Published in final edited form as:

Lifetime Data Anal. 2017 January ; 23(1): 3–24. doi:10.1007/s10985-016-9367-y.

Nonparametric and Semiparametric Regression Estimation for Length-biased Survival Data

Yu Shen,

Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas, 77030, USA

Jing Ning, and

Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas, 77030, USA

Jing Qin

Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, National Institute of Health, Bethesda, Maryland, 20892, USA

Abstract

For the past several decades, nonparametric and semiparametric modeling for conventional right-censored survival data has been investigated intensively under a noninformative censoring mechanism. However, these methods may not be applicable for analyzing right-censored survival data that arise from prevalent cohorts when the failure times are subject to length-biased sampling. This review article is intended to provide a summary of some newly developed methods as well as established methods for analyzing length-biased data.

Keywords

Length-biased sampling; Estimating equation; Left truncation; Likelihood; Nonparametric estimator; Semiparametric models

1 Introduction

An outcome-dependent sampling bias arises when observations are not randomly selected from the target population (Cox 1962, Chapter 5). Length-biased sampling represents a special case of left-truncated data, and has been recognized in various applications: biomedical research (Keiding et al 2002; Simon 1980; Zelen 2004), marketing research (Nowell and Stanley 1991), genome-wide linkage studies (Terwilliger et al 1997), labor economics (de Una-Álvarez et al 2003), and nanotechnology (Kvam 2008). Length bias is difficult to remove through the trial design and may confound the interpretation of disease-specific survival times.

The observed data are often from a cross-sectional cohort of patients diagnosed with a particular disease at the time of examination, who are then followed for the occurrence of a subsequent disease-related event (e.g. disease-specific death). Under this sampling design, patients with shorter survival times are selectively excluded, while those with longer survival

times are more likely to be included in the cohort. One example of such data is a study of dementia and the subsequent onset of death. In 1991, the Canadian Study of Health and Aging (CSHA) was initiated to estimate the prevalence of dementia and to understand the natural history of the disease in elderly Canadians (Wolfson et al 2001).

Using a prevalent cohort design, a total of 14,026 adults who were 65 years or older were randomly selected and recruited throughout Canada in the first phase of the CSHA. Among them, 10,263 agreed to participate in this multicenter epidemiological study. Focusing on the prevalent patients, 1132 were identified as having dementia at enrollment and were followed prospectively until the end of 1996. At enrollment, each caregiver provided an approximate date of the diagnosis of dementia for these patients (Asgharian et al 2014). While the majority of the patients with dementia died at the end of follow-up, a small proportion of them had been lost to follow-up, so that their survival times were right censored. Two major scientific objectives were to estimate the survival distribution of these patients measured from their diagnosis, and to determine how different types of dementia impact long-term survival. In addition, the study collected other baseline covariates such as gender, age at study enrollment and level of education.

As noted by investigators of the CSHA and in the literature, individuals with dementia in the CSHA had to survive long enough to be sequentially recruited into the study during 1991. In other words, the patients who had shorter survival times when measured from date of diagnosis were less likely to be recruited to the cohort. Therefore, the observed data are not a random sample of the target population, which was all elderly individuals with dementia who resided in Canada. This bias is common in cross-sectional prevalent cohort studies; the survival times from such cohort studies are subject to left truncation. Here, the truncation time is the duration from the diagnosis of dementia until enrollment in the study. In some applications, including the CSHA for elderly people, the incidence of disease onset follows a Poisson process, i.e. the disease incidence is constant over time (stable disease model), and the left-truncation time is uniformly distributed. Under this special condition, the probability of a survival time being sampled is proportional to its length; therefore, the survival times are known as length-biased times in this case. Length-biased sampling and the need to correct such bias in applications have been well recognized in epidemiology, marketing survey, environmental and labor economics studies (Kalbfleisch and Lawless 1989; Keiding et al 2002; Kvam 2008; Nowell and Stanley 1991; Simon 1980; de Una-Álvarez et al 2003; Zelen 2004). The assumption of a uniform truncation distribution (i.e. length-biased sampling) can be examined by formal goodness-of-fit tests (Addona and Wolfson 2006; Mandel and Betensky 2007; Martin and Betensky 2005). For the CSHA, Asgharian et al (2006) validated the stationarity assumption for the observed data, which are thus length-biased data.

In this article, we review the current state of methodologic development for statistical estimation and inference on nonparametric estimation and semiparametric modeling for length-biased, right-censored data.

2 Nonparametric Methods

2.1 Basic Notations and Definitions

Consider a cross-sectional sampling of individuals with a given disease from the population and then following those individuals prospectively until they experience a failure event, which may be subject to right censoring. As depicted in the diagram of Figure 1, the observable data are the time A from disease onset to enrollment, the time V from enrollment to death, and the censoring time C from enrollment to loss to follow-up. Let $\delta = I(V < C)$ be the censoring indicator and assume that (A, V) is independent of C. Let $Y = \min(A + V, A + C)$. Denote the observed data as (Y_i, A_i, δ_i), $i = 1, 2, ..., n$.

With length-biased sampling, the population survival time $\tilde{T}$, measured from disease diagnosis to death, can be observed only among those with $\tilde{T} > \tilde{A}$. The observable survival time is length-biased and equal to $T = A + V$. In contrast to conventional right-censored data, another complication in analyzing such data is the potential dependence between the failure time T and the right censoring time $A + C$, measured from the initiating event (diagnosis) to the event of failure.

The density function, and survival function for (unbiased) failure time $\tilde{T}$ are denoted by $f(t) = dF(t)/dt$ and $S(t)$, respectively. The density function of the observed biased T is defined as $g(y) = dG(y)/dy$, where $dG(y) = ydF(y)/\mu$, and the survival function of $\tilde{T}$ is $\mu = \int tf(t)dt$. Based on the renewal theory under the stable disease model (Vardi 1982, 1989), the joint distribution of (A, V) is

$$\frac{f(a+v)}{\mu}, \quad a, v>0.$$

Equivalently, the joint density of (A, T) can be expressed as

$$f_{A,T}(a,t)=f_A(a)f_{T|A}(t|a)=\frac{S(a)}{\mu}\frac{f(t)}{S(a)} \quad (1)$$

where

$$S(t)=\int_t^\infty dF(u)=\mu\int_t^\infty u^{-1}dG(u). \quad (2)$$

2.2 Conditional Likelihood Methods

Several articles published over the past three decades have described statistical methodology development on the estimation of nonparametric survival distributions for left-truncated data, and have covered length-biased data as a special case (Keiding 1992; Wang 1991; Wang et al 1993). The large sample properties of such nonparametric survival function

estimators have been well established (Gross and Lai 1996; Keiding 1992; Keiding and Gill 1990; Lagakos et al 1988; Turnbull 1976; Wang 1991; Wang et al 1986).

For length-biased data subject to right censoring, the full likelihood function for the observed data can be derived from (1) and is proportional to

$$L_F(S)=\prod_{i=1}^{n}\frac{S(A_i)}{\mu}\prod_{i=1}^{n}\frac{f(Y_i)^{\delta_i}\,S(Y_i)^{1-\delta_i}}{S(A_i)}. \quad (3)$$

Given truncation time $A = a$, the conditional likelihood of (Y, δ) is the second component of (3),

$$L_C(S)=\prod_{i=1}^{n}\frac{f(Y_i)^{\delta_i}\,S(Y_i)^{1-\delta_i}}{S(A_i)}. \quad (4)$$

As described in detail by Woodroofe (1985) and Wang (1991), a nonparametric product-limit estimator for $S(t)$ can be constructed on the basis of L_C. Letting $t_{(j)}$ denote the distinct ordered failure times from uncensored Y_i, the derived nonparametric maximum likelihood estimator (NPMLE) is similar to the Kaplan-Meier estimator after replacing the risk set $R(t) = \{i : Y_i \geq t\}$ with $R_T(t) = \{i : A_i \leq t \leq Y_i\}$,

$$\hat{S}(t)=\prod_{t_{(j)}<t}\left[1-\frac{\sum_{i=1}^{n}I\{Y_i=t_{(j)},\delta_i=1\}}{\sum_{i=1}^{n}I\{(A_i,Y_i)\in R_T(t_{(j)})\}}\right].$$

Considering length-biased data to be a special case of left-truncated data, the nonparametric survival function of $S(.)$ can be estimated using this approach. (The nonparametric estimator is also referred to as the truncation product-limit estimator.) Without specifying the distribution of the random truncation time, the above methods are efficient, conditional on the observed truncation times.

2.3 Full Likelihood Methods

It is clear that the aforementioned conditional likelihood approach for length-biased data is not fully efficient, since it does not use the likelihood contribution due to A. Under length-biased sampling, the distribution of $\tilde{A}$ follows a uniform distribution.

Vardi proposed the NPMLE from full likelihood function approaches under the stationarity assumption (Gill et al 1988; Vardi 1982, 1989). More recently, Qin et al. (2011) developed an alternative nonparametric full likelihood method. We describe the methods next.

2.3.1 EM Algorithm to Estimate the Cumulative Distribution Function of the Biased Failure Time T—

With a multiplicative censoring model, Vardi considered that

uncensored failure times $\{R_1, \cdots, R_{n_1}\}$ are independently generated from the cumulative distribution function (CDF) G, and censored times $\{Z_1, \cdots Z_{n_2}\}$ as "incomplete observations" (or "contaminated observations") and come from a product of an independent uniform (0,1) random variable U and R, $Z = U * R$. Here, R is also generated from the CDF G. Under this data generation model, the density function of Z is

$$h(z)=\int_z^\infty y^{-1}dG(y), z>0.$$

The nonparametric likelihood function given the observed data $\{R_1, \cdots, R_{n_1}\}$, $\{Z_1, \cdots Z_{n_2}\}$ is

$$L_V(G)=\prod_{j=1}^{n_1}[dG(R_j)]\prod_{i=1}^{n_2}\left[\int_{Z_i}^\infty y^{-1}dG(y)\right]. \quad (5)$$

Using the notation introduced earlier for left-truncated data (also length-biased data), the full likelihood for the observed data (Y_i, δ_i), $i = 1, \cdots, n = n_1 + n_2$ can be equivalently expressed as

$$L_V(G)=\prod_{i=1}^{n}[dG(Y_i)]^{\delta_i}\left[\int_{y\geq Y_i} y^{-1}dG(y)\right]^{1-\delta_i}. \quad (6)$$

Vardi (1989) proposed an expectation-maximization (EM) algorithm for the NPMLE of G based on the above full likelihood function. If $dG(.)$ is replaced by $dF(.)$ in (6), then the likelihood expression (3) is equivalent to (6), in which the likelihood contributions from the uniformly "contaminated observation" Z and "uncontaminated" observation R correspond to the contributions from right censored and uncensored failure time data, respectively. A subtle difference between the likelihood expressions (3) and (6) is that the numbers of censored and uncensored failure times (n_1 and n_2) are random in (3), but are fixed in (6). However, maximizing the two likelihood functions remains the same.

It is sufficient to consider the discrete version of distribution G on the point masses at $t_1 < t_2 <.... < t_k$, where $t_1, ..., t_k$ are the unique *failure and censoring times* for $\{Y_1, ..., Y_n\}$, $k \leq n$. Maximizing (6) is reduced to the problem of maximizing the discrete version of the CDF of T, $p_j = G(dt_j)$, subject to the constraints $p_j \geq 0$ and $\sum_{j=1}^k p_j = 1$, $j = 1, \cdots, k$. For this EM algorithm, $\{R_1, \cdots, R_{n_1}\}$, $\{\tilde{R}_1, \cdots \tilde{R}_{n_2}\}$ are considered as the "complete data" and $\{R_1, \cdots, R_{n_1}, Z_1, \cdots Z_{n_2}\} = \{(Y_i, \delta_i), i = 1, \cdots, n\}$ as the "incomplete data". The iterative EM algorithm to solve p_j follows.

Step 1. Select an arbitrary $p_j^{(0)}$ that satisfies $\sum_{j=1}^k p_j^{(0)} = 1$, $p_j^{(0)} \geq 0$.

Step 2. Solve $p_j^{(1)}$ by maximizing (5), so that we replace $p_j^{(0)}$ with

$$\hat{p}_j^{(1)}=\frac{1}{n}\sum_{i=1}^{n}\left[\delta_i I(Y_i=t_j)+(1-\delta_i)\frac{\hat{p}_j^{(0)}t_j^{-1}I(Y_i\geq t_j)}{\sum_{l=i}^{k}\hat{p}_l^{(0)}t_l^{-1}}\right]. \quad (7)$$

Given a convergence criterion, p_j can be solved iteratively and the maximum likelihood estimation (MLE) of p_j is denoted by $\hat{p}_j$. The NPMLE is consistently estimated by $\hat{G}(t)=\sum_{j=1}^{k}\hat{p}_j I(t_j>t)$. Asgharian et al (2002, 2005) provided the asymptotic properties of the NPMLE for the unbiased survival function in the presence of right censoring.

Using the fundamental relationship between G and S in (2), the NPMLE for S can be derived as follows,

$$\hat{S}(t)=1-\frac{\int_t^{\infty}u^{-1}d\hat{G}(u)}{\int_0^{\infty}u^{-1}d\hat{G}(u)},\ \text{since}\ dF(t)=t^{-1}dG(t)/\int t^{-1}dG(t).$$

Note that the NPMLE of S is estimated via the CDF of G for observed bias sample T. It is thus difficult to impose constraints on F (i.e., S) because they may not be easily translated to the constraints on G.

2.3.2 EM Algorithm to Estimate the CDF of Unbiased Failure Time $\tilde{T}$—In contrast to Vardi's method for estimating the NPMLE of G, Qin and colleagues proposed a different EM algorithm by directly estimating the NPMLE of F for the unbiased failure time $\tilde{T}$ in the target population (Qin et al 2011). They considered the 'incomplete' data from a different perspective for length-biased data. Specifically, they defined the observed biased sample on n subjects, denoted by $\boldsymbol{O}=\{(Y_1, \delta_1, A_1), \cdots, (Y_n, \delta_n, A_n), A_i \leq Y_i, i=1, \cdots, n\}$, as incomplete data due to left truncation; whereas the data on m subjects are left truncated. Here, the left-truncated data are denoted by $\boldsymbol{O}^*=\{(T_1^*, A_1^*), \ldots, (T_m^*, A_m^*), A_i^*>T_i^*, i=1, 2, \ldots, m\}$, where m follows a negative binomial distribution with parameter π. In essence, the length-biased observations (A, T) can be considered to be generated from a model with

$$\tilde{A}\sim U(0,\hat{\tau}),\ \ \tilde{T}\sim F,\ \text{on}\ (0,\hat{\tau}), \quad (8)$$

where $\hat{\tau}=t_k$, $\tilde{A}$ and $\tilde{T}$ are independent, and (A, T) is observed if and only if $\tilde{T}\geq\tilde{A}$. Similar to Vardi's setting, it is sufficient to consider the discrete version of F on the point masses, t_1, $\cdots$, t_k, and define $dF(t_i)=q_i$ and $\sum_{i=1}^{k}q_i=1$. The probability of having a length-biased observation under this setting is $\pi=P(\tilde{T}\geq\tilde{A})=E(\tilde{T})/\hat{\tau}$.

Thus, the "complete" data are defined as $\{\boldsymbol{O}, \boldsymbol{O}^*\}$. The log-likelihood based on the complete data is

$$\sum_{j=1}^{k}\left[\sum_{i=1}^{n}I(T_i=t_j)+\sum_{i=1}^{m}I(T_i^*=t_j)\right]\log q_j, \quad (9)$$

where $T_i \geq A_i$, $i = 1, 2, \cdots, n$ and $T_l^* < A_l^*$, $l = 1, \cdots, m$. Conditional on the observed data $\boldsymbol{O}$, one can derive the expectation of right-censored T as well as left-truncated T^*. Based on the conditional expectations, the following iterative EM algorithm is proposed to solve the MLE of $dF(t_j) = q_j$ for $j = 1, \cdots, k$.

Step 1. Select an arbitrary $q_j^{(0)}$ that satisfies $\sum_{j=1}^{k} q_j^{(0)}=1,\ q_j^{(0)} \geq 0$.

Step 2. Iteratively replace $q_j^{(0)}$ with

$$\hat{q}_j^{(1)}=\frac{\hat{\pi}^{(0)}}{n}\left\{\sum_{i=1}^{n}\left[\delta_i I(Y_i=t_j)+(1-\delta_i)\frac{\hat{q}_j^{(0)}I(Y_i\leq t_j)}{\sum_{j=1}^{k}\hat{q}_j^{(0)}I(Y_i\leq t_j)}\right]+\frac{n}{\hat{\pi}^0}(1-\frac{t_j}{\hat{\tau}})\hat{q}_j^{(0)}\right\}, \quad (10)$$

where $\hat{\pi}^{(0)}=\sum_{j=1}^{k} t_j\hat{q}_j^{(0)}/\hat{\tau}$.

Let $\hat{q}_j$ denote the MLE of q_j, $j = 1, \cdots, k$, the NPMLE $\hat{F}(t)=\sum_{j=1}^{k}\hat{q}_j I(t_j \leq t)$, and $\hat{p} = \int t d\hat{F}(t)/\hat{\tau}$. As a very different missing data mechanism is assumed here, a different EM-algorithm is proposed for estimating the NPMLE of F. It is not surprising that the derived NPMLE estimator $d\hat{G}(t) = td\hat{F}(t)/\hat{\mu}$, where $\hat{\mu} = \hat{\pi}\hat{\tau}$ is proved to be equivalent to the NPMLE $\hat{G}$ of Vardi. However, the EM algorithm by Qin et al. has the following advantages over Vardi's NPMLE: 1. It directly estimates the target distribution function F, so that constraints on F can be easily imposed; and 2. The EM algorithm based on the full likelihood (9) can be generalized to other semiparametric models.

3 Semiparametric Cox Model: Estimation Methods

The Cox proportional hazards model is the semiparametric regression model that is most commonly used in survival analyses (Cox 1972). It has been a main focus for analyzing length-biased data by developing valid approaches to modeling unbiased failure times with risk factors estimated under the Cox model. Assume that failure times in the target population, $\tilde{T}$, follow the proportional hazards model

$$\lambda(t|\boldsymbol{x})=\lambda_0(t)\exp(\boldsymbol{\beta}_0^T\boldsymbol{x}), \quad (11)$$

where $\lambda_0(t)$ is an unspecified baseline hazard function and $\boldsymbol{\beta}_0$ is a vector-valued unknown regression coefficient for $\boldsymbol{X}$. Note that the Cox model structure assumed for the target population is often different from the one for the observed length-biased data. Under length-

biased sampling, one can only observe T among those $\tilde{T} > \tilde{A}$, and $\boldsymbol{X}$ is the baseline covariate vector. It is reasonable to assume that C and (A, V) are independent given covariate $\boldsymbol{X}$, and that the censoring distribution is independent of covariate $\boldsymbol{X}$.

3.1 Conditional Approach for General Left-Truncated Data

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, similar to (4) for the case without covariates. Such a formulation has been utilized in analyzing left-truncated data (e.g. (Andersen et al 1993; Wang et al 1993)):

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

where $S(t|\boldsymbol{x})$ is the survival distribution for the unbiased failure time and $\mu(\boldsymbol{x}) = \int_0^\infty S(t|\boldsymbol{x})dx$ given $\boldsymbol{x}$. Given truncation time $A = a$, the conditional likelihood of Y is proportional to

$$L_C(\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{f(Y_i|\boldsymbol{X}_i,\boldsymbol{\beta})^{\delta_i} S(Y_i|\boldsymbol{X}_i,\boldsymbol{\beta})^{1-\delta_i}}{S(A_i|\boldsymbol{X}_i,\boldsymbol{\beta})}. \quad (12)$$

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:

$$L_C(\boldsymbol{\beta}, \Lambda_0) = L_P(\boldsymbol{\beta}) L_M(\boldsymbol{\beta}, \Lambda_0),$$

where

$$L_P(\boldsymbol{\beta}) = \prod_i \left[\frac{\exp(\boldsymbol{\beta}^T \boldsymbol{X}_i)}{\sum_{j\in R(Y_i)} \exp(\boldsymbol{\beta}^T \boldsymbol{X}_j)}\right]^{\delta_i}, \quad (13)$$

and the residual likelihood, L_M, is the marginal likelihood for truncation time A defined by

$$L_M(\boldsymbol{\beta}, \hat{\Lambda}_{\boldsymbol{\beta}}) = \prod_{i=1}^{n} \frac{S(A_i|\boldsymbol{X}_i)}{\mu_{\boldsymbol{\beta},\Lambda}(\boldsymbol{X}_i)} = \prod_{i=1}^{n} \frac{\exp\{-\Lambda(A_i)\exp(\boldsymbol{\beta}^T \boldsymbol{X}_i)\}}{\int_0^\infty \exp\{-\Lambda(u)\exp(\boldsymbol{\beta}^T \boldsymbol{X}_i)\}du}.$$

Note that the partial score equation from $L_P(\boldsymbol{\beta})$ is

$$U_P(\boldsymbol{\beta})=\sum_{i=1}^{n}\delta_i\left[\boldsymbol{X}_i-\frac{\sum_{j=1}^{n}\boldsymbol{X}_j\exp(\boldsymbol{\beta}^T\boldsymbol{X}_j)I(Y_j\geq Y_i\geq A_j)}{\sum_{j=1}^{n}\exp(\boldsymbol{\beta}^T\boldsymbol{X}_j)I(Y_j\geq Y_i\geq A_j)}\right]=0. \quad (14)$$

Wang et al (1993) proved that the partial likelihood estimator solved from $\boldsymbol{U}_P$ is approximately as efficient as the maximizer of L_C, when the distribution $\tilde{A}$ is unknown. Here, 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 $R(y) = \{j : A_j \leq y \leq Y_j\}$.

3.2 Estimating Equations with Inverse Weighting (or Weighted Risk Set)

3.2.1 Weighted Estimating Equation Methods for Uncensored Data—Under proportional hazards models, Wang (1996) was among the first to construct a pseudo-likelihood, which may be viewed as an inverse weighting estimating equation, to estimate the covariate effects on the unbiased failure outcomes without right censoring. Under length-biased sampling, Wang (1996) derived a score equation based on the pseudo-likelihood function, which can be expressed as the following estimating equation,

$$U_W(\boldsymbol{\beta})=\sum_{i=1}^{n}\left[\boldsymbol{X}_i-\frac{\sum_{j=1}^{n}\boldsymbol{X}_j\exp(\boldsymbol{\beta}^T\boldsymbol{X}_j)T_j^{-1}I(T_j\geq T_i)}{\sum_{j=1}^{n}\exp(\boldsymbol{\beta}^T\boldsymbol{X}_j)T_j^{-1}I(T_j\geq T_i)}\right]=0. \quad (15)$$

Wang (1996) and Tsai (2009) observed that (14) reduced to (15) asymptotically, when A_j is integrated out in (14) using the fact that A_j follows a uniform distribution of $U(0, T_j)$ given T_j. However, different methods are required when T_j is subject to right-censoring.

Ghosh (2008) proposed an estimating equation approach for modeling the regression coefficient of A and $\boldsymbol{X}$ to estimate the natural history of tumor growth under the Cox model in the context of length-biased sampling. While the method may serve the purpose of modeling the data of the tumor size without forward recurrence times (i.e. follow-up times), thus without right censoring, it could lead to a severely biased estimator for general right-censored, length-biased data.

3.2.2 Pseudo-Partial Estimating Equation—Tsai (2009) proposed the pseudo-partial likelihood approach by embedding the biased sampling data into left-truncated data using a missing data mechanism. Considering a different right-censoring schema, it was assumed that the truncation time and censoring time $(\tilde{A}, \tilde{C})$ have joint distribution function $G_{\tilde{A},\tilde{C}}(a, c)$, and $\tilde{T}$ and $(\tilde{A}, \tilde{C})$ are mutually independent. Here, $\tilde{C}$ is different from the previously defined residual censoring time C. Specifically, the failure, censoring and truncation times are not sampled from the joint distribution but from the conditional distribution, given $\{\tilde{T} \geq \tilde{A}, \tilde{C} \geq \tilde{A}\}$. By embedding biased-sampling data into the left-truncation model (13), the author used the log-partial likelihood derived by Cox (1972) for conventional right-censored data as the working likelihood, as follows,

$$\ell_\tau(\boldsymbol{\beta})=\sum_{i=1}^{n}\boldsymbol{\beta}^T\boldsymbol{X}_j-\sum_{i=1}^{n}E\left\{\log\sum_{j=1}^{n}\exp(\boldsymbol{\beta}^T\boldsymbol{X}_j)I(A_j\le T_i\le Y_j)|t_1,\cdots t_n,\boldsymbol{X}_1,\cdots,\boldsymbol{X}_n\right\}. \quad (16)$$

The second term of (16) is estimated by using the Monte Carlo method. For length-biased sampling with right censoring independently applied to the residual survival time V, the author derived the pseudo estimating equation as

$$\boldsymbol{U}_\tau(\boldsymbol{\beta})=\sum_{i=1}^{n}\delta_i\left\{\boldsymbol{X}_i-\frac{\sum_{j=1}^{n}\exp(\boldsymbol{\beta}^T\boldsymbol{X}_j)I(Y_j\ge Y_i)\hat{W}_t(Y_j,\delta_j,Y_i)}{\sum_{j=1}^{n}\boldsymbol{X}_j\exp(\boldsymbol{\beta}^T\boldsymbol{X}_j)I(Y_j\ge Y_i)\hat{W}_t(Y_j,\delta_j,Y_i)}\right\}$$

where

$$\hat{W}_t(t,\delta,a)=\delta\left\{\frac{\hat{w}_c(t)-\hat{w}_c(t-a)}{\hat{w}_c(t)}\right\}+(1-\delta)\left\{\frac{\hat{S}_C(t-a)-\hat{S}_C(t)}{1-\hat{S}_C(t)}\right\}$$

and $\hat{S}_C(t)$ is the Kaplan-Meier estimator of the residual censoring time C, and $\hat{w}_c(t)=\int_0^t\hat{S}_C(u)du$. The above estimating equation derived from the proposed profile pseudo-partial likelihood of $\boldsymbol{\beta}$ is reduced to (15) under length-biased sampling without right censoring. When T is subject to right censoring, it is less intuitive by assuming that the biased data (T_i, A_i) are obtained by applying an independent censoring mechanism to the data before the data are sampled with bias.

3.2.3 Weighted Estimating Equation Methods—To estimate covariate coefficients for length-biased data with right censoring, Qin and Shen (2010) constructed two types of **i**nversely-weighted estimating equations under the proportional hazards model. The density function of an unbiased $\tilde{T}$ given $\boldsymbol{X}$ is denoted by $f(t/\boldsymbol{x})$ and the corresponding survival function by $S(t/\boldsymbol{x})$. Assuming (A, V) is independent of residual censoring time C, the probability of observing a pair of uncensored data is

$$\mathrm{pr}(A=a,V=y-a,C\ge y-a|\boldsymbol{X}=\boldsymbol{x})=f(y|\boldsymbol{x})S_C(y-a)/\mu(\boldsymbol{x}),$$

where S_C is the survival distribution for residual censoring variable C. Consequently, the probability of observing the length-biased failure time at y can be obtained by integrating out a:

$$\mathrm{pr}(Y=y,\delta=1|\boldsymbol{X}=\boldsymbol{x})=\frac{f_U(y|\boldsymbol{x})w_c(y)}{\mu(\boldsymbol{x})},\ \ w_c(y)=\int_0^y S_C(t)dt. \quad (17)$$

The following two inversely weighted estimating equations play a similar role in adjusting for the dependent censoring distribution in length-biased data, and both are asymptotically consistent. However, the empirical performance of the estimators is different. For the estimating equation U_1,

$$U_1(\boldsymbol{\beta})=\sum_{i=1}^{n}\delta_i\left[\boldsymbol{X}_i-\frac{\sum_{j=1}^{n}I(Y_j\geq Y_i)\delta_j\boldsymbol{X}_j\exp(\boldsymbol{\beta}^T\boldsymbol{X}_j)\{Y_jS_C(Y_j-A_j)\}^{-1}}{\sum_{j=1}^{n}I(Y_j\geq Y)\delta_j\exp(\boldsymbol{\beta}^T\boldsymbol{X}_j)\{Y_jS_C(Y_j-A_j)\}^{-1}}\right]=0. \quad (18)$$

When S_C is unknown, we can replace it with its consistent Kaplan-Meier estimator for the residual censoring time, which leads to an asymptotic unbiased estimating equation. It is clear that the above weight function $\{Y_j S_C(Y_j - A_j)\}^{-1}$ can be unstable at the tail when $S_C(u) \rightarrow 0$. The authors proposed a more robust estimating equation U_2 as follows, where $w_c(.)$ is an integral of the survival function of $S_C(Y_j - A_j)$ with the numerical stability at the tail,

$$U_2(\boldsymbol{\beta})=\sum_{i=1}^{n}\delta_i\left[\boldsymbol{X}_i-\frac{\sum_{j=1}^{n}I(Y_j\geq Y_i)\delta_j\{w_c(Y_j)\}^{-1}\boldsymbol{X}_j\exp(\boldsymbol{\beta}^T\boldsymbol{X}_j)}{\sum_{j=1}^{n}I(Y_j\geq Y_i)\delta_j\{w_c(Y_j)\}^{-1}\exp(\boldsymbol{\beta}^T\boldsymbol{X}_j)}\right]=0. \quad (19)$$

Note that the weight function $w_c(t)$ is very robust even at tail, when $t \rightarrow \infty$, $w_c(t)$ approximates to the mean of residual censoring time.

3.3 Likelihood-based Approaches

The above estimating equations (referred to as adjusted at-risk set) (Qin and Shen 2010; Tsai 2009) are inefficient as the covariates of the censored subjects are not included in the estimating equations. They also require the censoring time to be independent of the covariates unless a dependence structure is assumed when estimating the weight functions. The following three likelihood-based approaches have the advantage of avoiding the estimation of the censoring distribution and achieving more efficiency compared to the previously mentioned methods.

3.3.1 Full Likelihood Method Based on EM Algorithm—Qin et al (2011) proposed an EM algorithm to maximize the full likelihood to estimate both $\boldsymbol{\beta}$ and baseline hazard function $\Lambda_0(t)$ under the Cox model. In contrast to Vardi's EM algorithm to estimate the nonparametric MLE of G, this alternate algorithm to directly estimate the MLE of F can be modified and extended to handle the semiparametric MLE using the full likelihood.

For random but length-biased samples of n subjects, the observed data consist of $\{\mathcal{O}_i \equiv (A_i, Y_i, \delta_i, \boldsymbol{X}_i), i = 1, \cdots, n\}$, which are n independent and identically distributed (i.i.d.) copies of $\mathcal{O} \equiv (A, Y, \delta, \boldsymbol{X})$. The full likelihood function of the observed data is proportional to

$$L_F=\prod_{i=1}^{n}\frac{f^{\delta_i}(Y_i|\boldsymbol{X}_i)S^{(1-\delta_i)}(Y_i|\boldsymbol{X}_i)}{\mu_{\beta,\Lambda}(\boldsymbol{X}_i)}, \quad (20)$$

where $\mu_{\beta,\Lambda}(\boldsymbol{X}_i)=\int_0^\infty tf(t|\boldsymbol{X}_i)dt=\int_0^\infty S(t|\boldsymbol{Y}_i)dt$. The estimation of the MLE of $\boldsymbol{\beta}$ and the infinite dimensional parameter S can be computationally intractable if directly maximizing (20) or solving its score equations.

Generalizing the approach in Section 2.3.2, but under the semiparametric Cox model, Qin and colleagues proposed the EM algorithm to impute the "missing data," which are the truncated latent data that correspond to each covariate. For $i = 1, \cdots, n$, let T^*_{ij}, $j = 1, 2, ..., m_i$ be the truncated latent data that correspond to covariate $\boldsymbol{X}_i$. The EM algorithm is used to estimate the discretized version of the baseline hazard function $\Lambda(u) = \Sigma_{u \geq t_j} \lambda_j$, where λ_j is the positive jump at the ordered unique time t_j for $j = 1, \cdots, k$, and $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_k)$. For notational convenience, denote $f_i(t) = dF(t/\boldsymbol{X}_i)$. The log-likelihood based on the complete data is then

$$\sum_{j=1}^{k}\sum_{i=1}^{n}\left[I(T_i=t_j)+\sum_{l=1}^{m_i}I(T^*_{il}=t_j)\right]\log f_i(t_j).$$

Conditional on the observed data relative to the ith subject, $\mathcal{O}_i = \{Y_i, A_i, \delta_i, \boldsymbol{X}_i\}$, the expectation of the latent variable can be expressed as

$$\begin{aligned} w_{ij}&=E\left[I(T_i=t_j)+\sum_{l=1}^{m_i}I(T^*_{il}=t_j)|\boldsymbol{\mathcal{O}}_i\right]\\ &=\delta_i I(Y_i=t_j)+(1-\delta_i)\frac{p_{ij}I(Y_i\leq t_j)}{\sum_{j=1}^{k}p_{ij}I(Y_i\leq t_j)}+\frac{\hat{\tau}}{\mu_i}(1-t_j/\hat{\tau})p_{ij}, \end{aligned} \quad (21)$$

where

$$f_i(t_j)=p_{ij}=\lambda_j\exp(\boldsymbol{\beta}'\boldsymbol{X}_i)\exp\left\{-\sum_{l=1}^{j}\lambda_l\exp(\boldsymbol{\beta}'\boldsymbol{X}_i)\right\}, \text{ and } \mu_i=\sum_{j=1}^{k}t_jp_{ij}.$$

Thus, the expected complete-data log-likelihood function conditional on the observed data is

$$\ell_E(\boldsymbol{\beta},\boldsymbol{\lambda})=\sum_{i=1}^{n}\sum_{j=1}^{k}w_{ij}\log p_{ij}. \quad (22)$$

In the M-step, maximizing $\ell_E(\boldsymbol{\beta}, \boldsymbol{\lambda})$ with respect to the baseline hazard function at t_j leads to a closed form of λ_j as a function of $\boldsymbol{\beta}$,

$$\lambda_j(\boldsymbol{\beta}) = \frac{\sum_{i=1}^{n} w_{ij}}{\sum_{l=j}^{k} \sum_{i=1}^{n} w_{il} \exp(\boldsymbol{\beta}^T \boldsymbol{X}_i)}.$$

After inserting $\boldsymbol{\beta}$ into the score equation of $\boldsymbol{\beta}$ derived from (22), $\boldsymbol{\beta}$ can be solved using existing software for the analysis of conventional right-censored data under the Cox model, which is a computational advantage. With the estimated $\boldsymbol{\beta}$ and $\lambda_j(\boldsymbol{\beta})$, one can update the expected likelihood via updating w_{ij}. The estimators of $\boldsymbol{\beta}$ and $\lambda_j(\boldsymbol{\beta})$ are obtained by repeating the iterative steps.

3.3.2 Composite Partial Likelihood Method—Huang and Qin (2012) introduced the composite partial likelihood approach for estimating covariate coefficients under the Cox model for length-biased data. In general, directly maximizing the full likelihood function (20) with respect to $(\Lambda, \boldsymbol{\beta})$ is computationally intensive, even though it leads to the most efficient estimators. The full likelihood can be factored to $L_F = L_C \times L_M$, where L_C is as defined in (12), and can be further factored to the partial likelihood L_P and a remaining ancillary term as described in Section 3.1. As demonstrated in the literature (Qin and Shen 2010; Shen et al 2009), the estimator of $\boldsymbol{\beta}$ obtained from the partial likelihood L_P is inefficient. Using the exchangeable (or symmetric) property of (A, V), their joint density is $f(a + v/x)/\mu(x)$ and their marginal distributions are

$$f_A(t|x) = f_V(t|x) = S(t|x)/\mu(x).$$

Therefore, the density of $T = A + V$ given A is the same as T given V. Based on the partial likelihood function L_P and the symmetric distribution property of (A, V) when there is no right censoring, Huang and Qin (2012) proposed following the composite likelihood by doubling the information of A using V. Specifically, the likelihood for the pooled data $\{(T_i, A_i, \boldsymbol{X}_i), i = 1, \cdots n\}$ and $\{(T_i, V_i, \boldsymbol{X}_i), i = 1, \cdots n\}$ is

$$L_{2C}(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left\{ \frac{f(T_i|\boldsymbol{X}_i)}{S(A_i|\boldsymbol{X}_i)} \times \frac{f(T_i|\boldsymbol{X}_i)}{S(V_i|\boldsymbol{X}_i)} \right\}. \quad (23)$$

When V is subject to right censoring, the symmetric property of (A, V) does not hold since A is not subject to right censoring. However, it is interesting that the conditional density of $A = a$ given $V = v$ and $\delta = 1$ has the same density function as the conditional density of V given A without right censoring,

$$\frac{f(a+v|x)}{S(v|x)}, a>0, v>0.$$

Using the above property, a different composite conditional likelihood for right-censored, length-biased data is formulated by augmenting the information among the uncensored subjects only,

$$L_{2C}^{C}(\boldsymbol{\beta})=\prod_{i=1}^{n}\left\{\frac{f(Y_i|\boldsymbol{X}_i)}{S(A_i|\boldsymbol{X}_i)}\times\frac{f(Y_i|\boldsymbol{X}_i)}{S(V_i|\boldsymbol{X}_i)}\right\}^{\delta_i}\left\{\frac{S(Y_i|\boldsymbol{X}_i,\boldsymbol{\beta})}{S(A_i|\boldsymbol{X}_i,\boldsymbol{\beta})}\right\}^{1-\delta_i}. \quad (24)$$

Specifically, the pooled data are $\{(Y_i, A_i, \delta_i, \boldsymbol{X}_i), i = 1, \cdots n\}$ combined with the uncensored subset $\{(Y_i, V_i, \boldsymbol{X}_i)$, among $\delta_i = 1, i = 1, \cdots, n\}$. By inserting the Breslow-type estimator for the baseline hazard function $\Lambda(.)$ into the above likelihood function, the corresponding score equation for estimating $\boldsymbol{\beta}$ yields

$$\boldsymbol{U}_{C}(\boldsymbol{\beta})=\sum_{i=1}^{n}\delta_i\left[\boldsymbol{X}_i-\frac{\sum_{j=1}^{n}\boldsymbol{X}_j\exp(\boldsymbol{\beta}^T\boldsymbol{X}_j)\{I(A_j\le Y_i\le Y_j)+\delta_j I(V_j\le Y_i\le Y_j)\}}{\sum_{j=1}^{n}\exp(\boldsymbol{\beta}^T\boldsymbol{X}_j)\{I(A_j\le Y_i\le Y_j)+\delta_j I(V_j\le Y_i\le Y_j)\}}\right]=0. \quad (25)$$

Note the similarity and difference between the conditional estimating equation for general left-truncated data (14) and the above augmented estimating equation for the augmented data. For the variance estimator of $\hat{\boldsymbol{\beta}}$, the augmented and original data have overlapping information; therefore, the pooled data should be treated as clustered survival data to adjust for the correlation. The composite likelihood approach can improve statistical efficiency compared to some of the estimating equation approaches without modeling the censoring distribution.

3.3.3 Maximum Pseudo-Profile Likelihood Method—Under the proportional hazards model, Huang et al (2012) introduced a maximum pseudo-profile likelihood approach, which can improve efficiency and naturally handle time-dependent covariates. By generalizing the profile likelihood method via replacing the nuisance parameters in the full or partial likelihood with a consistent estimator, the authors replaced the baseline hazard function with infinite dimensional parameters in the full likelihood with a Breslow-type estimator for the baseline hazard function to attain a pseudo-profile likelihood function. For a known $\boldsymbol{\beta}$, a natural consistent estimate of $\Lambda_0(t)$ can be obtained from the conditional (truncation) likelihood $L_C(\boldsymbol{\beta}, \Lambda_0)$ defined in Section 3.1,

$$\hat{\Lambda}_{\boldsymbol{\beta}}(t)=\int_0^t\frac{d\sum_{i=1}^{n}\delta_i I(Y_i\le u)}{\sum_{i=1}^{n}\exp(\boldsymbol{\beta}^T\boldsymbol{X}_i)I(A_i\le u\le Y_i)} \quad (26)$$

in the class of nondecreasing right-continuous functions that jump only at uncensored failure times. Profiling out $\Lambda_0(.)$ from the conditional likelihood $L_C(\boldsymbol{\beta}, \Lambda_0)$ leads to the partial likelihood, $L_P(\boldsymbol{\beta})$ defined in (13). This is why the estimators of $\boldsymbol{\beta}$ obtained from maximizing L_P retain the same efficiency as those from L_C, as proved by Wang et al (1993) when there is no information on the distribution of the truncation time. Under length-biased sampling, the maximum partial likelihood estimator obtained from (13) is not efficient, since the likelihood contribution from $L_M(\boldsymbol{\beta}, \hat{\Lambda}_{\boldsymbol{\beta}})$ is ignored. The resultant pseudo-profile likelihood follows,

$$L_C(\boldsymbol{\beta}, \hat{\Lambda}_{\boldsymbol{\beta}}) = L_P(\boldsymbol{\beta}) L_M(\boldsymbol{\beta}, \hat{\Lambda}_{\boldsymbol{\beta}}) = L_P(\boldsymbol{\beta}) \prod_{i=1}^{n} \frac{\exp\left\{-\hat{\Lambda}_{\boldsymbol{\beta}}(A_i)\exp(\boldsymbol{\beta}^T \boldsymbol{X}_i)\right\}}{\int_0^\infty \exp\left\{-\hat{\Lambda}_{\boldsymbol{\beta}}(u)\exp(\boldsymbol{\beta}^T \boldsymbol{X}_i)\right\} du}.$$

Similar to the other two methods in Section 3.3 (Huang and Qin 2012; Qin et al 2011), this method is more robust compared with the methods that use estimating equations, when the censoring distribution depends on covariates and/or the censoring proportion is high.

4 Semiparametric Accelerated Failure Time Model: Estimation Methods

The accelerated failure time (AFT) model, which linearly relates covariates to the logarithm of the survival time, has been one of the regression models most commonly used for analyzing right-censored survival data besides the proportional hazards model (Kalbfleisch and Prentice 2002). Assuming that the failure time in the target population follows the AFT model,

$$\log \tilde{T} = \boldsymbol{X}^T \boldsymbol{\beta}_0 + \varepsilon, \quad (27)$$

where $\boldsymbol{\beta}_0$ is a $p \times 1$ parameter vector and ε is independent of $\boldsymbol{X}$ with an unspecified probability density distribution function $q(.)$.

4.1 Estimating Equation Methods without Right Censoring

Chen (2010) considered a special case for length-biased data without right censoring. Under the semiparametric AFT model, the author derived the hazard-based estimating equation for length-biased data,

$$\boldsymbol{U}_A(\boldsymbol{\beta}) = \sum \int_0^\infty \boldsymbol{X}_i \left\{ dN_i(ye^{-\boldsymbol{\beta}^T \boldsymbol{X}_i}) - \Delta_i(ye^{-\boldsymbol{\beta}^T \boldsymbol{X}_i}) d\hat{\Lambda}_\eta(y;\boldsymbol{\beta}) \right\} = 0,$$

where $N_i(y) = I(T_i \le y)$, $\Delta_i(y) = I(T_i > y)$, and $\hat{\Lambda}_\eta(y;\boldsymbol{\beta}) = \int_0^y \left\{ \frac{\sum_i dN_i(ue^{-\boldsymbol{\beta}_0^T \boldsymbol{X}_i})}{\sum_i \Delta_i(ue^{-\boldsymbol{\beta}_0^T \boldsymbol{X}_i})} \right\}$. The model structure for the observed T (length-biased) is generally different from that for $\tilde{T}$ (unbiased) in the target population when the failure time is subject to right censoring. However, Chen

(2010) observed there is a unique feature for length-biased data under the AFT model: the observed length-biased failure time follows an AFT model with the same regression coefficients except for the intercept if $\tilde{T}$ follows an AFT model. By using this invariance property, Mandel and Ritov (2010) also proposed to use the standard least square method for analyzing length-biased data under the AFT model. The numerical studies showed that the least squares estimator outperformed the estimator of Chen (2010), with smaller standard errors. However, both methods are not applicable to length-biased data with potential right censoring.

4.2 Conditional Estimating Equation Approach

For general left-truncated and right-censored data, Lai and Ying (1991) proposed a rank-based estimating equation for $\boldsymbol{\beta}$, based on the constructed at-risk set at t as

$$R^*(t,\boldsymbol{b})=\{i\le n: A_i\exp(-\boldsymbol{X}_i^T\boldsymbol{b})<t\le Y_i\exp(-\boldsymbol{X}_i^T\boldsymbol{b})\},$$

$$\boldsymbol{U}_{LT}(\boldsymbol{\beta})=\sum_{i=1}^{n}\delta_i\left[\boldsymbol{X}_i-\frac{\sum_{j=1}^{n}\boldsymbol{X}_j I\{j\in R^*(Y_i e^{-\boldsymbol{X}_i^T\boldsymbol{\beta}},\boldsymbol{\beta})\}}{\sum_{j=1}^{n} I\{j\in R^*(Y_i e^{-\boldsymbol{X}_i^T\boldsymbol{\beta}},\boldsymbol{\beta})\}}\right]. \quad (28)$$

For length-biased data, this estimating equation ignores the information on the truncation times and may result in efficiency loss. By effectively utilizing information contained in the truncation time, Ning et al (2014b) modified equation (28) by replacing the indicator function $I\{j\in R^*(Y_i e^{-\boldsymbol{X}_i^T\boldsymbol{\beta}},\boldsymbol{\beta})\}$ with its conditional expectation given the observed information under the stationarity assumption,

$$\boldsymbol{U}_M(\boldsymbol{\beta})=\sum_{i=1}^{n}\int_0^{\infty}\left\{\boldsymbol{X}_i-\frac{\sum_{j=1}^{n}\boldsymbol{X}_j R(Y_j e^{-\boldsymbol{X}_j^T\boldsymbol{\beta}},t)}{\sum_{j=1}^{n} R(Y_j e^{-\boldsymbol{X}_j^T\boldsymbol{\beta}},t)}\right\} dN_i(\boldsymbol{\beta},t),$$

where $N_i(\boldsymbol{\beta},t)=I\left(Y_i e^{-\boldsymbol{X}_i^T\boldsymbol{\beta}}\ge t,\delta_i=1\right)$, and

$$R(y,t)=\delta I(y\ge t)\left\{\frac{\hat{w}_c(y)-\hat{w}_c(y-t)}{\hat{w}_c(y)}\right\}+(1-\delta)I(y\ge t)\left\{\frac{\hat{S}_C(y-t)-\hat{S}_C(y)}{1-\hat{S}_C(y)}\right\}.$$

The modified estimating equation relies on both the magnitude and rank of the failure times and achieves more efficiency compared to the estimators obtained from rank-based equation (28).

4.3 Estimating Equations with Inverse Weight

In the presence of right censoring, Shen et al (2009) constructed the following straightforward inverse-weighting estimating equation to account for the informative censoring due to the sampling constraint,

$$\boldsymbol{U}_I(\boldsymbol{\beta})=\sum_{i=1}^{n}\frac{\delta_i \boldsymbol{X_i}}{\hat{w}_c(Y_i)}(\log Y_i-\boldsymbol{X_i}'\boldsymbol{\beta})=0. \quad (29)$$

Although it is not the most statistically efficient estimator, the inverse-weighted estimator is the most computationally efficient, with a simple closed form of the solution from the above estimating equation,

$$\hat{\boldsymbol{\beta}}_I=\left\{\sum_{i=1}^{n}\frac{\delta_i \boldsymbol{X_i}\boldsymbol{X_i}'}{\hat{w}_c(Y_i)}\right\}^{-1}\sum_{i=1}^{n}\frac{\delta_i \boldsymbol{X_i}\log Y_i}{\hat{w}_c(Y_i)}.$$

Similar to the inverse-weighted estimating equation approach for the proportional hazards model, this estimating equation approach requires the assumption that covariates are independent of the residual censoring time. In addition, since (29) does not use the covariate data from right-censored subjects, the estimator can be inefficient.

4.4 Buckley-James-Type Estimator

Under the AFT model (27), Ning et al (2011) constructed a Buckley-James-type estimating equation and developed an iterative algorithm to obtain the root of the estimating equation to overcome the aforementioned limitations from the above estimating equations. The covariate-specific density function of the unbiased failure time $\tilde{T}$ and the corresponding length-biased density function can be expressed by

$$\begin{aligned} f(t|\boldsymbol{x}) &= \tfrac{1}{t}q(\log t-\boldsymbol{x}^T\boldsymbol{\beta}_0),\ t>0 \\ g(t|\boldsymbol{x}) &= \tfrac{q(\log t-\boldsymbol{x}^T\boldsymbol{\beta}_0)}{\mu(\boldsymbol{x})},\ t>0, \end{aligned}$$

where $\mu(\boldsymbol{x}) = \int tf(t|\boldsymbol{x})dt = \int q(\log y - \boldsymbol{x}^T\boldsymbol{\beta}_0)dy$. Define $\tilde{T}_0 = \tilde{T}\exp(-\boldsymbol{X}^T\boldsymbol{\beta}_0)$, $T_0 = T\exp(-\boldsymbol{X}^T\boldsymbol{\beta}_0)$, and $\mu = \int q(\log t)dt$. Then the density function of $\tilde{T}_0$ is $dF_0(t;\boldsymbol{\beta}) = q(\log t)/t$.

Using the same principle as for traditional right-censored survival data, the Buckley-James-type estimating equation for length-biased data can be constructed as follows:

$$\boldsymbol{U}_0(\boldsymbol{\beta})=\sum_{i=1}^{n}\boldsymbol{X}_i\left\{\delta_i\frac{\log Y_i-\boldsymbol{X}_i^T\boldsymbol{\beta}}{Y_i\exp(-\boldsymbol{X}_i^T\boldsymbol{\beta})}+(1-\delta_i)\frac{\int_{Y_{i0}}^{\infty}u^{-1}\log u dF_0(u;\boldsymbol{\beta})}{1-F_0(Y_{i0};\boldsymbol{\beta})}\right\},$$

where the second term represents that the conditional expectation of the right-censored likelihood is based on the transformed data, $Y_{i0}=\min(T_i e^{-\boldsymbol{X}_i^T\beta_0}, C_i e^{-\boldsymbol{X}_i^T\beta_0})$. For the unknown distribution F_0, we can use Vardi's method (Vardi 1989) to derive its NPMLE using the transformed i.i.d. data $\{Y_{i0}, \delta_i\}$.

The Buckley-James-type estimator $\hat{\boldsymbol{\beta}}_{BJ}$ is then defined as the root of $\boldsymbol{U}_{BJ}(\boldsymbol{\beta}) = 0$, where

$$\boldsymbol{U}_{BJ}(\boldsymbol{\beta})=\sum_{i=1}^{n}\boldsymbol{X}_i\left\{\delta_i\frac{\log Y_i-\boldsymbol{X}_i^T\boldsymbol{\beta}}{Y_i\exp(-\boldsymbol{X}_i^T\boldsymbol{\beta})}+(1-\delta_i)\frac{\int_{Y_{i0}}^{\infty}u^{-1}\log u\, d\hat{F}_0(u;\boldsymbol{\beta})}{1-\hat{F}_0(Y_{i0};\boldsymbol{\beta})}\right\}. \quad (30)$$

Compared with the inverse-weighted estimating equation approach, the Buckley-James-type method effectively utilizes all information from censored observations, with their conditional expectations in the estimating equations. Another advantage of the Buckley-James-type method is that the estimation procedure does not require the assumption that the residual censoring time is independent of the covariate.

4.5 Estimating Equations from Embedded Likelihood Functions

As demonstrated in Section 4.4, using the Buckley-James estimation approach, one appealing feature of the AFT model is that the observed failure time data can be transformed to i.i.d. random variables without covariate effects. Using this unique feature, Ning et al (2014a) proposed a class of estimating equations based on the score functions for the transformed i.i.d. data, which are derived from the full likelihood function under commonly used semiparametric models such as the proportional hazards or proportional odds model.

Under the AFT model, the transformed time $T_0 = \tilde{T}e^{-\boldsymbol{X}^T\boldsymbol{\beta}^0}$, which has null effect for the covariates, can be assumed to follow the Cox proportional hazards model or other popular semi-parametric models. For illustration, one can assume the null embedded model as the Cox model,

$$\lambda(t|\boldsymbol{X})=\lambda_0(t)\exp(\boldsymbol{X}^T\boldsymbol{\alpha}), \quad (31)$$

where $\lambda_0(t)$ is an unspecified baseline hazard function and $\lambda(t|\boldsymbol{X})$ is the hazard function given covariate $\boldsymbol{X}$.

Under the proportional hazards embedded model assumption with a null covariate effect, the full log-likelihood for the transformed, observed data, $\{ Y_{i0}=Y_ie^{-\boldsymbol{X}_i^T\boldsymbol{\beta}_0}, A_{i0}=A_ie^{-\boldsymbol{X}_i^T\boldsymbol{\beta}_0}, \delta_i, \boldsymbol{X}_i, i = 1, \cdots, n\}$ can be expressed as

$$\begin{aligned}\tilde{l}_{PH}(\Lambda_0,\boldsymbol{\alpha})&=\sum_{i=1}^{n}\Big[\delta_i\left\{\log\lambda_0(Y_{i0})+\boldsymbol{X}_i^T\boldsymbol{\alpha}\right\}\\&\quad-\Lambda_0(Y_{i0})\exp(\boldsymbol{X}_i^T\boldsymbol{\alpha})\\&\quad-\log\int\exp\left\{-\Lambda_0(s)\exp(\boldsymbol{X}_i^T\boldsymbol{\alpha})\right\}ds\Big]\\&\quad+\sum_{i=1}^{n}(1-\delta_i)\boldsymbol{X}_i^T\boldsymbol{\beta}_0.\end{aligned}$$

The corresponding score function of $\boldsymbol{\alpha}$ evaluated at the null covariate effects ($\boldsymbol{\alpha} = 0$) has a mean of zero (Ning et al 2010) and can be used as an unbiased estimating equation for solving $\boldsymbol{\beta}$,

$$\tilde{\boldsymbol{E}}_{PH}(\Lambda_0, \boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{X}_i \left\{ \delta_i - \Lambda_0(Y_i e^{-\boldsymbol{X}_i^T \boldsymbol{\beta}}) + \frac{\int \Lambda_0(s) \exp\{-\Lambda_0(s)\}\, ds}{\int \exp\{-\Lambda_0(s)\}\, ds} \right\}. \quad (32)$$

The unknown baseline hazard function $\Lambda_0(t)$ in equation (32) can be estimated through

$$\hat{\Lambda}_0(t) = -\log \left\{ \frac{\int_t^{\infty} u^{-1} d\hat{G}(u)}{\int_0^{\infty} u^{-1} d\hat{G}(u)} \right\},$$

where $\hat{G}(.)$ is the NPMLE introduced in Section 2.3.1.

Note that although the Cox model is chosen to illustrate the principle behind the method, other semiparametric models such as the proportional odds model can be used in this framework as an embedded model. Compared to the inverse-weighted and Buckley-James-type estimating equations, the score-based estimating equations lead to more efficiency gain, which may be achieved because the proposed estimating equations are directly derived from the embedded full likelihood function.

5 Other Semiparametric Models and Developments

Semiparametric linear transformation models, which include the proportional hazards model and proportional odds model as special cases, have been used in conventional survival data analyses for decades. The semiparametric transformation model may be specified as $H(\tilde{T}) = -\boldsymbol{X}^T \boldsymbol{\beta} + \varepsilon$, where $H(.)$ is an unknown increasing function, and ε has a known density function. For right-censored, length-biased data under the transformation models, Shen et al (2009) proposed a rank-based estimating equation approach, Wang and Wang (2015) constructed estimating equations based on counting processes, Liu et al (2012) introduced a general estimation and inference procedure by using an imputation method, and Liu et al (2015) described a maximum likelihood method for general truncated data under the semiparametric transformation model. Kim et al (2013) proposed an inference procedure using weighted estimating equations under several biased sampling schemes, including length-biased sampling, in which they assumed a different right-censoring mechanism. Cheng and Huang (2014) then proposed a method that combines two estimation procedures: the martingale estimating equation based on the partial likelihood function and the pseudo-partial score equation. To handle varying coefficients, Lin and Zhou (2014); Zhang et al (2014) proposed the local inverse probability weighted estimating equation for right-censored, length-biased data under the semiparametric linear transformation models and the Cox model, respectively.

Extending the work of Wang (1996), Shen (2009) proposed respective estimating equations for the additive hazards and proportional hazards models. Using a pairwise pseudolikelihood to eliminate nuisance parameters from the marginal likelihood, Huang and Qin (2013) proposed an estimating function in obtaining the coefficient parameters under the additive hazards model for left-truncated, right-censored data.

Other semiparametric models, such as the semiparametric density ratio model and the proportional mean residual model, have been proposed for analyzing right-censored, length-biased data (Chan et al 2012; Davidov et al 2010; Shen et al 2012). Additional developments have addressed the estimation of distributions for baseline covariates in the target population given the observed biased data under length-biased sampling (Chan and Wang 2012), and have investigated issues associated with efficiency for covariate estimates under parametric models (Bergeron et al 2008; Cook and Bergeron 2011). Keiding et al (2011) investigated the parametric AFT regression models to analyze backward recurrence times in a pregnancy study.

6 Discussion

In summary, methodologic development in semiparametric and nonparametric modeling of length-biased data has made considerable progress in recent years in many different directions. In particular, nonparametric and semiparametric maximum likelihood inference based on the full likelihood method has attained both robustness and efficiency compared to methods based on estimating equations or other types of likelihood functions. Future research will include additional promising methodologic developments as well as related software for the implementation of such methods.

More importantly, we should educate non-statistician collaborators to be aware of sampling bias when reporting analytic results from prevalent cohort studies and cancer screening trials, and to properly adjust for such biases. Although the statistics and epidemiology literature on biased sampling can be traced back many decades and has been widely noted by statisticians (Asgharian et al 2014; Cox 1962; Simon 1980), there is a need for practitioners to properly implement such methods.

Acknowledgments

The work was supported in part by the U.S. NIH grants CA079466 and CA016672. The authors thank Professor Asgharian and the investigators from the Canadian Study of Health and Aging for generously sharing the dementia data. The data reported in this article were collected as part of the Canadian Study of Health and Aging. The core study was funded by the Seniors' Independence Research Program, through the National Health Research and Development Program (NHRDP) of Health Canada Project 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.

References

Addona V, Wolfson DB. A formal test for the stationarity of the incidence rate using data from a prevalent cohort study with follow-up. Lifetime data analysis. 2006; 12(3):267–284. [PubMed: 16917734]

Andersen, P., Borgan, O., Gill, R., Keiding, N. Statistical Models Based on Counting Processes. Springer-Verlag; New York: 1993.

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

Asgharian M, Wolfson DB, et al. Asymptotic behavior of the unconditional npmle of the length-biased survivor function from right censored prevalent cohort data. The Annals of Statistics. 2005; 33(5): 2109–2131.

Asgharian M, Wolfson DB, Zhang X. Checking stationarity of the incidence rate using prevalent cohort survival data. Statistics in medicine. 2006; 25(10):1751–1767. [PubMed: 16220462]

Asgharian M, Wolfson C, Wolfson DB. Analysis of biased survival data: the canadian study of health and aging and beyond. Stat Action Can Outlook. 2014:193–208.

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

Chan KCG, Wang MC. Estimating incident population distribution from prevalent data. Biometrics. 2012; 68(2):521–531. [PubMed: 22313264]

Chan KCG, Chen YQ, Di CZ. Proportional mean residual life model for right-censored length-biased data. Biometrika. 2012:ass049.

Chen YQ. Semiparametric regression in size-biased sampling. Biometrics. 2010; 66(1):149–158. [PubMed: 19432792]

Cheng YJ, Huang CY. Combined estimating equation approaches for semiparametric transformation models with length-biased survival data. Biometrics. 2014; 70(3):608–618. [PubMed: 24750126]

Cook RJ, Bergeron PJ. Information in the sample covariate distribution in prevalent cohorts. Statistics in medicine. 2011; 30(12):1397–1409. [PubMed: 21259303]

Cox, DR. Renewal theory. Methuen; London: 1962.

Cox DR. Regression models and life-tables. Journal of the Royal Statistical Society Series B (Methodological). 1972:187–220.

Davidov O, Fokianos K, Iliopoulos G. Order-restricted semiparametric inference for the power bias model. Biometrics. 2010; 66(2):549–557. [PubMed: 19522874]

Ghosh D. Proportional hazards regression for cancer studies. Biometrics. 2008; 64(1):141–148. [PubMed: 17573863]

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

Gross ST, Lai TL. Nonparametric estimation and regression analysis with left-truncated and right-censored data. Journal of the American Statistical Association. 1996; 91(435):1166–1180.

Huang, Cy, Qin, J. Composite partial likelihood estimation under length-biased sampling, with application to a prevalent cohort study of dementia. Journal of the American Statistical Association. 2012; 107(499):946–957. [PubMed: 24000265]

Huang CY, Qin J. Semiparametric estimation for the additive hazards model with left-truncated and right-censored data. Biometrika. 2013:ast039.

Huang CY, Qin J, Follmann DA. A maximum pseudo-profile likelihood estimator for the cox model under length-biased sampling. Biometrika. 2012:asr072.

Kalbfleisch JD, Lawless JF. Inference based on retrospective ascertainment: An analysis of the data on transfusion-related aids. Journal of the American Statistical Association. 1989; 84:360–372.

Kalbfleisch, JD., Prentice, RL. The Statistical Analysis of Failure Time Data. 2. Wiley-Interscience [John Wiley & Sons]; Hoboken, NJ: 2002.

Keiding, N. Independent delayed entry. In: Klein, JP., Goel, P., editors. Survival Analysis: State of the Art. Kluwer Academic Publishers; Boston: 1992. p. 309-326.

Keiding N, Gill RD. Random truncation models and markov processes. Annals of Statistics. 1990; 18:582–602.

Keiding N, Kvist K, Hartvig H, Tvede M, Juul S. Estimating time to pregnancy from current durations in a cross-sectional sample. Biostatistics. 2002; 3:565–578. [PubMed: 12933598]

Keiding N, Fine JP, Hansen OH, Slama R. Accelerated failure time regression for backward recurrence times and current durations. Statistics & Probability Letters. 2011; 81(7):724–729.

Kim JP, Lu W, Sit T, Ying Z. A unified approach to semiparametric transformation models under general biased sampling schemes. Journal of the American Statistical Association. 2013; 108(501): 217–227. [PubMed: 23667280]

Kvam P. Length bias in the measurements of carbon nanotubes. Technometrics. 2008; 50(4):462–467.

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

Lai TL, Ying Z. Rank regression methods for left-truncated and right-censored data. Annals of Statistics. 1991; 19:531–556.

Lin C, Zhou Y. Analyzing right-censored and length-biased data with varying-coefficient transformation model. Journal of Multivariate Analysis. 2014; 130:45–63.

Liu H, Qin J, Shen Y. Imputation for semiparametric transformation models with biased-sampling data. Lifetime data analysis. 2012; 18(4):470–503. [PubMed: 22903245]

Liu H, Ning J, Qin J, Shen Y. Semiparametric maximum likelihood inference for truncated or biased-sampling data. Statistica Sinica. 2015 in press.

Mandel M, Betensky RA. Testing goodness of fit of a uniform truncation model. Biometrics. 2007; 63(2):405–412. [PubMed: 17688493]

Mandel M, Ritov Y. The accelerated failure time model under biased sampling. Biometrics. 2010; 66(4):1306–1308. [PubMed: 19995351]

Martin EC, Betensky RA. Testing quasi-independence of failure and truncation times via conditional kendall's tau. Journal of the American Statistical Association. 2005; 100(470):484–492.

Ning J, Qin J, Shen Y. Non-parametric tests for right-censored data with biased sampling. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2010; 72(5):609–630. [PubMed: 21031144]

Ning J, Qin J, Shen Y. Buckley–james-type estimator with right-censored and length-biased data. Biometrics. 2011; 67(4):1369–1378. [PubMed: 21385160]

Ning J, Qin J, Shen Y. Score estimating equations from embedded likelihood functions under accelerated failure time model. Journal of the American Statistical Association. 2014a; 109(508): 1625–1635. [PubMed: 25663727]

Ning J, Qin J, Shen Y. Semiparametric accelerated failure time model for length-biased data with application to dementia study. Statistica Sinica. 2014b; 24(1):313. [PubMed: 24478570]

Nowell C, Stanley LR. Length-biased sampling in mall intercept surveys. Journal of Marketing Research. 1991; 28(4):475–479.

Qin J, Shen Y. Statistical methods for analyzing right-censored length-biased data under Cox model. Biometrics. 2010; 66(2):382–392. [PubMed: 19522872]

Qin J, Ning J, Liu H, Shen Y. Maximum likelihood estimations and em algorithms with length-biased data. Journal of the American Statistical Association. 2011; 106(496):1434–1449. [PubMed: 22323840]

Shen PS. Hazards regression for length-biased and right-censored data. Statistics & Probability Letters. 2009; 79(4):457–465.

Shen Y, Ning J, Qin J. Analyzing length-biased data with semiparametric transformation and accelerated failure time models. Journal of the American Statistical Association. 2009; 104(487): 1192–1202. [PubMed: 21057599]

Shen Y, Ning J, Qin J. Likelihood approaches for the invariant density ratio model with biased-sampling data. Biometrika. 2012; 99(2):363–378. [PubMed: 23843663]

Simon R. Length biased sampling in etiologic studies. American Journal of Epidemiology. 1980; 111(4):444–452. [PubMed: 7377187]

Terwilliger JD, Shannon WD, Lathrop GM, Nolan JP, Goldin LR, Chase GA, Weeks DE. True and false positive peaks in genomewide scans: Applications of length-biased sampling to linkage mapping. American Journal of Human Genetics. 1997; 61(2):430–438. [PubMed: 9311749]

Tsai WY. Pseudo-partial likelihood for proportional hazards models with biased-sampling data. Biometrika. 2009; 96(3):601–615. DOI: 10.1093/biomet/asp026 [PubMed: 22422175]

Turnbull BW. The empirical distribution function with arbitrarily grouped, censored and truncated data. Journal of the Royal Statistical Society Series B (Methodological). 1976:290–295.

de Una-Álvarez J, Otero-Giráldez MS, Álvarez-Llorente G. Estimation under length-bias and right-censoring: An application to unemployment duration analysis for married women. Journal of Applied Statistics. 2003; 30(3):283–291.

Vardi Y. Nonparametric estimation in the presence of length bias. Annals of Statistics. 1982; 10(2): 616–620.

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

Wang MC. Nonparametric estimation from cross-sectional survival data. Journal of the American Statistical Association. 1991; 86:130–143.

Wang MC. Hazards regression analysis for length-biased data. Biometrika. 1996; 83(2):343–354.

Wang MC, Jewell NP, Tsai WY. Asymptotic properties of the product limit estimate under random truncation. Annals of Statistics. 1986; 14:1597–1605.

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

Wang X, Wang Q. Estimation for semiparametric transformation models with length-biased sampling. Journal of Statistical Planning and Inference. 2015; 156:80–89.

Wolfson C, Wolfson DB, Asgharian M, M'Lan CE, Østbye T, Rockwood K, Hogan Df. A reevaluation of the duration of survival after the onset of dementia. New England Journal of Medicine. 2001; 344(15):1111–1116. [PubMed: 11297701]

Woodroofe M. Estimating a distribution function with truncated data. The Annals of Statistics. 1985:163–177.

Zelen M. Forward and backward recurrence times and length biased sampling: Age specific models. Lifetime Data Analysis. 2004; 10(4):325–334. [PubMed: 15690988]

Zhang F, Chen X, Zhou Y. Proportional hazards model with varying coefficients for length-biased data. Lifetime data analysis. 2014; 20(1):132–157. [PubMed: 23649724]

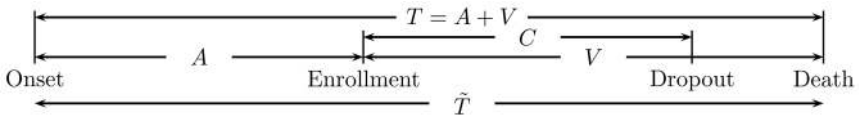

Figure 1.
Data sampling with right censoring