag{16}$$

where

$$\boldsymbol{a}(\beta, t) = -\lim_{n\to\infty}\frac{1}{n^2}\sum_{i=1}^n\sum_{k=1}^n \frac{I(Y_k - A_k \geqslant t)\delta_k\delta_i I(Y_k \geqslant Y_i) w_c(Y_i) W_{1k}\boldsymbol{X_k}\exp\left(\beta^T \boldsymbol{X_k}\right)}{S_1^{(0)}(\beta, Y_i)},$$

where $\widehat{M}_{Ci}(t) = I(C_i \leqslant t, \delta_i = 0) - \int_0^t I(C_i \geqslant u) d\Lambda_c(u)$, $\Lambda_C(t)$ is the hazard function for C, and $\pi(t) = S_C(t)S_V(t)$. The second term in (16) is derived using the fact that the Kaplan-Meier estimator can be approximated by a sum of martingale integrals,

$$\sqrt{n}\frac{\widehat{S}_c(t) - S_c(t)}{S_c(t)} = -\frac{1}{\sqrt{n}}\sum_{i=1}^n \int_0^t \frac{dM_{Ci}(s)}{\pi(s)} + o_p(1).$$

More details can be found in Section 1 of the Supplementary Materials. Under the regularity conditions (a)-(d), $n^{-1/2}\tilde{\boldsymbol{U}}_1(\boldsymbol{\beta})$ converges weakly to a Gaussian process with mean zero and variance-covariance matrix $\mathbf{\Sigma}_1$, where

$$\mathbf{\Sigma}_1 = E\left[\int_0^\tau \left\{\{\boldsymbol{X}_i - \boldsymbol{e}_i(\beta, t)\} dM_{1i}(t) + \frac{\boldsymbol{a}(\beta, t)}{\pi(t)} dM_{Ci}(t)\right\}\right]^{\otimes 2}.$$

Consistency and uniqueness of $\boldsymbol{\beta}_2$

Consider a likelihood function

$$L_2(\beta) = \frac{1}{n}\sum_{i=1}^n \int_0^\tau \left[(\beta - \beta_0)^T \boldsymbol{X}_i - \log\left\{\frac{\bar{S}_2^{(0)}(\beta, t)}{\bar{S}_2^{(0)}(\beta_0, t)}\right\}\right] dN_i(t).$$

The rest of the derivation is then similar to that for $\boldsymbol{\beta}_1$.

Weak Convergence of $\tilde{U}_2(\beta)$

Similar to the asymptotic representation for (8), the estimating equation (11) can be approximated by the following i.i.d. summation of a mean zero stochastic process, $\tilde{\boldsymbol{U}}_2(\boldsymbol{\beta})$ can be approximated by the sum of i.i.d mean zero process,

$$\begin{aligned}\frac{1}{\sqrt{n}}\tilde{U}_2(\beta) = & \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^\tau \{\boldsymbol{X}_i - \boldsymbol{E}_2(\beta,t))\}\, dN_i(t) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^\tau \left\{\boldsymbol{E}_2(\beta,t) - \tilde{E}_2(\beta,t))\right\} dN_i(t) \\ = & \frac{1}{\sqrt{n}}\int_0^\tau \{\boldsymbol{X}_i - \boldsymbol{e}_2(\beta,t))\}\, dM_{2i}(t) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^\tau \boldsymbol{H}(\beta,t)\,\frac{dM_{Ci}(t)}{\pi(t)} + o_p(1),\end{aligned} \tag{17}$$

where $M_{Ci}(t) = I(V_i \leqslant t, \delta_i = 0) - \int_0^t I(V_i \geqslant u)\, d\Lambda_c(u)$,

$$\boldsymbol{H}(\beta,t) = \lim_{n\to\infty}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{k=1}^{n}\frac{w_c(Y_i)\,R_k(Y_i)\,\boldsymbol{X_k}\exp\left(\beta^T\boldsymbol{X_k}\right)w_c^{-2}(Y_k)\,h_k(t)}{S_2^{(0)}(\beta,Y_i)},$$

and $h_k(t) = I(t \leqslant Y_k)\int_t^{Y_k} S_c(u)\,du$. The second term in (17) explains the uncertainty induced by $\hat{w}_c(t)$. The last equation holds because all $w_c(t)$ and $\hat{w}_c(t)$ are canceled out inside $\boldsymbol{E}_2$ and $\hat{\boldsymbol{E}}_2$, and $(\hat{w}_c(Y_k) - w_c(Y_k))$ can be expressed as an i.i.d. sum of martingales (Pepe & Fleming, 1991),

$$n^{1/2}(\widehat{w}_c(Y_k) - w_c(Y_k)) = n^{-1/2}\sum_{j=1}^{n}\int_0^\tau h_k(t)\,\frac{dM_{Cj}(t)}{\pi(t)}.$$

More details can be found in Section 2 of the Supplementary Materials.

References

Andersen, PK.; Borgan, O.; Gill, RD.; Keiding, N. Statistical models based on counting processes. Springer-Verlag Inc.; 1993.

Asgharian M, M'Lan CE, Wolfson DB. Length-biased sampling with right censoring: an unconditional approach. J. Am. Statist. Assoc. 2002; 97:201–209.

Asgharian M, Wolfson DB. Asymptotic behavior of the unconditional NPMLE of the length-biased survivor function from right censored prevalent cohort data. Ann. Statist. 2005; 33:2109–2131.

Asgharian M, Wolfson DB, Zhang X. Checking stationarity of the incidence rate using prevalent cohort survival data. Stat. Med. 2006; 25:1751–1767. [PubMed: 16220462]

Bergeron P-J, Asgharian M, Wolfson DB. Covariate bias induced by length-biased sampling of failure times. J. Am. Statist. Assoc. 2008; 103:737–742.

Cox DR. Regression models and life-tables (with discussion). J. R. Statist. Soc. B. 1972; 34:187–220.

de Una-Alvarez J, Otero-Giraldez MS, Alvarez-Llorente G. Estimation under length-bias and right-censoring: an application to unemployment duration analysis for married women. J. Applied Statist. 2003; 30:283–291.

Gill RD, Vardi Y, Wellner JA. Large-sample theory of empirical distributions in biased sampling models. Ann. Statist. 1988; 16:1069–1112.

Ghosh D. Proportional hazards regression for cancer studies. Biometrics. 2008; 64:141–148. [PubMed: 17573863]

Kalbfleisch JD, Lawless JF. Regression models for right truncated data with applications to AIDS incubation times and reporting lags. Statistica Sinica. 1991; 1:19–32.

Keiding, N. Independent delayed entry. In: Klein, JP.; Goel, PK., editors. Survival Analysis: State of the Art. Kluwer Academic Publishers Group; Norwell, Massachusetts: 1992. p. 309-325.

Lagakos SW, Barraj LM, De Gruttola V. Nonparametric analysis of truncated survival data, with applications to AIDS. Biometrika. 1988; 75:515–523.

Lancaster T. Econometric methods for the duration of unemployment. Econometrica. 1979; 47:939–956.

Pepe MS, Fleming TR. Weighted Kaplan -Meier statistics: large sample and optimality considerations. J. Roy. Statist. Soc. B. 1991; 53:341–352.

Shen Y, Ning J, Qin J. Analyzing length-biased data with semiparametric transformation and accelerated failure time models. J. Am. Statist. Assoc. 2009 in press.

Turnbull BW. The empirical distribution function with arbitrarily grouped, censored and truncated data. J. R. Statist. Soc. B. 1976; 38:290–295.

Vardi Y. Nonparametric estimation in the presence of length bias. Ann. Statist. 1982; 10:616–620.

Vardi Y. Empirical distributions in selection bias models. Ann. Statist. 1985; 13:178–203.

Vardi Y. Multiplicative censoring, renewal processes, deconvolution and decreasing density: Nonparametric estimation. Biometrika. 1989; 76:751–761.

Wang MC. A semiparametric model for randomly truncated data. J. Am. Statist. Assoc. 1989; 84:742–748.

Wang MC. Hazards regression analysis for length-biased data. Biometrika. 1996; 83:343–354.

Wang MC, Brookmeyer R, Jewell NP. Statistical models for prevalent cohort data. Biometrics. 1993; 49:1–11. [PubMed: 8513095]

Wolfson C, Wolfson DB, Asgharian M, M'Lan CE, Ostbye T, Rockwood K, Hogan DB, For the Clinical Progression of Dementia Study Group. A reevaluation of the duration of survival after the onset of dementia. New Engl. J. Med. 2001; 344:1111–1116. [PubMed: 11297701]

Zelen M. Forward and backward recurrence times and length biased sampling: age specific models. Lifetime Data Anal. 2004; 10:325–334. [PubMed: 15690988]

Zelen M, Feinleib M. On the theory of screening for chronic diseases. Biometrika. 1969; 56:601–614.

Table 1

Empirical estimators and coverage probabilities of the 95% confidence interval under three estimating equations; C%: censoring %; EE: estimating equation; 95% CP: 95% coverage probability; RE: ratios of empirical variances (e.g for $\hat{\alpha}_1$, ratio of variance estimators from EE-II and EE-I: $\widehat{\sigma}_2^2/\widehat{\sigma}_1^2$)

(α_1, α_2)	C%	EE	$(\hat{\alpha}_1, \hat{\alpha}2)$		95% CP		RE		$(\hat{\alpha}_1, \hat{\alpha}2)$		95% CP		RE	
					***n* = 200**						***n* = 400**			
(1,1)	20	I	0.991	0.972	.974	.959	.89	.89	0.989	0.970	.978	.956	.92	.81
		II	1.020	1.016	.985	.970	1.0	1.0	1.010	1.010	.984	.978	1.0	1.0
		LT	1.008	1.015	.934	.948	.56	.67	1.003	1.004	.953	.946	.58	.65
		III	0.346	0.362	.078	.381			0.345	0.349	.001	.071		
	35	I	0.936	0.920	.930	.912	.79	.78	0.934	0.901	.920	.895	.73	.59
		II	1.024	1.026	.974	.958	1.0	1.0	1.011	0.999	.973	.973	1.0	1.0
		LT	1.008	1.013	.948	.952	.72	.85	1.006	1.000	.961	.951	.76	.71
		III	0.293	0.315	.068	.425			0.297	0.296	.000	.099		
	50	I	0.838	0.813	.841	.866	.70	.71	0.841	0.835	.762	.838	.56	.61
		II	1.019	1.003	.953	.944	1.0	1.0	1.016	1.007	.958	.956	1.0	1.0
		LT	1.003	1.021	.961	.945	.92	.90	1.002	1.001	.950	.945	.96	.77
		III	0.201	0.230	.072	.423			0.218	0.240	.000	.136		
(2,2)	20	I	1.989	1.951	.968	.961	1.0	.90	2.017	1.960	.976	.962	1.0	.95
		II	2.013	2.001	.972	.971	1.0	1.0	2.044	2.018	.974	.968	1.0	1.0
		LT	2.024	2.031	.958	.956	.68	.69	2.055	2.029	.951	.954	.67	.72
		III	0.585	0.691	.005	.016			0.593	0.672	.007	.008		
	35	I	1.961	1.839	.963	.920	.99	.82	1.938	1.810	.953	.895	.94	.74
		II	2.042	2.018	.971	.962	1.0	1.0	2.019	1.980	.972	.963	1.0	1.0
		LT	2.058	2.052	.945	.947	.63	.79	2.022	1.998	.950	.966	.63	.79
		III	0.539	0.569	.006	.009			0.538	0.556	.003	.008		
	50	I	1.833	1.674	.894	.837	.90	.75	1.858	1.664	.892	.826	.90	.73
		II	2.016	1.992	.962	.955	1.0	1.0	2.045	1.987	.956	.961	1.0	1.0
		LT	2.039	2.015	.948	.951	.71	.81	2.055	2.005	.941	.934	.70	.75
		III	0.435	0.422	.006	.018			0.434	0.392	.006	.017		

Table 2

Simulation results without stationarity assumption with sample size 200

(α_1, α_2)	**C%**	**EE**	**($\hat{\alpha}_1, \hat{\alpha}_2$)**		**95% CP**		**MSE**	
(1,1)	20	I	0.946	0.898	0.952	0.942	1.170	1.367
		II	0.938	0.882	0.965	0.948	1.184	1.388
		LT	1.016	1.013	0.948	0.930	1.045	1.163
	35	I	0.894	0.839	0.884	0.869	1.305	1.587
		II	0.920	0.853	0.934	0.932	1.238	1.505
		LT	1.016	1.029	0.945	0.939	1.056	1.167
	50	I	0.813	0.766	0.805	0.844	1.532	1.834
		II	0.917	0.853	0.901	0.926	1.272	1.565
		LT	1.009	1.002	0.943	0.942	1.090	1.279

Table 3

Estimates (standard errors) of regression coefficients for dementia data under Cox model, using “probable Alzheimers disease” as baseline

	Length-bias adjusted analyses			**Naive Cox model**
	EE-I	**EE-II**	**EE-LT**	
Vascular dementia	0.137 (.101)	0.074 (.101)	0.087 (.103)	0.076 (.103)
Possible Alzheimers disease	−0.109 (.093)	−0.134 (.091)	−0.037 (.093)	−0.182 (.093)