Published in final edited form as:

J Am Stat Assoc. 2009 September 1; 104(487): 1192–1202. doi:10.1198/jasa.2009.tm08614.

Analyzing Length-biased Data with Semiparametric Transformation and Accelerated Failure Time Models

Yu Shen,

Department of Biostatistics M. D. Anderson Cancer Center The University of Texas, Houston, TX 77030 yshen@mdanderson.org Phone: 713-794-4159, Fax: 713-563-4242

Jing Ning, and

Department of Biostatistics M. D. Anderson Cancer Center The University of Texas, Houston, TX 77030

Jing Qin

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

SUMMARY

Right-censored time-to-event data are often observed from a cohort of prevalent cases that are subject to length-biased sampling. Informative right censoring of data from the prevalent cohort within the population often makes it difficult to model risk factors on the unbiased failure times for the general population, because the observed failure times are length biased. In this paper, we consider two classes of flexible semiparametric models: the transformation models and the accelerated failure time models, to assess covariate effects on the population failure times by modeling the length-biased times. We develop unbiased estimating equation approaches to obtain the consistent estimators of the regression coefficients. Large sample properties for the estimators are derived. The methods are confirmed through simulations and illustrated by application to data from a study of a prevalent cohort of dementia patients.

Keywords

Accelerated failure time models; Estimating equation; Informative censoring; Length-biased sampling; Prevalent cohort; Right-censored data; Transformation models

1 Introduction

Length-biased data are often encountered in observational studies, when the observed samples are not randomly selected from the population of interest but with probability proportional to their length (Cox and Miller, 1965; Zelen and Feinleib, 1969; Simon, 1980; Vardi, 1982, 1985, 1989; Sansgiry and Akman, 2000; Zelen, 2004). The data set considered is often from a cross-sectional cohort of subjects diagnosed to have the disease at the time of examination, which is then followed for the occurrence of a subsequent event.

Authors' Footnote: Yu Shen is Professor, Department of Biostatistics, M.D. Anderson Cancer Center, Holcombe Blvd, Unit 1411, Houston, TX 77030 (yshen@mdanderson.org); Jing Ning is Postdoctoral Fellow, Department of Biostatistics, M.D. Anderson Cancer Center, Holcombe Blvd, Unit 1411, Houston, TX 77030 (jning@mdanderson.org); and Jing Qin is Mathematical Statistician, National Institute of Allergy and Infectious Diseases, Bethesda, MD 20817 (jingqin@niaid.nih.gov). This research was partially supported by National Institute of Health grant R01-CA079466.

Two examples of such data are seen in a study of dementia and subsequent onset of death, and a study of the natural history of lung cancer. In the first example, about 10,000 Canadians over the age of 65 were recruited and screened for prevalence of dementia. Investigators recorded the approximate date of onset of dementia and the subsequent time of death or censoring for individuals within the study population who were found to have dementia (Asgharian and Wolfson, 2005). The second example involves a prevalent cohort for elderly lung cancer patients. Individuals are sampled from Surveillance Epidemiology and End Results (SEER) database. The sampling (or recruitment) times can be considered to follow a uniform distribution independent of failure times, only those individuals who had been diagnosed to have lung cancer before the recruitment time and who had been alive are eligible for inclusion. The data on each subject in the prevalent cohort included an initiating event (e.g., diagnosis of lung cancer) and failure event (e.g., death) for subjects who had been sampled at an intermediate time. A common feature for data from the prevalent cohort is that the time duration measured from the first (or initiating) event to the terminal event is subject to potential left truncation and right censoring. Length-biased sampling occurs because the “observed” time intervals from initiation to failure within the prevalent cohort tend to be longer than those arising from the underlying distribution of the general population, because individuals diagnosed with the disease have to survive to the examination or sampling time.

Extensive literature has focused on estimating the unbiased distribution given length-biased sampling, using methods conditional on the observed truncation times (Turnbull, 1976; Lagakos et al., 1988; Wang 1991) or an unconditional approach (Vardi, 1982, 1985; Gill et al., 1988). The latter approach often assumes that the initiation times follow a stationary Poisson process (known as the stationarity assumption), which is satisfied in many practical applications. Under the stationarity assumption, Asgharian et al. (2002) and Asgharian and Wolfson (2005) provided an unconditional nonparametric maximum likelihood estimation (NMLE) for the unbiased survival function in the presence of right censoring, and derived the asymptotic properties of the NMLE under that setting. Of note, a majority of the related literature has focused on one-sample estimation of the unbiased survival distribution given length-biased data. Only limited literature is available that describes modeling risk factors on the distribution of the underlying population. A seminal work by Wang (1996) described a semiparametric proportional hazards model to assess risk factors on length-biased data using a bias-adjusted risk set in constructing the pseudo-likelihood. However, a constraint not allowing data to be right censored had to be imposed due to the dependent censoring mechanism with the observed length-biased data. This subtle feature for informative censoring induced by the sampling scheme has been avoided by not allowing right censoring (Vardi, 1982, 1985; Wang, 1996) or has simply been ignored, as noted by Asgharian and Wolfson (2005).

In this article, we fill the gap by providing semiparametric models (transformation and accelerated failure time models) to examine risk factors on the population failure time via the observed length-biased data, in the presence of informative right censoring. We use the terminology of “length-biased” data for left-truncated and right-censored data under the stationarity assumption. The transformation models and the accelerated failure time models have been extensively investigated for traditional right-censored survival data (e.g., Prentice 1978; Buckley and James 1979; Dabrowska and Doksum, 1988; Ritov 1990; Tsiatis 1990; Lai and Ying 1991; Cheng et al., 1995; Fine, 1999; Lin and Ying 1995; Kalbfleisch and Prentice 2002). However, the literature does not indicate that these models have been used for length-biased data. We estimate the covariate coefficients through constructed unbiased estimating equations, with an inverse weight chosen to adjust the bias induced by the length-biased data and dependent censoring. We present the estimation procedures and the large sample properties of the estimators in Section 2. Under some mild regularity conditions, we show that the resulting estimators are consistent and asymptotically normal. In Section 3, we summarize the simulation

results, and we include an application of the proposed method to the dementia study in Section 4. We offer a discussion in Section 5.

2 Notations and Models

Assume $\tilde{T}$ to be the time measured from the initiating event to failure within the population of interest, A to be the time of examination for the disease (measured from the initiating event), V to be the time measured from examination to failure, and C to be the time from examination to censoring. With length-biased sampling, one can only observe T among those $\tilde{T} > A$, where $T = A + V$ is the observed failure time. Here, A is also referred to as the truncation variable (sampling time or backward recurrence time) and V is the residual survival time (or forward recurrence time). See the following diagram describing the observed dementia data.

Let $Y_i = \min(T_i, A_i + C_i)$, $T_i = A_i + V_i$, $\delta_i = I(V_i \leq C_i)$, and $\mathbf{Z}_i$ be the observed data for the i -th subject, where $\mathbf{Z}_i$ is a vector of covariates for the i -th subject and $i = 1, 2, \dots, n$. It is reasonable to assume that the censoring time measured from the recruitment time, C , and (A, V) are independent given covariate $\mathbf{Z}$.

Recall that the observed failure time data $(T_1, \dots, T_n)$ are a biased subset for population sample $\tilde{T}$. The unbiased density function for $\tilde{T}$, f_U can be related to the length-biased density, g (conditional on $\tilde{T} > A$) under the stationarity assumption,

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

Given covariates, $\mathbf{Z} = \mathbf{z}$, the density of T can be similarly expressed as

$$g(t|\mathbf{z}) = \frac{t f_U(t|\mathbf{z})}{\int_0^\infty u f_U(u|\mathbf{z}) du} := \frac{t f_U(t|\mathbf{z})}{\mu(\mathbf{z})}, \quad (1)$$

where $g(t|\mathbf{z})$ and $f_U(t|\mathbf{z})$ denote the length-biased sampling density and the population (unbiased) density given $\mu(\mathbf{z}) < \infty$.

2.1 Transformation Models

Under the transformation models, it is assumed that the true underlying failure time $\tilde{T}$ is linearly related to the covariates with various specified error distributions (Cheng, Wei and Ying, 1995). Specifically,

$$H(\tilde{T}) = -\mathbf{Z}^T \boldsymbol{\beta} + \epsilon_T,$$

where H is an unknown increasing function, ϵ_T has a known density function and $\boldsymbol{\beta}$ is the vector of regression coefficients. When ϵ_T has an extreme value distribution, $\tilde{T}$ follows the proportional hazards model, whereas when ϵ_T has a logistic distribution, $\tilde{T}$ follows the proportional odds model. It is of interest to assess the covariate coefficients $\boldsymbol{\beta}$ on $\tilde{T}$.

We first consider a special case for the length-biased data without right censoring. For the unbiased variable $\tilde{T}$, we have the following conditional expectation given covariates,

$$\begin{aligned} E \left[I \tilde{T}_i \geq \tilde{T}_j \middle| \mathbf{Z}_i, \mathbf{Z}_j \right] &= P \left[H(\tilde{T}_i) \geq H(\tilde{T}_j) \middle| \mathbf{Z}_i, \mathbf{Z}_j \right] \\ &= P \left(\epsilon_{T_i} - \epsilon_{T_j} \geq \mathbf{Z}_{ij}^T \boldsymbol{\beta} \right) := \xi \left(\mathbf{Z}_{ij}^T \boldsymbol{\beta} \right), \end{aligned}$$

where the data vectors for the i th and j th subjects are independent, $\mathbf{Z}_{ij} = \mathbf{Z}_i - \mathbf{Z}_j$, and

$$\xi \left(\mathbf{Z}_{ij}^T \boldsymbol{\beta} \right) = \int \int I(t \geq s) f_U(t | \mathbf{Z}_i) f_U(s | \mathbf{Z}_j) ds dt. \quad (2)$$

Using an approach similar to that of Cheng et al. (1995), we construct an unbiased estimating function for $\boldsymbol{\beta}$, given the observed length-biased data

$$U_1(\boldsymbol{\beta}) = \sum_{i < j} q(\mathbf{Z}_{ij}) \mathbf{Z}_{ij} \left\{ \frac{I(T_i \geq T_j) - \xi(\mathbf{Z}_{ij}^T \boldsymbol{\beta})}{T_i T_j} \right\},$$

where q is a positive, scalar weight function. The estimating equation is unbiased because

$$E \left[\frac{I(T_i \geq T_j) - \xi(\mathbf{Z}_{ij}^T \boldsymbol{\beta})}{T_i T_j} \middle| \mathbf{Z}_i, \mathbf{Z}_j \right] = \int \int \frac{I(t \geq s) - \xi(\mathbf{Z}_{ij}^T \boldsymbol{\beta})}{ts \mu(\mathbf{Z}_i) \mu(\mathbf{Z}_j)} f_U(t | \mathbf{Z}_i) f_U(s | \mathbf{Z}_j) ds dt = 0$$

by using (1) and (2), when there is no right censoring for the length-biased data.

When the observed failure time T from length-biased sampling is subject to potential right censoring, there are two difficulties. First, the indicators $\{T_i \geq T_j\}$ are not always observable. Second, the censoring variable can be mechanistically dependent on the failure time, because $\text{Cov}(T, A + C) = \text{Cov}(A + V, A + C) = \text{Var}(A) + \text{Cov}(V, A) > 0$ except for trivial cases, (see, e.g. Asgharian, 2003).

The primary purpose of this section is to derive unbiased estimating functions for covariates $\boldsymbol{\beta}$ when the length-biased data are subject to dependent right censoring. Under the stationarity assumption, the joint distribution of (A, V) and (A, T) given $\mathbf{Z}$ follows,

$$f_{A,V}(a, v | \mathbf{Z} = \mathbf{z}) = f_U(a + v | \mathbf{z}) I(a > 0, v > 0) / \mu(\mathbf{z}) \quad (3)$$

as shown in earlier work (Zelen, 2004; Asgharian and Wolfson, 2005). We first assume that the censoring variable C is independent of the covariates. By extending the results of Asgharian and Wolfson (2005), we have the covariate-specific joint probability to observe a failure at y in the length-biased sampling, since the residual censoring time C is assumed to be independent of (A, V) given covariate $\mathbf{Z} = \mathbf{z}$,

$$\Pr(Y \in (y, y + dy), A \in (a, a + da), C \geq y - a | \mathbf{z}) = f_{A,V}(a, y | \mathbf{z}) S_C(y - a) da dy,$$

where $S_C(t)$ is the survival function for the residual censoring variable, C . Replacing the bivariate density $f_{A,V}$ in the above formula by (3), we have

$$\Pr(Y \in (t, t+dt), \delta=1|z) = \int_0^t f_{A,V}(t-v, v|z) S_C(v) dv dt = \frac{f_U(t|z) w(t) dt}{\mu(z)},$$

where $w(t) = \int_0^t S_C(v) dv$. Given the above equations, the following conditional expectation for the observed pairwise length-biased failure times is zero, since

$$\begin{aligned} E \left[\delta_i \delta_j \frac{I(Y_i \geq Y_j) - \xi(Z_{ij}^T \beta)}{w(Y_i)w(Y_j)} \middle| Z_i, Z_j \right] &= \int \int \left[w(t) w(s) f_U(t|Z_i) f_U(s|Z_j) \frac{I(t \geq s) - \xi(Z_{ij}^T \beta)}{w(t)w(s)\mu(Z_i)\mu(Z_j)} \right] dt ds \\ &= \frac{\int \int \{I(t \geq s) - \xi(Z_{ij}^T \beta)\} f_U(t|Z_i) f_U(s|Z_j) ds dt}{\mu(Z_i)\mu(Z_j)} = 0. \end{aligned}$$

The last equation holds by (2). If the censoring distribution is known, the estimating equation can then be constructed as

$$\tilde{U}_T(\beta) = \sum_{i,j} q(Z_{ij}) Z_{ij} \delta_i \delta_j \left[\frac{I(Y_i \geq Y_j) - \xi(Z_{ij}^T \beta)}{w(Y_i)w(Y_j)} \right].$$

When the censoring distribution is unknown, it is natural to replace the unknown censoring distribution by its consistent Kaplan-Meier estimator. Thus, an asymptotic unbiased estimating equation follows

$$U_T(\beta) = \sum_{i,j} q(Z_{ij}) Z_{ij} \delta_i \delta_j \left[\frac{I(Y_i \geq Y_j) - \xi(Z_{ij}^T \beta)}{\hat{w}(Y_i)\hat{w}(Y_j)} \right], \quad (4)$$

where $\hat{w}(t) = \int_0^t \hat{S}_C(v) dv$, and $\hat{S}_C$ is the Kaplan-Meier estimator for censoring variable C . Under the following regularity conditions:

- a. $\xi(\cdot) < \infty$ is a twice continuously differentiable function;
- b. $\mathbf{Z}$ is a $p \times 1$ vector of bounded covariates, not contained in a $(p-1)$ -dimensional hyperplane;
- c. $\sup[t : \Pr(V > t) > 0] \sup[t : \Pr(C > t) > 0] = t_0$, and $\Pr(\delta = 1) > 0$;
- d. $\int_0^{t_0} \left[\left\{ \int_t^{t_0} S_C(u) du \right\}^2 / \{S_C^2(t) S_V(t)\} \right] dS_C(t) < \infty$, where $S_V(t)$ is the survival function for the residual failure time;
- e. $\Gamma_T \equiv \lim_{n \rightarrow \infty} \left\{ n^{-2} \sum_{i < j} q(Z_{ij}) \delta_i \delta_j Z_{ij}^{\otimes 2} \frac{\xi'(Z_{ij}^T \beta)}{\hat{w}(Y_i)\hat{w}(Y_j)} \right\}$ is nonsingular.

The estimating equations (4) have, asymptotically, a unique solution β if the weight $q(\cdot)$ is positive under regularity conditions (a)-(c) and (e). A sketch of the proof is provided in the Appendix. In addition, we have the following weak convergence property for the estimating function.

THEOREM 1. Under regularity conditions (a)-(d),

$$n^{-3/2}U_T(\beta_0) \rightarrow N(0, \Sigma_T),$$

in distribution as $n \rightarrow \infty$, where t_0 is a finite time point defined in (c).

The derivation of the weak convergence of $n^{-3/2}U_T(\beta_0)$ is obtained by the properties of U-statistics and the asymptotic martingale integral representation of $\sqrt{n}(\widehat{w}(t) - w(t))$ (Pepe and Fleming, 1989, 1991). Because summation of the pairwise indicators in $U_T(\beta)$ is over $\binom{n}{2} = O(n^2)$ items, the rate of convergence is in the order of $n^{-3/2}$. Based on the above distributional properties, the variance-covariance of $n^{-3/2}U_T(\beta_0)$, Σ_T can be consistently estimated by the empirical form of the U-statistic variance

$$\widehat{\Sigma}_T = \frac{1}{n^3} \sum_{i < j < k} \left\{ \widehat{\eta}_{ij}(\beta) \widehat{\eta}_{ik}^T(\beta) + \widehat{\eta}_{ik}(\beta) \widehat{\eta}_{ij}^T(\beta) + \widehat{\eta}_{ik}(\beta) \widehat{\eta}_{jk}^T(\beta) + \widehat{\eta}_{jk}(\beta) \widehat{\eta}_{ik}^T(\beta) \right\},$$

where

$$\begin{aligned} \widehat{\eta}_{ij}(\beta_0) &= q(Z_{ij}) Z_{ij} a_{ij}(\beta) + \int_0^\infty \frac{\widehat{b}(t)}{\widehat{\pi}(t)} [d\widehat{M}_i(t) + d\widehat{M}_j(t)], \\ \widehat{M}_i(t) &= I(Y_i - A_i \leq t, \Delta_i = 0) - \int_0^t I(Y_i - A_i \geq u) d\widehat{\Lambda}_C(u), \end{aligned}$$

and $\widehat{\Lambda}_C(t)$ is the Nelson estimate for the cumulative hazards function of the censoring variable C , $\widehat{\pi}(t) = \sum_{i=1}^n I(Y_i - A_i \geq t) / n$,

$$\widehat{b}(t) = \lim_{n \rightarrow \infty} \frac{1}{n^2} \sum_{i < j} q(Z_{ij}) Z_{ij} \widehat{a}_{ij}(\beta) \{ \widehat{h}_i(t) + \widehat{h}_j(t) \},$$

$$\widehat{a}_{ij}(\beta) = \frac{\delta_i \delta_j [I(Y_i \geq Y_j) - \xi(Z_{ij}^T \beta)]}{\widehat{w}(Y_i) \widehat{w}(Y_j)}.$$

Note that regularity condition (c) ensures that the support for the censoring variable C be shorter than the support of V , so that $\widehat{b}(t) / \widehat{\pi}(t)$ is well defined for any $t \in [0, t_0]$. Most of observational studies and clinical trials have limited follow-up, thus this condition can be easily satisfied.

By the Taylor series expansion of $U_T(\beta)$ around β_0 , $\sqrt{n}(\widehat{\beta} - \beta_0)$ is asymptotically equivalent to $n^{-3/2} \Gamma_T^{-1} U_T(\beta_0)$ by the Slutsky theorem and the delta method. Thus, under the regularity conditions (a)-(e), the distribution of $\sqrt{n}(\widehat{\beta} - \beta_0)$ is approximated by a normal distribution as

$$\sqrt{n}(\widehat{\beta} - \beta_0) \rightarrow N(0, \Gamma_T^{-1} \Sigma_T \Gamma_T^{-1}),$$

where Γ_T is the Hessian matrix of the estimating function (4) and can be consistently estimated by

$$\widehat{\Gamma}_T = \left\{ n^{-2} \sum_{i < j} q(Z_{ij}) \delta_i \delta_j Z_{ij}^{\otimes 2} \frac{\xi'(Z_{ij}^T \widehat{\beta})}{\widehat{w}(Y_i) \widehat{w}(Y_j)} \right\}.$$

Provided that the covariate vectors are not all on a hyperplane, the identifiability of β is guaranteed given conditions (c) and (e).

2.2 Accelerated Failure Time Models

The accelerated failure time (AFT) model relates the logarithm of the failure time linearly to the covariates (Cox and Oakes 1984; Kalbfleisch and Prentice 2002),

$$\log \tilde{T} = Z^T \alpha + \epsilon_A, \quad (5)$$

where ϵ_A has a unknown distribution function with mean zero. In contrast to the transformation models, the distribution of ϵ_A in the AFT model is not specified other than the mean zero requirement, but the transformation to $\tilde{T}$ is specified. Under the AFT model, the effect of the covariates on the failure time is direct acceleration or deceleration of the time to failure, which is easier to interpret and more relevant to clinicians.

Because of the dependent right censoring induced from length-biased sampling, unconventional approaches to adjust for the non-informative censoring are required under the AFT models. We assume that the censoring time is independent of the covariates. Based on the joint distribution of (A, Y) and C conditional on covariates Z , we can derive the expectation of the following quantity to be zero,

$$\begin{aligned} E \left\{ \frac{\delta(\log Y - Z^T \alpha)}{w(Y)} \middle| Z = z \right\} &= E \left\{ \int_0^\infty \int_0^\infty P(Y=y, A=a, \delta=1 | Z=z) \frac{(\log y - z^T \alpha)}{w(y)} dady \right\} \\ &= E \left\{ \frac{1}{\mu(z)} \int_0^\infty \int_0^y f_U(y | Z=z) S_C(y-a) \frac{(\log y - z^T \alpha)}{w(y)} dady \right\} \\ &= E \left\{ \frac{1}{\mu(z)} \int_0^\infty f_U(y | Z) (\log y - z^T \alpha) dy \right\} = 0. \end{aligned}$$

Accordingly, an unbiased estimating equation for α can be constructed as

$$\tilde{U}_A(\alpha) = \sum_{i=1}^n q(Z_i) \delta_i Z_i \frac{(\log Y_i - Z_i^T \alpha)}{w(Y_i)} = 0, \quad (6)$$

where q is a positive, scalar weight function. By replacing the unknown quantity, $w(Y_i)$, by its consistent estimator $\hat{w}(Y_i)$, an asymptotic unbiased estimating equation follows:

$$U_A(\alpha) = \sum_{i=1}^n q(Z_i) \delta_i Z_i \frac{(\log Y_i - Z_i^T \alpha)}{\hat{w}(Y_i)} = 0. \quad (7)$$

Of note, the above estimating equation has a closed-form solution for α ,

$$\widehat{\alpha} = \left\{ \sum_{i=1}^n \frac{q(Z_i) \delta_i Z_i Z_i^T}{\widehat{w}(Y_i)} \right\}^{-1} \sum_{i=1}^n \frac{q(Z_i) \delta_i Z_i \log Y_i}{\widehat{w}(Y_i)}.$$

Let α_0 be the true value of the regression coefficient vector. We impose the following regularity conditions besides conditions (b)-(d) for a rigorous justification of the asymptotic properties of $\widehat{\alpha}$:

- f** $\det \left(E \left[\left\{ \delta Z \left(\log Y - Z^T \alpha_0 \right) \right\} / \{w(Y)\} \right]^{\otimes 2} \right) < \infty$;
- g** $\det \left(\int_0^{\tau_0} D^{\otimes 2}(s) / \{S_c^2(s)\} dS_c(s) \right) < \infty$, where
- $$D(t) = E \left[\left\{ q(Z) \delta Z I(Y \geq s) \int_t^Y S_c(u) du \left(\log Y - Z^T \alpha_0 \right) \right\} / \{w^2(Y)\} \right];$$
- h** $\Gamma_A \equiv -\lim_{n \rightarrow \infty} \left\{ \frac{1}{n} \sum_{i=1}^n \frac{q(Z_i) \delta_i Z_i Z_i^{\otimes 2}}{\widehat{w}(Y_i)} \right\}$ is nonsingular. We show the consistency of $\widehat{\alpha}$ in the supplementary technical report by an argument similar to that for the consistency of $\widehat{\beta}$ under the transformation models. The asymptotic normality of $\widehat{\alpha}$ is stated next.

THEOREM 2. Under regularity conditions (b)-(d), and (f)-(h), $\sqrt{n}(\widehat{\alpha} - \alpha_0)$ converges weakly to a normal distribution with mean zero and variance-covariance matrix $\Gamma_A^{-1} \Sigma_A \Gamma_A^{-1}$, in which Γ_A is the Hessian matrix of the $U_A(\alpha_0)$ and Σ_A is the asymptotic variance-covariance matrix of $n^{-1/2}U_A(\alpha_0)$.

A detailed proof for Theorem 2 is provided in the Appendix. Essentially, by Taylor series expansion,

$$n^{-1/2}U_A(\widehat{\alpha}) = n^{-1/2}U_A(\alpha_0) - \frac{1}{n}\Gamma_n(\alpha_0) \sqrt{n}(\widehat{\alpha} - \alpha_0) + o_p(1),$$

where $\Gamma_n(\alpha_0)$ is the first derivative of $U_A(\alpha_0)$. By the asymptotic martingale expression of $\widehat{w}(y)$, we show the asymptotic i.i.d. representation of the difference between $n^{-1/2}U_A(\alpha_0)$ and $n^{-1/2}\widetilde{U}_A(\alpha_0)$,

$$n^{-1/2}U_A(\alpha_0) - n^{-1/2}\widetilde{U}_A(\alpha_0) = n^{-1/2} \sum_{i=1}^n \int_0^\infty \frac{D(s)}{S_c(s) S_v(s)} dM_i(s) + o_p(1).$$

This representation, together with the consistency of $\frac{1}{n}\Gamma_n(\alpha_0)$ and an application of the classic Central Limit and Slutsky's Theorem, implies that $\sqrt{n}(\widehat{\alpha} - \alpha_0)$ converges in distribution to a zero-mean normal distribution.

The Hessian matrix Γ_A and the variance-covariance matrix Σ_A can be consistently estimated by

$$\widehat{\Gamma}_A = -\frac{1}{n} \sum_{i=1}^n \frac{q(\mathbf{Z}_i) \delta_i \mathbf{Z}_i^{\otimes 2}}{\widehat{w}(Y_i)},$$

$$\widehat{\Sigma}_A = \frac{1}{n} \sum_{i=1}^n \left\{ q(\mathbf{Z}_i) \delta_i \mathbf{Z}_i \frac{(\log Y_i - \mathbf{Z}_i^T \widehat{\alpha})}{\widehat{w}(Y_i)} + \int_0^\infty \frac{\widehat{D}(t) d\widehat{M}_i(t)}{\sum_{i=1}^n I(Y_i - A_i \geq t)/n} \right\}^{\otimes 2},$$

respectively, where

$$\widehat{D}(t) = \frac{1}{n} \sum_{i=1}^n q(\mathbf{Z}_i) I(t \leq Y_i) \delta_i \mathbf{Z}_i \int_t^{Y_i} \widehat{S}_c(u) du \frac{(\log Y_i - \mathbf{Z}_i^T \widehat{\alpha})}{\widehat{w}^2(Y_i)},$$

$$\widehat{M}_i(t) = I(Y_i \leq t, \Delta_i = 0) - \int_0^t I(Y_i \geq u) d\widehat{\Lambda}_c(u),$$

and $\widehat{\Lambda}_c(u)$ is the Nelson-Aalen estimator for the cumulative hazard function of the censoring times.

2.3 Remarks

We have restricted our attention to the setting in which the censoring distribution is independent of the covariates for both model structures. However, it is not conceptually difficult to generalize derivations to the setting with a covariate-dependent censoring distribution. Without loss of generality, we assume that the covariate $\mathbf{Z}$ can be discretized to a finite number of possible values for the covariate-specific censoring distribution if the censoring variable is not independent of $\mathbf{Z}$. Taking the transformation model as an example, when sample size is large enough within each subgroup according to discretized covariates, the estimating equations $U_T(\beta)$ under the transformation models can be extended accordingly for a covariate-specific censoring distribution as follows,

$$U_T^*(\beta) = \sum_{i,j} q(\mathbf{Z}_{ij}) \mathbf{Z}_{ij} \delta_i \delta_j \left[\frac{I(Y_i \geq Y_j) - \xi(\mathbf{Z}_{ij}^T \beta)}{\widehat{w}(Y_i | \mathbf{Z}_i) \widehat{w}(Y_j | \mathbf{Z}_j)} \right], \quad (8)$$

where $\widehat{w}(t | \mathbf{Z}_i) = \int_0^t \widehat{S}_c(u | \mathbf{Z}_i) du$, and $\widehat{S}_c(u | \mathbf{Z}_i)$ is the Kaplan-Meier estimator for the censoring variable C given pairs $(Y_i - A_i, \delta_i)$ for which $\mathbf{Z}_i = \mathbf{Z}(i = 1, \dots, n)$. Note that covariate $\mathbf{Z}$ does not have to be discretized in modeling the failure time. We may also fit a semiparametric (e.g. Cox model, accelerated failure time model) or parametric model to take into consideration the covariate dependence of the censoring variable, and plug covariate-specific censoring distribution $S_c(t | \mathbf{Z})$ into the estimating equation (8) for the transformation models or into the estimating equation (7) for the AFT models. Alternatively, one may consider the profile likelihood approach for length-biased data, in which the distribution of censoring variable C is not required to be estimated. Further research is required to study the tradeoff between robustness and efficiency of various alternative methods; but this is beyond the scope of this paper.

Another note is that either the transformation model assumption or the accelerated failure time model assumption is not invariant for population data and length-biased data in general. For example, the proportional hazards model assumption for the population samples would not lead to the same model assumption for the length-biased samples. Unless evaluating trivial cases for which the risk factors have no impact on the failure times, the model assumptions may be satisfied using both the population sample and the length-biased sample. Another

special parametric case is when $H(\cdot)$ is of a log-transformation and ε follows the extreme value distribution, the covariate-specific density functions for T and T both follow the Gamma distribution, but with different parameters.

3 Simulations

3.1 Data Generation

Generating length-biased failure time data is often difficult from a renewal process. Using the following approach, we can simplify the data generating procedure to obtain right-censored, length-biased data under the semiparametric models. We first generated independent pairs of (A_i, T_i) from $A \sim U(0, \tau)$ and $T \sim f_U(t|z)$, and we kept the pairs with $A_i < \tilde{T}_i$, which are the length-biased samples, denoted by $T_i = A_i + V_i$. For simplicity of notation, we use $f_U(t)$ instead of $f_U(t|z)$ in this section. Thus, the length-biased sample follows the conditional distribution,

$$\Pr(T=t|A<T) = \Pr(T=t, A<T) / \Pr(A<T)$$

where $\Pr(T=t, A < T) = \Pr(T=t, A < t) = f_U(t)t/\tau$, for $t < \tau$, $f_U(t)$ for $t \geq \tau$, and

$$\Pr(A < T) = \int_0^\tau \Pr(T=t, A < t) dt = \int_0^\tau f_U(t) t dt / \tau + \int_\tau^\infty f_U(t) dt.$$

Thus,

$$\Pr(T=t|A<T) = \begin{cases} \frac{tf_U(t)}{\int_0^\tau f_U(t)tdt + \tau \int_\tau^\infty f_U(t)dt} & t < \tau \\ \frac{\tau f_U(t)}{\int_0^\tau f_U(t)tdt + \tau \int_\tau^\infty f_U(t)dt} & t \geq \tau. \end{cases}$$

To obtain length-biased data, we should choose τ larger than the upper bound of T , so that $P(T=t|A < T) = 0$ for $t > \tau$. For example, if T is the length-biased outcome for age of death, then we shall let $\tau > 100$ years. Under both proportional hazards and proportional odds models with a baseline Gamma survival function (i.e. $H(t) = 2 - \log(t)$), $\tau \int_\tau^\infty f_U(t) dt = \tau(1 - F_U(\tau))$ approximates to zero if $\tau \rightarrow \infty$. Therefore, while the probability mass at tail ($t \geq \tau$) is negligible, for $t < \tau$

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

The censoring variables measured from the examination time (C_i) are generated from uniform distributions for various censoring percentages. The censoring indicator is obtained by $\delta_i = I(T_i \leq C_i + A_i)$. We considered regression models with two independent covariates. Z_1 is generated from a Bernoulli variable and Z_2 from a uniform variable on $(0,1)$. Each scenario comprised 1000 runs. Cohort sizes of 200 or 400 were used.

3.2 Transformation Models

First, we considered the underlying population distribution of $\tilde{T}$ to follow a proportional odds model. The average bias, 95% coverage probability, and mean squared errors (MSEs) are summarized in Table 1. We found that the biases decreased with the increase of cohort size, and the MSEs slightly increased with the percentage of censoring. In general, the coverage

probabilities were reasonably close to the nominal level but tended to be slightly larger than .95, because the mean standard errors obtained from $\widehat{\Gamma}_T^{-1}\widehat{\Sigma}_T\widehat{\Gamma}_T^{-1}$ were somewhat overestimated for the empirical ones, especially with small right censoring percentage.

We generated unbiased data from another important special case, the proportional hazards model. We also assumed that the baseline cumulative hazard function was t^2 , so that the covariate specific survival function followed a Weibull distribution. We compared the estimators obtained from the proposed method with those from the naive Cox proportional hazards model when ignoring length-biased sampling. The results are summarized in Table 2. When $\beta = (0, 0)$, the estimators from the proposed method and the usual Cox model are comparable and virtually unbiased with appropriate coverage probabilities. Under the null hypothesis, it is clear that the length-biased samples and the population samples both follow the proportional hazards model. In contrast, when $\beta \neq 0$, the length-biased samples will not follow the proportional hazards model, if the population samples follow the proportional hazards model. Therefore, the estimators obtained from the naive Cox model can be substantially biased (overestimated in this case) compared with those from the proposed method. We found that the 95% coverage probabilities varied from 28% to 84%.

3.3 AFT Models

Next we evaluated the performance of the proposed estimators for the AFT models and compared them with the estimators from the following naive estimating equation ignoring length-biased sampling,

$$U_{A2}(\alpha) = \sum_{i=1}^n \delta_i Z_i \frac{(\log Y_i - Z_i^T \alpha)}{\widehat{S}_c(Y_i)}.$$

Note that U_{A2} is an asymptotic unbiased estimating equation for traditional survival data under the AFT models. Here, the distribution of the random error, ε_A , was assumed to be Uniform $(-0.5, 0.5)$. Table 3 summarizes the performance of estimators with and without adjusting the length-biased sampling. The proposed estimators achieved outstanding accuracy: the empirical biases were less than 1% and the empirical coverage probabilities of the 95% confidence intervals were quite consistent with the nominal level for the scenarios investigated.

By the naive estimating equation U_{A2} , the estimators for the intercept were substantially biased (overestimated) and the associated inference were inaccurate, especially when censoring percentage was larger than 10%. For settings with null covariate effects, both estimation procedures led to virtually unbiased estimators for covariate coefficients with appropriate coverage probabilities. In the presence of non-zero covariate effects, the average biases of the estimated covariate effects by the estimating equation U_{A2} increased and the associated coverage probabilities decreased with increasing degree of censoring.

3.4 Censoring Times Dependent on Covariate

We have also performed additional simulation studies to evaluate the robustness of the estimating equations to mis-specification of the censoring model used. We simulated survival times from the model described in Table 2 (proportional hazards) and Table 3 (AFT), but the censoring times were dependent on one of the two covariates, Z_2 . The choice of covariate coefficient for Z_2 , γ reflected the degree of dependence between the censoring time and the covariate. In the lower panels of Table 2 and Table 3, we list both biases and empirical coverage probabilities, while ignoring the censoring dependent on Z_2 in the corresponding estimating equations. With a total sample size of 200, the results show that the estimation procedures are

reasonably robust to the dependent censoring in general, though the biases slightly increase when the dependence to Z_2 gets stronger. Thus, if there is no strong indication for dependent censoring, we may use the overall nonparametric estimator for the censoring distribution, as shown by our empirical study.

4 A Real Data Example

To illustrate the proposed methods, we applied our methods to analyze a large prevalent cohort study with time to death data for subjects diagnosed with dementia. Dementia is a progressive degenerative medical condition, and is one of the leading causes of all deaths in the United States and Canada. The Canadian Study of Health and Aging (CSHA) study was a multicenter epidemiologic study, in which more than 14,000 subjects who were 65 years or older were randomly chosen to receive an invitation for a health survey throughout Canada, and a total of 10,263 subjects agreed to participate (Wolfson et al, 2001). The participants were then screened for dementia starting in 1991, and 1132 people were identified as having the disease.

All dementia patients had been followed until 1996 and their dates of dementia onset were ascertained from their medical records, and their dates of death or last follow-up were recorded prospectively from the time of screening. The relevant data included three dementia categories (probable Alzheimer's disease, possible Alzheimer's disease and vascular dementia), approximate date of onset, date of screening for dementia, date of death or censoring and death indicator variable. Excluding subjects with missing date of disease onset or missing classification of dementia sub-type, there were a total of 818 patients remained. We used data collected as part of the CSHA to examine if the survival distributions were different among three sub-types of dementia diagnosis. Given prevalent cases ascertained cross-sectionally, the stationarity assumption, which is defined as that the incidence of dementia did not change over the period of the study, was validated in Addona and Wolfson (2006). It is clear that patients with worse prognosis of dementia were more likely to die before the study recruitment, therefore the remaining cohort of patients was a collection of length-biased samples.

Among them, 393 patients had probable Alzheimer's disease, 252 had possible Alzheimer's disease and 173 had vascular dementia. At the end of this study, 638 out of 818 patients died and the others were right censored. Using the type of probable Alzheimer's disease as the baseline, we defined two indicator variables for the other two subtypes of dementias. We assess whether the subtype of dementia at diagnosis affects time to death. Since there was no evidence of dependence between the censoring time measured from the study recruitment and the subtype of dementias, it is appropriate to use the estimating equation (4) or equation (7) to make inferences about regression coefficients. Moreover, given the sampling mechanism for this study the residual censoring time is independent of the residual survival time, and the truncation time A .

We analyzed the above data set using the proportional hazards model, the proportional odds model, and the accelerated failure time model. Applying the proposed methods for length-biased data, the estimated covariate effects of two subtypes of dementias and their standard errors are listed in Table 4. Under the proportional hazards model, the estimated covariate coefficient (s.e) is .241 (.211) for vascular dementia at diagnosis, and -.110 (.201) for possible Alzheimer's disease relative to the probable Alzheimer's disease (with $q = 1$). The results under the proportional odds model and the accelerated failure time model with length-biased adjustment lead to conclusions similar to those obtained under the proportional hazards model. The data show that the diagnosis of three subtypes of dementia had little difference in long-term survival using the methods that adjusted for length-biased sampling, which was consistent with the nonparametric survival estimators provided in Wolfson et al (2001).

We also analyzed the the same data set using the estimating equations proposed in Cheng et al (1995) under the proportional hazards and proportional odds models and the estimating equation U_{A2} under the accelerated failure time model for traditional right-censored data (without considering length-biased sampling). And the results are listed in Table 4 under the column of “Cheng” or “Eq. U_{A2} ”. Similar to the results from the simulation studies, under the proportional hazards and proportional odds models, the covariate effects could be overestimated using the “naive” approaches under the same model assumptions (Cheng et al., 1995), when the length-biased sampling is ignored. Comparing the results from the naive methods, we found a weaker relationship between subtype of dementias at diagnosis and the elapsed time to death under the bias-adjusted models (smaller Z-values), though all methods and models did not yield statistically significant difference among the subtypes of dementia. Under the accelerated failure time model, the difference between the estimated intercepts with and without length-biased adjustment indicates that the use of an approach ignoring the length-biased sampling may lead to a substantial overall underestimation of the deleterious effects of dementia. Specifically, when ignoring the length-biased sampling, the estimated median survival duration was 8.88 years (95% CI: 8.43-9.34) for probable Alzheimer's disease, 9.00 years (95% CI: 8.13-9.56) for vascular dementia, and 9.40 years (95% CI: 8.75-10.09) for possible Alzheimer's disease. When adjusting for length-biased sampling, the estimated median survival duration was 3.71 years (95% CI: 2.95-4.67) for probable Alzheimer's disease, 3.44 years (95% CI: 3.02-3.92) for vascular dementia, and 4.22 years (95% CI: 3.49-5.10) for possible Alzheimer's Disease.

5 Discussion

We have proposed the inference methods to evaluate covariate effects on time-to-event data with right censoring under the semiparametric transformation models and the semiparametric accelerated failure time models, when the observed data are subject to length-biased sampling. Under both models, we impose the semiparametric model structure on the population sample, which is often of primary interest, instead of on the length-biased sample. The interpretation of the regression coefficients in the population model is straightforward. Under the transformation models, one can take advantage of the stochastic ordering preserved for length-biased times and unbiased times, thereby construct the estimating equations using observed length-biased times. The estimation procedures can be easily implemented and applied for various models under the class of transformation models, which include important special cases such as proportional hazards and proportional odds models. In contrast to the pairwise comparison of the data in the estimating equations under the transformation models, the proposed estimating equation approach under the AFT model is based on the least-squares principle. It is appealing for practical applications because of the elegant closed form of the solution for the estimator of α , which can be obtained easily without using an iterative numerical algorithm.

When directly using the Kaplan-Meier estimator, $\hat{S}_C(t)$, as an inverse weight, the stability of the weight function at the tail is of practical concern. Our weight function, in contrast, is much more robust especially regarding its behavior at the tail, because the weight is an inverse of the integral of the Kaplan-Meier estimator, $\hat{w}(t) = \int_0^t \hat{S}_C(u) du$. The potential instability for the estimating equation caused by $S_C(t) \rightarrow 0$ under heavy censoring is moderated by integrating the estimated survival distribution over the observed failure times, because the area under the survival curve, $\int_0^t S_C(x) dx$, would not approximate to zero under either heavy or light censoring. Some theoretical comparison of the asymptotic efficiency between estimating equations using $\hat{S}_C(t)$ and $\hat{w}(t)$ as inverse weights is provided in the supplementary technical report. One common challenge in the use of Kaplan-Meier estimator is the undefined nonparametric estimator beyond the largest observed censoring time. One possible solution is

to use a parametric model fitted with the data up to the largest observed censoring time, and then to extrapolate the curve for the censoring distribution. However, from our empirical studies, the impact of using the imperfect survival distribution estimator (extrapolated from the last observed censoring time) seems to be limited for the moderately and heavily censored C .

Theoretically, we can find the optimal weight function q in the estimating equations to achieve the maximum efficiency for the estimation of regression coefficients. Because the optimal weight function consists of the true value of the regression coefficient vector, the use of such an optimal weight is less feasible. As in the literature (e.g., Cheng et al 1995; Fine 1999), $q = 1$ is a practical and reasonable choice in applications. An alternative approach is to use an initial consistent estimator of regression coefficients for the optimal weight function and calculate a single Newton-like update of the solution to the estimating equation (Gray, 2000).

One potential challenge with the proposed semiparametric models is the inclusion of time-varying covariates for length-biased data, similar to that for traditional survival data in Cheng et al (1995). It is possible to extend the maximum likelihood approaches of Zeng and Lin (2007a, 2007b), which allow time varying covariates under the transformation models or the AFT models for traditional survival data, to length-biased data. It is of interest to estimate the distribution of the baseline function, thus predicting unbiased covariate-specific survival distribution from the linear transformation model with length-biased data. However, a substantial research effort would be required to establish the large sample properties and computational algorithms in this more complicated setting. Specifically, modern empirical process approaches are required to prove the tightness of the estimated quantities involving t . It is our intention in future research to develop methods for the purpose of prediction.

The issues of biased sampling discussed in this work extend beyond length-biased data. The importance of correctly modeling the risk factors in a cohort with *selection bias* has been discussed in recent literature (Begg, 2002), in which Begg pointed out the over-representation of risk factors such as genetic abnormalities for breast cancer in population-based case proband studies, because adjustments are not made under the size-biased sampling scheme. Although this paper and Begg's work share a common concern regarding the potential bias in estimating covariate effects using naive inference methods under biased sampling (length-biased sampling is a special type of biased sampling mechanism), our data structure with time-to-event outcome is different from the one with a binary outcome under *size-biased* sampling discussed by Begg (2002), and the proposed method for length-biased data with right-censoring may not be directly applicable to estimation with the size-biased sampling schemes. New method for correcting the bias caused by size-biased sampling is needed.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported in part by the U.S. National Institutes of Health. We thank the editor, the associate editor and three referees for their constructive suggestions that improved the article. We also thank Professor Masoud Asgharian and investigators of the Canadian Study of Health and Aging (CHSA) for providing us the dementia data from CHSA. The data reported in the example were collected as part of the CHSA. The core study was funded by the Seniors' Independence Research Program, through the National Health Research and Development Program 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

Consistency of $\widehat{\beta}$

Under regularity conditions (a)-(c) and (e), similar to the work of Cheng et al. (1995), we can prove there is a unique solution to the equation $U_T(\beta) = 0$ if the weights in $U_T(\beta)$ are positive. With probability one, the quantity, $n^{-2}U_T^T(\beta)(\beta - \beta_0)$, converges to

$$\int_{z_1, z_2} q(z_{12}) (z_{12}^T \beta - z_{12}^T \beta_0) \frac{\{\xi(z_{12}^T \beta_0) - \xi(z_{12}^T \beta)\}}{\mu(z_1) \mu(z_2)} dF(z_1) dF(z_2),$$

where F is the distribution function of $\mathbf{Z}$ and $\mathbf{Z}_{12} = \mathbf{Z}_1 - \mathbf{Z}_2$. The above quantity is nonnegative, because $\zeta(\cdot)$ is a decreasing function. Thus, $\widehat{\beta}$ is consistent, as long as the weight, $q(\cdot)$ is positive.

Asymptotic distribution of $U_T(\beta_0)$

We derive the asymptotic properties of $U_T(\beta_0)$ when the censoring distribution is independent of the covariate vector $\mathbf{Z}$. The uniform consistency of the Kaplan-Meier estimator, $\widehat{S}_C(t)$, over $[0, t_0]$ (e.g., Gill 1982) entails the uniform consistency of $\widehat{w}(t)$, i.e.,

$$\sup_{t \leq t_0} |\widehat{w}(t) - w(t)| \rightarrow 0,$$

because $\sup_{t \leq t_0} |\int_0^t (\widehat{S}_C(u) - S_C(u)) du| \leq \sup_{t \leq t_0} |\int_0^t \widehat{S}_C(u) - S_C(u) du| \rightarrow 0$ in probability given condition (c). Recall that

$$\frac{U_T(\beta_0)}{n^{3/2}} = \frac{1}{n^{3/2}} \sum_{i,j} q(\mathbf{Z}_{ij}) \mathbf{Z}_{ij} \delta_i \delta_j \left[\frac{I(Y_j \geq Y_i) - \xi(\mathbf{Z}_{ij}^T \beta_0)}{\widehat{w}(Y_i) \widehat{w}(Y_j)} \right].$$

Note that condition $Pr(\delta = 1) > 0$ assures the existence of the estimating equation $U_T(\beta)$ in (2.4). Using arguments similar to those in Cheng et al (1995) or Fine (1999), together with the uniform consistency of $\widehat{w}(t)$ to $w(t)$, it is easy to show that

$$\frac{U_T(\beta_0)}{n^{3/2}} = \frac{1}{n^{3/2}} \sum_{i,j} q(\mathbf{Z}_{ij}) \mathbf{Z}_{ij} \delta_i \delta_j \left[\frac{I(Y_j \geq Y_i) - \xi(\mathbf{Z}_{ij}^T \beta_0)}{w(Y_i) w(Y_j)} \right] \times \left\{ \frac{w(Y_i) - \widehat{w}(Y_i)}{w(Y_i)} + \frac{w(Y_j) - \widehat{w}(Y_j)}{w(Y_j)} + 1 \right\} + o_p(1). \quad (9)$$

Following from a martingale integral representation for $\sqrt{n}(\widehat{w}(t) - w(t))$ by Pepe and Fleming (1989, 1991), we can re-express $\sqrt{n}(\widehat{w}(t) - w(t))$ as a martingale integral via integration by parts

$$\begin{aligned} \sqrt{n}(w(Y_i) - \widehat{w}(Y_i)) &= n^{-1/2} \sum_{k=1}^n \int_0^{Y_i} \left[\int_t^{Y_i} S_C(u) du \right] \frac{dM_k(t)}{\pi(t)} + o_p(1) \\ \sqrt{n} \frac{w(Y_i) - \widehat{w}(Y_i)}{w(Y_i)} &= n^{-1/2} \sum_{k=1}^n \int_0^{Y_i} \frac{h_i(t)}{\pi(t)} dM_k(t) + o_p(1), \end{aligned} \quad (10)$$

where $h_i(t) = I(t \leq Y_i) \left[\int_t^{Y_i} S_c(u) du \right] / w(Y_i)$,
 $\pi(t) = S_c(t) S_v(t)$, $M_k(t) = I(Y_k - A_k \leq t, \Delta_k = 0) - \int_0^t I(Y_k - A_k \leq u) d\Lambda_c(u)$ is the martingale for the residual censoring variable, and $\Lambda_c(u)$ is the corresponding cumulative hazard function.

To simplify the notation, let

$$a_{ij}(\beta_0) = \frac{\delta_i \delta_j [I(Y_i \geq Y_j) - \xi(Z_{ij}^T \beta_0)]}{w(Y_i) w(Y_j)},$$

$$b(t) = \lim_{n \rightarrow \infty} \frac{1}{n^2} \sum_{i < j}^n q(Z_{ij}) Z_{ij} a_{ij}(\beta_0) \{h_i(t) + h_j(t)\}.$$

By inserting (10), $a_{ij}(\beta_0)$, and $b(t)$ into (9), we see that $n^{-3/2} U_T(\beta_0)$ can be asymptotically decomposed to a U-statistic $\tilde{U}(\beta_0)$ and a martingale integral $V(\beta_0)$,

$$\begin{aligned} n^{-3/2} U_T(\beta_0) &= n^{-3/2} \sum_{i < j} q(Z_{ij}) Z_{ij} a_{ij}(\beta_0) + n^{-1/2} \sum_{k=1}^n \int_0^\infty \frac{b(t)}{\pi(t)} dM_k(t) + o_p(1) \\ &\equiv \tilde{U}(\beta_0) + V(\beta_0) + o_p(1), \end{aligned}$$

where

$$\tilde{U}(\beta_0) = \sum_{i,j} q(Z_{ij}) Z_{ij} \delta_i \delta_j \left[\frac{I(Y_i \geq Y_j) - \xi(Z_{ij}^T \beta_0)}{w(Y_i) w(Y_j)} \right]$$

and $V(\beta_0) = n^{-1/2} \sum_{k=1}^n \int_0^\infty \frac{b(t)}{\pi(t)} dM_k(t)$. Note that the martingale integral $V(\beta_0)$ can also be written in the form of a U-statistic, so that

$$n^{-3/2} U_T(\beta_0) = n^{-3/2} \sum_{i < j} \left\{ q(Z_{ij}) Z_{ij} a_{ij}(\beta_0) + \int_0^\infty \frac{b(t)}{\pi(t)} [dM_i(t) + dM_j(t)] \right\} + o_p(1).$$

Given the above multivariate U-statistic representation, weak convergence of $n^{-3/2} U_T(\beta_0)$ follows from the standard distribution theory (e.g., Wei and Johnson 1985; Serfling 1980, pp. 192). Both $\tilde{U}(\beta_0)$ and $V(\beta_0)$ are naturally tight because they do not involve t . The asymptotic distribution of $U_T(\beta_0)$ with discrete covariates can be derived similarly.

Asymptotic distribution of $\hat{\alpha}$

By Taylor series expansion,

$$\frac{1}{\sqrt{n}} U_A(\hat{\alpha}) = \frac{1}{\sqrt{n}} U_A(\alpha_0) - \frac{1}{n} \Gamma_n(\alpha_0) \sqrt{n}(\hat{\alpha} - \alpha_0) + o_p(1),$$

where $\Gamma_n(\alpha_0)$ is the first derivative of $U_A(\alpha_0)$ and $\frac{1}{n} \Gamma_n(\alpha_0)$ converges in probability to the Hessian matrix of the $U_A(\alpha_0)$, Γ_A . By the uniform consistency of $\hat{w}(t)$ to $w(t)$, it is easy to show

$$\frac{1}{\sqrt{n}}U_A(\alpha_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n q(Z_i) \delta_i Z_i \frac{(\log Y_i - Z_i^T \alpha_0)}{w(Y_i)} \left\{ 1 + \frac{w(Y_i) - \widehat{w}(Y_i)}{w(Y_i)} \right\} + o_p(1). \quad (11)$$

Define $D(t) = E \left[\left\{ q(Z) \delta Z I(Y \geq s) \int_t^Y S_c(u) du (\log Y - Z^T \alpha_0) \right\} / \{w^2(Y)\} \right]$. Then by inserting (10) into (11), we have

$$\frac{1}{\sqrt{n}}U_A(\alpha_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \left\{ q(Z_i) \delta_i Z_i \frac{(\log Y_i - Z_i^T \alpha_0)}{w(Y_i)} + \int_0^\infty \frac{D(t) dM_i(t)}{\pi(t)} \right\} + o_p(1).$$

Hence, under regularity conditions (c)-(e), $n^{-1/2}U_A(\alpha_0)$ is asymptotically normally distributed by the Central Limit Theorem. This, combined with an application of Slutsky's theorem, implies that $\sqrt{n}(\widehat{\alpha} - \alpha_0)$ converges weakly to a normal distribution with mean zero and variance-covariance matrix $\Gamma_A^{-1} \Sigma_A \Gamma_A^{-1}$, in which Σ_A is the asymptotic variance-covariance matrix of $n^{-1/2}U_A(\alpha_0)$.

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 Anal* 2006;12:267–284. [PubMed: 16917734]
- Asgharian M. Comment on “Goodness-of-fit tests for parametric models based on biased samples” (2002V30 p475-490). *The Canadian J. Statist* 2003;31:349–350.
- Asgharian M, M'lan CM, 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 npml 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. *Statist. Med* 2006;25:1751–1767.
- Begg CB. On the use of familial aggregation in population-based case probands for calculating penetrance (with editorial). *J Natl Cancer Inst* 2002;94:1221–1226. [PubMed: 12189225]
- Buckley J, James I. Linear Regression with Censored Data. *Biometrika* 1979;66:429–36.
- Cheng SC, Wei LJ, Ying Z. Analysis of transformation models with censored data. *Biometrika* 1995;82:835–845.
- Cox, DR.; Miller, HD. *The theory of stochastic processes*. Chapman and Hall Ltd; London; New York: 1977.
- Cox, DR.; Oakes, D. *Analysis of Survival Data*. Chapman and Hall; London: 1984.
- Dabrowska DM, Doksum KA. Estimation and testing in a two-sample generalized odds-rate model. *J Am. Statist. Assoc* 1988;83:744–749.
- Fine JP. Analysing competing risks data with transformation models. *J. R. Statist. Soc. B* 1999;61:817–830.
- Gill, RD. *Censoring and Stochastic Integrals*. Mathematisch Centrum; Amsterdam: 1982. Mathematical Centre Tract No. 124.
- Gill RD, Vardi Y, Wellner JA. Large sample theory of empirical distributions in biased sampling models. *Ann Statist* 1988;16:1069–1112.
- Gray RJ. Estimation of regression parameters and the hazard function in transformed linear survival models. *Biometrics* 2000;56:571–576. [PubMed: 10877318]
- Kalbfleisch, JD.; Prentice, RL. *The Statistical Analysis of Failure Time Data*. Wiley; New York: 2002.

- 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. Large Sample Theory of a Modified Buckley-James Estimator for Regression Analysis with Censored Data. *Annals of Statistics* 1991;19:1370–1402.
- Lin DY, Ying Z. Semiparametric Inference for the Accelerated Life Model with Time-dependent Covariates. *Journal of Statistical Planning and Inference* 1995;44:47–63.
- Pepe M, Fleming TR. Weighted Kaplan-Meier statistics: A class of distance tests for censored survival data. *Biometrics* 1989;45:497–507. [PubMed: 2765634]
- Pepe M, Fleming TR. Weighted Kaplan-Meier statistics: Large sample and optimality considerations. *J. R. Statist. Soc. B* 1991;53:341–352.
- Prentice RL. Linear Rank Tests with Right Censored Data. *Biometrika* 1978;65:167–79.
- Ritov Y. Estimation in a Linear Regression Model with Censored Data. *Annals of Statistics* 1990;18:303–28.
- Serfling, RJ. Approximation theorems of mathematical statistics. John Wiley & Sons; New York; Chichester: 1981.
- Sansgiry P, Akman O. Transformations of the lognormal distribution as a selection model. *American Statist* 2000;54:307–309.
- Simon R. Length-biased sampling in etiologic studies. *Am J Epidemiol* 1980;111:444–452. [PubMed: 7377187]
- Tsiatis AA. Estimating Regression Parameters Using Linear Rank Tests for Censored Data. *Annals of Statistics* 1990;18:354–372.
- 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 (Com: p204-205). *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. Nonparametric estimation from cross-sectional survival data. *J Am. Statist. Assoc* 1991;86:130–143.
- Wang MC. Hazard regression analysis for length-biased data. *Biometrika* 1996;83:343–354.
- Wei LJ, Johnson WE. Combining dependent tests with incomplete repeated measurements. *Biometrika* 1985;72:359–64.
- Wolfson C, Wolfson DB, Asgharian M, M'Lan CE, Ostbye T, Rockwood K, Hogan DB, 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, Feinleib M. On the theory of screening for chronic diseases. *Biometrika* 1969;56:601–614.
- Zelen M. Forward and backward recurrence times and length biased sampling : Age specific models. *Lifetime Data Anal* 2004;10:325–34. [PubMed: 15690988]
- Zeng D, Lin DY. Maximum likelihood estimation in semiparametric regression models with censored data. *J. R. Statist. Soc. B* 2007a;69:507–564.
- Zeng D, Lin DY. Efficient Estimation for the Accelerated Failure Time Model. *J Am. Statist. Assoc* 2007b; 102:1387–1396.

Table 1

Simulation results under the proportional odds model: Estimated point estimates, coverage probabilities and MSE

β	Cohort size	Cen	$\hat{\beta}$	95% CP	MSE
(0,0)	200	0%	(.011,.016)	(.965,.981)	(.260,.746)
		10%	(.033,.005)	(.981,.976)	(.273,.727)
		30%	(.010,.009)	(.974,.972)	(.267,.788)
	400	0%	(.006,.010)	(.976,.983)	(.156,.413)
		10%	(-.013,-.005)	(.983,.986)	(.165,.436)
		30%	(-.003,-.005)	(.974,.984)	(.195,.426)
(1,1)	200	0%	(.993,.966)	(.974,.975)	(.269,.773)
		10%	(1.030,1.026)	(.963,.976)	(.274,.788)
		30%	(1.004,0.996)	(.974,.972)	(.287,.828)
	400	0%	(1.009,0.998)	(.981,.983)	(.147,.462)
		10%	(1.006,1.007)	(.978,.978)	(.149,.430)
		30%	(1.002,1.016)	(.971,.978)	(.156,.450)
(2,2)	200	0%	(2.020,2.027)	(.956,.966)	(.332,.940)
		10%	(2.039,2.053)	(.955,.971)	(.339,.963)
		30%	(2.030,1.986)	(.962,.970)	(.352,.998)
	400	0%	(2.003,2.008)	(.965,.975)	(.184,.532)
		10%	(2.014,1.999)	(.961,.973)	(.188,.531)
		30%	(2.019,1.981)	(.966,.966)	(.198,.557)

NOTE: For $\hat{\beta}$, average standard errors in the above scenarios ranged from (.363,.621) for $\beta = (0, 0)$, $n = 400$ and cen = 0% to (0.587,.989) for $\beta = (2, 2)$, $n = 200$, and cen = 30%.

Table 2

Simulation results under the proportional hazards model: Biases and coverage probabilities for the proposed method (LB) with the naive Cox model

β	Cohort size	Cen	Bias		95% Cov. Prob.	
			Cox	LB	Cox	LB
(0,0)	200	0%	(.003, -.009)	(.009, -.005)	(.953, .954)	(.979, .983)
		10%	(-.000, .004)	(-.006, .001)	(.960, .952)	(.974, .978)
		30%	(.009, -.004)	(-.007, -.014)	(.964, .938)	(.959, .969)
	400	0%	(-.005, -.002)	(.001, -.011)	(.954, .947)	(.980, .977)
		10%	(-.005, -.002)	(-.003, .001)	(.951, .950)	(.987, .983)
		30%	(-.005, -.001)	(-.002, -.002)	(.949, .955)	(.962, .975)
(1,1)	200	0%	(.261, .253)	(.019, .013)	(.649, .836)	(.971, .967)
		10%	(.286, .279)	(.031, .032)	(.676, .840)	(.974, .980)
		30%	(.319, .320)	(.034, .061)	(.621, .821)	(.974, .974)
	400	0%	(.255, .252)	(.012, .001)	(.404, .715)	(.982, .982)
		10%	(.267, .262)	(.025, .007)	(.421, .725)	(.976, .967)
		30%	(.304, .313)	(.028, .029)	(.381, .694)	(.971, .961)
Censoring depend on Z_2 with an effect $\gamma = 0.04$						
(1,1)	200	10%	(.276, .268)	(.035, .019)	(.650, .840)	(.975, .976)
	30%	(.319, .309)	(.040, .012)	(.638, .827)	(.972, .977)	
Censoring depend on Z_2 with an effect $\gamma = 0.2$						
	200	10%	(.280, .258)	(.021, .030)	(.641, .846)	(.982, .981)
	30%	(.316, .292)	(.032, .066)	(.631, .855)	(.976, .980)	

NOTE: For $\widehat{\beta}$, average standard errors in the above scenarios ranged from (.207, .355) for $\beta = (0,0)$, $n = 400$ and cen = 0% to (0.351, .580) for $\beta = (1,1)$, $n = 200$, and cen = 30%; the corresponding MSEs ranged from (.070, .222) to (.221, .606).

Table 3

Simulation results under accelerated failure time model: Biases and coverage probabilities for the proposed method (Eq. U_A) and the naive method (Eq. U_{A2})

N	Cen	Bias			95% Cov. Prob.		
		Eq. U_{A2}	Eq. U_A	($\alpha_I, \alpha_1, \alpha_2$) = (1, 0, 0)	Eq. U_{A2}	Eq. U_A	
200	0%	(.082, -.000, .002)	(.001, -.000, .002)	(.556, .937, .930)	(.936, .949, .929)		
	10%	(.091, .002, .001)	(-.001, .002, .000)	(.501, .937, .942)	(.930, .935, .935)		
	30%	(.143, .001, -.000)	(.001, .002, -.001)	(.259, .942, .950)	(.930, .940, .945)		
400	0%	(.081, -.000, .001)	(.000, -.000, .001)	(.269, .959, .929)	(.946, .955, .932)		
	10%	(.093, -.000, .002)	(-.000, -.000, .002)	(.231, .948, .940)	(.929, .951, .938)		
	30%	(.156, -.001, .002)	(.000, -.001, .001)	(.037, .941, .945)	(.933, .950, .941)		
(1, 1, 1)							
200	0%	(.084, -.002, -.002)	(.006, -.003, -.005)	(.666, .941, .935)	(.924, .935, .925)		
	10%	(.083, .007, .008)	(.006, -.003, -.003)	(.679, .932, .931)	(.935, .947, .943)		
	30%	(.123, .048, -.038)	(.005, -.004, -.004)	(.531, .775, .884)	(.939, .941, .948)		
400	0%	(.083, .001, .001)	(.002, .001, .002)	(.450, .941, .943)	(.931, .936, .942)		
	10%	(.079, .010, .010)	(.002, .000, .002)	(.503, .919, .909)	(.936, .948, .939)		
	30%	(.102, .080, -.006)	(.002, -.000, -.003)	(.421, .393, .891)	(.946, .942, .943)		
(1, 1, 1)							
Censoring depend on Z_2 with an effect $\gamma = 0.04$							
200	10%	(.090, .069, -.005)	(.006, -.005, -.003)	(.633, .648, .898)	(.941, .946, .949)		
	30%	(.337, .001, -.300)	(.005, -.005, -.000)	(.014, .901, .176)	(.943, .942, .951)		
Censoring depend on Z_2 with an effect $\gamma = 0.2$							
200	10%	(.087, .067, .001)	(.005, -.006, -.000)	(.660, .654, .890)	(.940, .949, .953)		
	30%	(.330, .007, -.289)	(.002, -.008, .005)	(.015, .900, .218)	(.947, .939, .957)		

NOTE: The simulation model is $\log Y = \alpha_I + \alpha_1 Z_1 + \alpha_2 Z_2 + \varepsilon_I$. For $(\widehat{\alpha}_I, \widehat{\alpha}_1, \widehat{\alpha}_2)$, average standard errors of the proposed method ranged from (0.034, 0.030, 0.053) for $(\alpha_I, \alpha_1, \alpha_2) = (1, 0, 0)$, $n=400$, and cen=0% to (0.071, 0.054, 0.095) for $(\alpha_I, \alpha_1, \alpha_2) = (1, 1, 1)$, $n = 200$, cen=30%.

Table 4

Dementia example: Regression coefficient estimates and standard errors (SE)

	Int. (SE)	VD (SE)	PA (SE)
PH			
LB		.241 (.211)	-.110 (.201)
Cheng		.160 (.111)	-.159 (.097)
PO			
LB		.360 (.316)	-.164 (.302)
Cheng		.240 (.167)	-.239 (.146)
AFT			
LB	7.210 (.117)	-.075 (.143)	.129 (.152)
Eq. U_{A2}	8.083 (.026)	.014 (.058)	.058 (.045)

NOTE: Int., intercept; VD, vascular dementia; PA, possible Alzheimer; PH, proportional hazards model; PO, proportional odds model; AFT, accelerated failure time model; SE, standard error; and Cheng, the methods proposed in Cheng et al. (1995). LB refers to the proposed estimating method under each of the three models (with weight $q=1$). The estimated SE of the covariate coefficients for Cheng are obtained from the Jackknife method.