Available online at www.sciencedirect.com
SCIENCE DIRECT®

Journal of Statistical Planning and Inference 129 (2005) 145–157

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Estimating equation approach for regression analysis of failure time data in the presence of interval-censoring

Xingqiu Zhao[a], Hee-jeong Lim[b], Jianguo Sun[c,*]

[a]*Department of Mathematical and Statistical Sciences, 632 Central Academic Building, University of Alberta, Edmonton, Alberta, T6G 2G1 Canada*

[b]*Department of Mathematics and Computer Science, Northern Kentucky University, Highland Heights, KY 41099, USA*

[c]*Department of Statistics, University of Missouri, 146 Middlebush Hall, Columbia, MO 65211, USA*

Available online 25 August 2004

Abstract

This article discusses statistical inference for the proportional hazards model when there exists interval-censoring on both survival time of interest and covariates (J. Roy. Statist. Soc. B 34 (1972) 187; Encyclopedia of Biostatistics. Wiley, New York, 1998, pp. 2090–2095). In particular, we consider situations where observations on the survival time are doubly censored and observations on covariates are interval-censored. For inference about regression parameters, a general estimating equation approach is proposed. The proposed estimate of the parameter is a generalization of the maximum partial-likelihood estimate for right-censored failure time data with known or exactly observed covariates (The Statistical Analysis of Failure Time Data. Wiley, New York, 1980). The asymptotic properties of the proposed estimate are established and its finite sample properties are investigated through a simulation study.

Keywords: Estimating equations; Interval-censoring; Proportional hazards models; Regression analysis

* Corresponding author. Tel.: +1-573-882-6376; fax: +1-573-884-5524.
E-mail address: sunj@missouri.edu (J. Sun).

0378-3758/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2004.06.044

1. Introduction

The proportional hazards model is the most commonly used regression model in survival analysis and is defined as

$$\lambda(t) = Y(t)\lambda_0(t)\ \exp(\beta' Z) \tag{1}$$

given a vector of covariates Z (Cox, 1972, Andersen and Gill, 1982). In the above, $\lambda_0(t)$ is an unknown baseline hazard function, β denotes the vector of regression coefficients, and $Y(t)$ is a predictable process taking value 0 or 1 indicating (by the value 1) if a subject is under observation at time t. When right-censored failure time data are available, many authors have studied the inference problem about β (Kalbfleisch and Prentice, 1980). In this paper, we consider a more general situation where the survival time of interest is doubly censored (De Gruttola and Lagakos, 1989) and covariates are interval-censored (Sun, 1998).

By doubly censored survival time, we mean that the survival time of interest is defined as the elapsed time between two related events, originating and end events. Furthermore, observations on the occurrences of the two events are interval- and right-censored, respectively. By interval-censoring, we mean that the occurrence time of the originating event is observed only to belong to an interval. Note that if the occurrence of the originating event is observed exactly, we would have right-censored observations on the survival time. By interval-censored covariates, we mean that the covariates are scalar variables or times to certain events and their values are known or observed only to belong to some intervals instead of being exactly known. Our goal is to make inference about regression parameters β.

One field in which doubly censored failure time data often occur is epidemiological studies, where the originating and end events may represent infection and onset of certain diseases, respectively. In particular, many authors have discussed such data in the context of AIDS studies (De Gruttola and Lagakos, 1989, Kim et al., 1993, Sun et al., 1999). In this case, the two events correspond, respectively, to HIV infection and AIDS diagnosis. The survival time of interest, the time from HIV infection to the diagnosis of AIDS, is often referred to as AIDS incubation time and plays an important role in the study of AIDS epidemic. HIV infection time is usually interval-censored in these studies because HIV status can only be checked periodically. In the meantime, AIDS diagnosis times are commonly right-censored due to patient drop-out from the study or the end of the study.

Goggins et al. (1999) discussed an example about interval-censored covariates arising from an AIDS clinical trial. In the example, the problem of interest is to predict the onset of active cytomegalovirus (CMV) using CMV shedding assuming that they are related by the proportional hazards model. However, the exact time to CMV shedding is usually unobservable since its determination is through clinical screen of blood or urine, which can only be performed periodically. In other words, only interval-censored CMV shedding times are available. Gómez et al (2000) described a similar example also from an AIDS clinical trial.

Several methods have been proposed for inference about β when interval-censoring occurs only on either survival time of interest (Kim et al., 1993, Sun et al., 1999) or covariates (Goggins et al., 1999). One shortcoming of these methods is that their asymptotic properties are unknown. There seems no existing method in the literature for the situation considered here: interval-censoring on both the survival time of interest and the covariates. It includes as special cases the situations considered in Goggins et al. (1999), Gómez et al (2000), Kim et al. (1993) and Sun et al. (1999).

The remainder of the paper is organized as follows. We begin in Section 2 with introducing some notation and briefly reviewing the partial-likelihood method (Kalbfleisch and Prentice, 1980) and the approach proposed in Sun et al. (1999). Section 3 considers inference about β when only doubly censored data and interval-censored data are available for survival time and covariates, respectively. To estimate regression parameter β, a general estimating equation approach is proposed. The proposed method involves only regression parameter β and is a generalization of that given in Sun et al. (1999). The asymptotic consistency and normal distribution of the proposed estimate are established. In Section 4, we report some results from a simulation study conducted to assess finite sample properties of the proposed estimate and Section 5 contains some concluding remarks. As other authors, we assume that the survival time of interest is independent of the occurrence time of the originating event and that the mechanism yielding interval-censoring is independent of related variables.

2. Notation and review

Consider a survival study that involves n independent subjects and in which each subject experiences an originating event and an end event as before. For subject i, let X_i and S_i^* denote the times at which the originating and end events occur, respectively, and define $T_i = S_i^* - X_i$, the survival time of interest, for $i = 1, \ldots, n$. Suppose that the hazard function of the T_i's is given by model (1). For simplicity of presentation, we will assume that the Z_i's are scalars and some comments on this will be given below.

To describe observed data, suppose that for each subject, there exists a censoring time C_i which is independent of S_i^*, $i = 1, \ldots, n$. For S_i^*, suppose that we observe only $\{S_i = \min(S_i^*, C_i), \delta_i = I(S_i = S_i^*);\ i = 1, \ldots, n\}$, where $I(\cdot)$ is the indicator function. For X_i and Z_i, suppose that only intervals $[L_i, R_i]$ and $[U_i, V_i]$ are observed such that $X_i \in [L_i, R_i]$ and $Z_i \in [U_i, V_i]$. That is, we have right-censored data for the S_i^*'s and interval-censored data for the X_i's and Z_i's. If $L_i = R_i$, we then have usual right-censored failure time data on the T_i's and the situation reduces to that discussed in Goggins et al. (1999) and Gómez et al (2000). On the other hand, $U_i = V_i$ corresponds to situations discussed by Kim et al. (1993) and Sun et al. (1999).

For given X_i, define $Y_i(t|X_i) = I(S_i - X_i \geqslant t)$ and $N_i(t|X_i) = I(S_i - X_i \leqslant t, \delta_i = 1)$, $i = 1, \ldots, n$. Let $X = (X_1, \ldots, X_n)$, $Z = (Z_1, \ldots, Z_n)$, and

$$S^{(j)}(\beta, t|X, Z) = \frac{1}{n} \sum_{i=1}^{n} Y_i(t|, X_i) Z_i^j \mathrm{e}^{\beta Z_i},$$

$j = 0, 1$, where $Z_i^0 = 1$ and $Z_i^1 = Z_i$. Also let H and G denote the cumulative distribution functions of the X_i's and Z_i's, respectively. Note that Y_i is a predictive process and $N_i(t|X_i)$ is a counting process.

Suppose that $L_i = R_i$ and $U_i = V_i$, $i = 1, \ldots, n$. That is, we have right-censored data on the T_i's. In this case, the most commonly used estimate for β is the partial-likelihood estimate defined as the solution to the partial score equation

$$U(\beta|X, Z) = \int_0^\tau \sum_{i=1}^n \left\{ Z_i - \frac{S^{(1)}(\beta, t|X, Z)}{S^{(0)}(\beta, t|X, Z)} \right\} \mathrm{d}N_i(t|X_i) = 0 \tag{2}$$

(Kalbfleisch and Prentice, 1980), where τ denotes the longest possible follow-up time. On the other hand, if we only have $U_i = V_i$ with $L_i \leqslant R_i$, motivated by the marginal-likelihood idea, Sun et al. (1999) proposed to estimate β using the following estimating equation:

$$U(\beta, \hat{H}|Z) = \left(\prod_{l=1}^n \hat{a}_l^{-1} \right) \int_{L_1}^{R_1} \cdots \int_{L_n}^{R_n} U(\beta|x_i's, Z) \prod_{l=1}^n (\mathrm{d}\hat{H}(x_l)) = 0. \tag{3}$$

In the above, $\hat{a}_l = \int_{L_l}^{R_l} \mathrm{d}\hat{H}(x)$, $l = 1, \ldots, n$, and $\hat{H}$ denotes the nonparametric maximum-likelihood estimate of H based on interval-censored data on the X_i's, which will be commented below. Note that the function $U(\beta \mid X, Z)$ in (2) can be regarded as a conditional score function given X and Z or as a score function about both parameters β and X if we treat X as nuisance parameters. The function $U(\beta, \hat{H}|Z)$ in (3) is simply the integration of $U(\beta \mid X, Z)$ with respect to the unknown X conditional on observed data. It is apparent that if $L_i = R_i$, Eq. (3) reduces to Eq. (2).

Note that a major advantage of Eqs. (2) and (3) is that they are independent of the unknown baseline hazard function $\lambda_0(t)$, which makes the study of the asymptotic properties of the resulting estimates relatively easy compared to the maximum full-likelihood estimates. Also they can be easily solved. For the estimate given by Eq. (2), it has been shown to be efficient and its asymptotic properties have been established (Kalbfleisch and Prentice, 1980). Sun et al. (1999) discussed the asymptotic properties of the estimate resulting from Eq. (3) without rigorous proofs. In the next section, we generalize the above methods to the situation where both the X_i's and Z_i's are interval-censored and the asymptotic properties of the generalized estimate are rigorously studied.

3. Estimation of $\boldsymbol{\beta}$

Now we consider estimation of β assuming that both the X_i's and Z_i's are interval-censored. That is, observations on the T_i's and Z_i's are doubly and interval-censored, respectively. Let $\hat{G}$ denote the nonparametric maximum-likelihood estimate of G based on interval-censored data on the Z_i's, on which some comments will be given below. Note that for current situation, the estimating function $U(\beta, \hat{H}|Z)$ is not fully defined since the Z_i's are not observed. To estimate β, we propose to use the following estimating equation:

$$U(\beta, \hat{H}, \hat{G}) = \left(\prod_{l=1}^n \hat{b}_l^{-1} \right) \int_{U_1}^{V_1} \cdots \int_{U_n}^{V_n} U(\beta, \hat{H}|z_l's) \prod_{l=1}^n (\mathrm{d}\hat{G}(z_l)) = 0, \tag{4}$$

where $\hat{b}_l = \int_{U_l}^{V_l} \mathrm{d}\hat{G}(z)$, $l = 1, \ldots, n$.

The function $U(\beta, \hat{H}, \hat{G})$ can be regarded as an estimate of the function $U(\beta, \hat{H}|Z)$ with missing Z_i's. It is obvious that if usual right-censored data are observed and all covariates are known, it reduces to the partial-likelihood score function $U(\beta|X, Z)$ given in (2). As $U(\beta|X, Z)$ and $U(\beta, \hat{H}|Z)$, $U(\beta, \hat{H}, \hat{G})$ has an advantage that it does not involve the baseline hazard function $\lambda_0(t)$, which makes both implementation of the method and the study of the properties of the resulting estimate of β relatively easier.

Let β_0 denote the true value of β and $\hat{\beta}$ the estimator of β_0 given by the solution to Eq. (4). Then it can be shown that $\hat{\beta}$ is a consistent estimate of β_0. Furthermore, under mild regularity conditions, $n^{1/2}(\hat{\beta} - \beta_0)$ has an asymptotic normal distribution with mean zero and covariance matrix that can be consistently estimated by $A(\hat{\beta})\Gamma(\hat{\beta})A'(\hat{\beta})$, where $A(\beta) = \{-n^{-1}\partial U(\beta, \hat{H}, \hat{G})/\partial\beta\}^{-1}$ and $\Gamma(\beta) = n^{-1}\sum_{i=1}^{n}\hat{h}_i^2(\beta)$, where

$$\hat{h}_i(\beta) = \int_0^{\tau}\int_{U_1}^{V_1}\cdots\int_{U_n}^{V_n}\int_{L_1}^{R_1}\cdots\int_{L_n}^{R_n}\left\{z_i - \frac{S^{(1)}(\beta, t|x_i's, z_i's)}{S^{(0)}(\beta, t|x_i's, z_i's)}\right\}\{\mathrm{d}N_i(t|x_i) - \frac{Y_i(t|x_i)\exp(\beta' z_i)\,\mathrm{d}\bar{N}(t)}{nS^{(0)}(\beta, t|x_i's, z_i's)}\right\}\prod_{l=1}^{n}\frac{\mathrm{d}\hat{H}(x_l)}{\hat{a}_l}\prod_{l=1}^{n}\frac{\mathrm{d}\hat{G}(z_l)}{\hat{b}_l}$$

and $\bar{N}(t|x) = \sum_{i=1}^{n} N_i(t|x_i)$. The proofs of the above results are given in the appendix.

To obtain the estimator $\hat{\beta}$, one way is to directly solve the estimating Eq. (4) using existing optimization algorithms, which are available in most statistical software. This is usually feasible for small data sets and some large data sets for which the resulting estimators $\hat{H}$ and $\hat{G}$ do not have many jumps. For general situations, we propose to apply the following simple Monte Carlo method. Let K_1 and K_2 be given integers.

Step 1: For each $k_1 = 1, \ldots, K_1$ and $i = 1, \ldots, n$, randomly sample $X_i^{(k_1)}$ from $\hat{H}$ conditional on observed interval $[L_i, R_i]$, that is, $X_i^{(k_1)} \in [L_i, R_i]$.

Step 2: For each $k_2 = 1, \ldots, K_2$ and $i = 1, \ldots, n$, randomly sample $Z_i^{(k_2)}$ from $\hat{G}$ conditional on observed interval $[U_i, V_i]$, that is, $Z_i^{(k_2)} \in [U_i, V_i]$.

Step 3: Let $X^{(k_1)} = (X_1^{(k_1)}, \ldots, X_n^{(k_1)})$ and $Z^{(k_2)} = (Z_1^{(k_2)}, \ldots, Z_n^{(k_2)})$ and calculate $U(\beta|X^{(k_1)}, Z^{(k_2)})$, $k_1 = 1, \ldots, K_1$, $k_2 = 1, \ldots, K_2$.

Step 4: Solve the equation

$$\frac{1}{K_1 K_2}\sum_{k_1=1}^{K_1}\sum_{k_2=1}^{K_2} U(\beta|X^{(k_1)}, Z^{(k_2)}) = 0.$$

If K_1 and K_2 are large, we should expect that the left-hand side of the above equation will give a good approximation to $U(\beta, \hat{H}, \hat{G})$. In the above, we need to determine the maximum-likelihood estimates $\hat{H}$ and $\hat{G}$. For this, a few algorithms have been proposed in the literature. For example, Turnbull (1976) gave a simple self-consistency algorithm for the problem and Gentleman and Geyer (1994) developed a procedure using combination and optimization theory. In the simulation study reported below, the Turnbull's algorithm is used.

Sometimes, one may also be interested in estimating the cumulative baseline hazard function $\Lambda_0(t) = \int_0^t \lambda_0(s)\,ds$. Given $\hat{\beta}$, an estimate is given by

$$\hat{\Lambda}_0(t) = \int_{L_1}^{R_1} \cdots \int_{L_n}^{R_n} \int_{U_1}^{V_1} \cdots \int_{U_n}^{V_n} \int_0^t \frac{\sum_{i=1}^n dN_i(s|x_i)}{nS^{(0)}(\hat{\beta}, s|x, z)} \times \prod_{k=1}^n b_k^{-1}\, d\hat{G}(z_k) \prod_{l=1}^n a_l^{-1}\, d\hat{H}(x_l),$$

which reduces to Breslow's estimate of $\Lambda_0(t)$ if exact observations on both X_i and Z_i are observed.

4. Simulation study

A simulation study was conducted to evaluate finite sample properties of the proposed estimate $\hat{\beta}$. In the study, for simplicity, we assume that all concerned variables are discrete and take integer values and $L_i = R_i$. In the study, we mimicked clinical trials and generated the true covariates Z_i's from the discretized exponential distribution with the hazard rate of 0.25. For each Z_i, the censoring interval was constructed as $U_i = \max\{1, \min(Z_i - a_i^1, 10)\}$ and $V_i = \min\{Z_i + a_i^2, 10\}$, where a_i^1 and a_i^2 were generated from the uniform distribution $U\{0, 1, \ldots, b\}$, where b is a constant. Given Z_i, T_i was generated from the discretized Weibull distribution with the hazard function given in (1) and the common censoring time $C_i = 20$. In the study, we set $\lambda_0(t) = 2t/225$. The results reported below are based on the sample size of $n = 100$, $K_2 = 100$ and 1000 replications.

In the study, we mainly focused on the comparison of the proposed point estimate and the maximum partial-likelihood estimate, $\hat{\beta}_p$ say, of β that would be obtained if the covariate was exactly observed. Also we were interested in investigating the approximation of the asymptotic normal distribution given in the previous section to the finite distribution of the proposed estimate. Table 1 presents the means of $\hat{\beta}$ based on simulated data for different true values of β with $b = 1$. It also gives the 95% empirical coverage probabilities and the estimated size or power for testing $\beta=0$ using the standardized $\hat{\beta}$ as the test statistic based on simulated data. For the comparison, assuming that exact values of covariates are known, we also obtained the corresponding results for $\hat{\beta}_p$ and included them in Table 1. It can be seen

Table 1
Simulation results for the proposed estimate of regression parameter

	$\hat{\beta}$			$\hat{\beta}_p$		
True β	Mean	Size or power	95% CP	Mean	Size or power	95% CP
000	0.0013	0.052	0.948	0.0015	0.041	0.959
010	0.0945	0.774	0.941	0.0963	0.763	0.963
−010	−0.0955	0.578	0.951	−0.0969	0.584	0.952

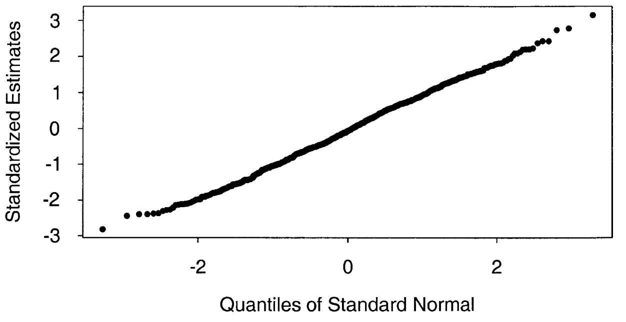

Fig. 1. Quantile plot with exact covariates and beta $= 0.0$.

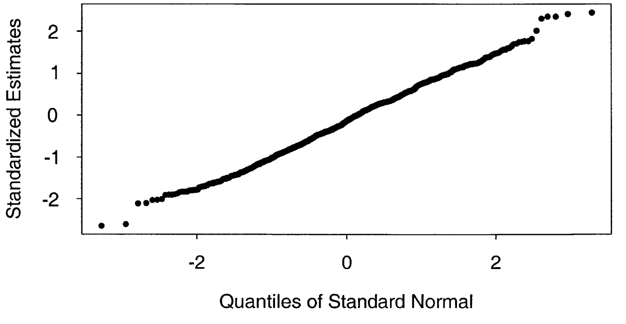

Fig. 2. Quantile plot with interval-censored covariates and beta $= 0.0$.

from the table that the results based on interval-censored covariates and exact covariates are quite close to each other for most cases considered, indicating that the proposed method works reasonably well.

For the assessment of the asymptotic normal distribution derived for the proposed estimate, the probability plots of the standardized $\hat{\beta}$ against the standard normal distribution were studied and compared to the corresponding plots of the standardized $\hat{\beta}_p$. Figs. 2, 4 and 6 display the quantile plots of the standardized $\hat{\beta}$ for the case of $\beta = 0,\ 0.1,\ -0.1$, respectively and the corresponding plots for the standardized $\hat{\beta}_p$ are given in Figs. 1, 3 and 5. It seems that the approximation to the finite sample distribution of $\hat{\beta}$ is satisfactory. In the simulation study, we also tried $K_2 = 200,\ 500$ and no significant differences were observed (Figs. 1–6). We also considered other set ups for generating data and obtained similar results.

5. Concluding remarks

This paper considered statistical inference about the proportional hazards model when there exists interval-censoring on observations on both survival time of interest and covari-

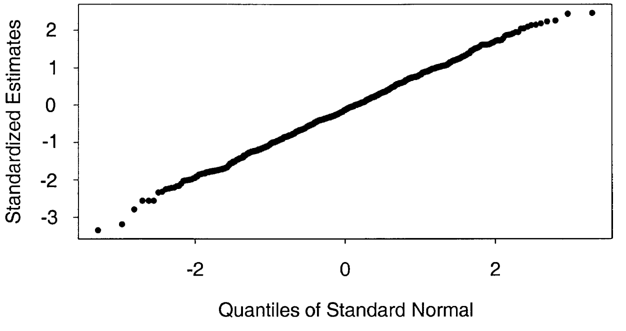

Fig. 3. Quantile plot with exact covariates and beta $= 0.1$.

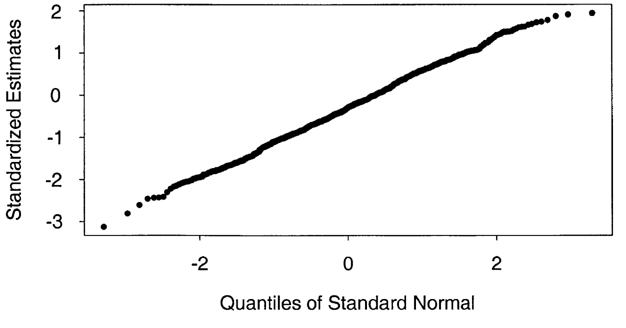

Fig. 4. Quantile plot with interval-censored covariates and beta $= 0.1$.

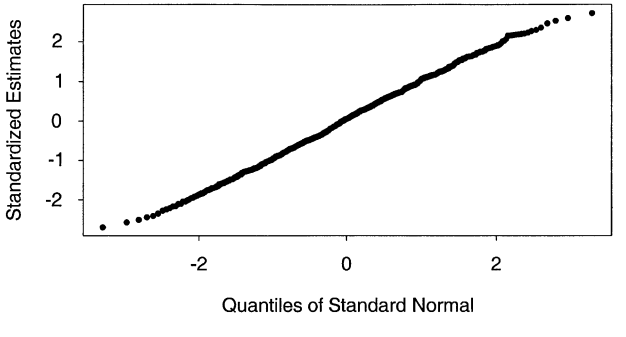

Fig. 5. Quantile plot with exact covariates and beta $= -0.1$.

ates. There is no existing research for the situation discussed here except that a few authors have discussed some special cases. To estimate regression parameters, we proposed an estimating equation approach and the proposed estimate is generalizations of the maximum partial-likelihood estimate (Kalbfleisch and Prentice, 1980) and the estimate given in Sun et al. (1999). In addition to the simplicity of the method, the asymptotic properties of the

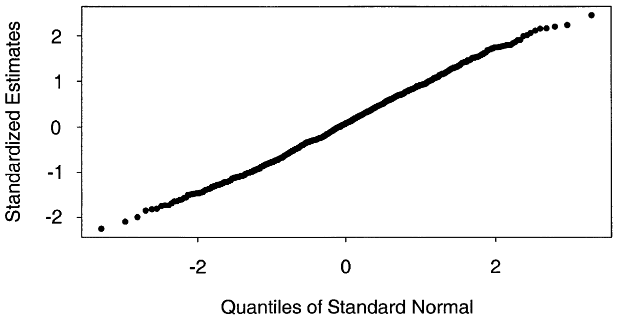

Fig. 6. Quantile plot with interval-censored covariates and beta $= -0.1$.

proposed estimate are established, while no rigorous asymptotic study was given in the studies that discussed these special cases (Goggins et al., 1999, Kim et al., 1993, Sun et al., 1999).

In the preceding sections, we have assumed that the covariate is a scalar. The proposed method can be easily generalized to the case where covariates are a vector. In this situation, $U(\beta, \hat{H}, \hat{G})$ will be a vector and the integral in Eq. (4) will be with respect to all components of the covariates. A similar method could also be developed for the case where only some components of the covariates are interval-censored and others are exactly observed.

A direction for future research is to generalize the proposed estimating Eq. (4) by incorporating a weight process. By this, we mean that instead of using $U(\beta|X, Z)$ in $U(\beta, \hat{H}|Z)$ and $U(\beta, \hat{H}, \hat{G})$, one can use the following weighted partial-likelihood score function

$$U_W(\beta|X, Z) = \int_0^\tau \sum_{i=1}^n W_i(t) \left\{ Z_i - \frac{S^{(1)}(\beta, t|X, Z)}{S^{(0)}(\beta, t|X, Z)} \right\} \mathrm{d}N_i(t|X_i),$$

where W_i is a weighting process. It is apparent that this will give more choices on the estimates of β and if both X_i and Z_i are exactly observed, the resulting weighted estimates of β will become weighted partial-likelihood estimates.

Appendix. Proofs

Let β_0, $\hat{\beta}$, $U(\beta|X, Z)$, and $U(\beta, \hat{H}, \hat{G})$ be defined as before and use the notation given in the previous sections. Assume that the regularity conditions given in Andersen and Gill (1982) for the case of known covariates and in Yu et al. (1998) for the strong consistency of $\hat{H}$ and $\hat{G}$ hold.

Consistency of $\hat{\beta}$: Define

$$A_n(\beta, t|X, Z) = n^{-1}\left[\sum_{i=1}^{n}\int_0^t (\beta-\beta_0)' Z_i \, \mathrm{d}N_i(u|X_i) - \int_0^t \log\left\{\frac{\sum_{i=1}^{n} Y_i(u|X_i)\mathrm{e}^{\beta' Z_i}}{\sum_{i=1}^{n} Y_i(u|X_i)\mathrm{e}^{\beta_0' Z_i}}\right\} \mathrm{d}\bar{N}(u|X_i)\right],$$

$$B_n(\beta, t|X, Z) = \int_0^t \left[(\beta-\beta_0)' S^{(1)}(\beta_0, u|X, Z) - \log\left\{\frac{S^{(0)}(\beta, u|X, Z)}{S^{(0)}(\beta_0, u|X, Z)}\right\} S^{(0)}(\beta_0, u|X, Z)\right]\lambda_0(u)\,\mathrm{d}u,$$

and

$$M_i(t|X_i, Z_i) = N_i(t|X_i) - \int_0^t \lambda_0(s) Y_i(s|X_i)\exp(\beta_0' Z_i)\,\mathrm{d}s,$$

$i = 1, 2, \ldots, n$. Then we have

$$\begin{aligned} &A_n(\beta, t|X, Z) - B_n(\beta, t|X, Z) \\ &\quad = n^{-1}\left[\sum_{i=1}^{n}\int_0^t \left\{(\beta-\beta_0)' Z_i - \log\frac{S^{(0)}(\beta, u|X, Z)}{S^{(0)}(\beta_0, u|X, Z)}\right\} \mathrm{d}M_i(u|X_i, Z_i)\right], \end{aligned}$$

which is a locally square integrable martingale. Note that

$$\begin{aligned} &\left\langle A_n(\beta, t|X, Z) - B_n(\beta, t|X, Z), A_n(\beta, t|X, Z) - B_n(\beta, t|X, Z)\right\rangle \\ &\quad = n^{-2}\sum_{i=1}^{n}\int_0^t \left[(\beta-\beta_0)' Z_i - \log\left\{\frac{S^{(0)}(\beta, u|X, Z)}{S^{(0)}(\beta_0, u|X, Z)}\right\}\right]^2 Y_i(u|X_i)\mathrm{e}^{\beta_0' Z_i}\lambda_0(u)\,\mathrm{d}u \\ &\quad = n^{-1}\int_0^t \left[(\beta-\beta_0)' S^{(2)}(\beta_0, u|X, Z)(\beta-\beta_0) - 2(\beta-\beta_0)' \right. \\ &\qquad \times S^{(1)}(\beta_0, u|X, Z)\log\left\{\frac{S^{(0)}(\beta, u|X, Z)}{S^{(0)}(\beta_0, u|X, Z)}\right\} \\ &\qquad \left. + \log^2\left\{\frac{S^{(0)}(\beta, u|X, Z)}{S^{(0)}(\beta_0, u|X, Z)}\right\} S^{(0)}(\beta_0, u|X, Z)\right]\lambda_0(u)\,\mathrm{d}u, \end{aligned}$$

where

$$S^{(2)}(\beta, t|X, Z) = \frac{1}{n}\sum_{i=1}^{n} Z_i^{\otimes 2} Y_i(t|X_i)\mathrm{e}^{\beta' Z_i}.$$

Thus we have that asymptotically,

$$A_n(\beta, t|X, Z) - B_n(\beta, t|X, Z) \longrightarrow 0$$

in probability.

Also note that asymptotically,

$$B_n(\beta, \tau|X, Z) \longrightarrow A(\beta, \tau)$$

in probability, where

$$A(\beta, \tau) = \int_0^\tau \left[(\beta - \beta_0)' s^{(1)}(\beta_0, u) - \log \left\{ \frac{s^{(0)}(\beta, u)}{s^{(0)}(\beta_0, u)} \right\} s^{(0)}(\beta_0, u) \right] \lambda_0(u)\, \mathrm{d}u$$

and $s^{(0)}$ and $s^{(1)}$ are the asymptotic limits of $S^{(0)}$ and $S^{(1)}$, respectively. Thus, we have

$$A_n(\beta, \tau|X, Z) \longrightarrow A(\beta, \tau)$$

and

$$\begin{aligned} A_n(\beta, \tau) = & \int_{U_1}^{V_1} \cdots \int_{U_n}^{V_n} \int_{L_1}^{R_1} \cdots \int_{L_n}^{R_n} A_n(\beta, \tau|X, Z) \\ & \times \prod_{l=1}^{n} \hat{a}_l^{-1} \mathrm{d}\hat{H}(X_l) \prod_{l=1}^{n} \hat{b}_l^{-1}\, \mathrm{d}\hat{G}(Z_l) \to A(\beta, \tau) \end{aligned}$$

both in probability due to the strong consistency of $\hat{H}$ and $\hat{G}$ (Yu et al., 1998). The consistency of $\hat{\beta}$ therefore follows from the above second equation and the fact that both $A_n(\beta, \tau)$ and $A(\beta, \tau)$ are concave functions of β with a unique maximum at $\beta = \hat{\beta}$ and $\beta = \beta_0$, respectively.

Asymptotic normality of $\hat{\beta}$: To prove the asymptotic normality, first note that the application of Taylor series expansion to $U(\beta, \hat{H}, \hat{G})$ yields, asymptotically,

$$n^{-\frac{1}{2}} U(\beta_0, \hat{H}, \hat{G}) = \left\{ -n^{-1} \frac{\partial U(\beta^*, \hat{H}, \hat{G})}{\partial \beta} \right\} \{ n^{\frac{1}{2}} (\hat{\beta} - \beta_0) \},$$

where β^* is on the segment between β_0 and $\hat{\beta}$. Following Anderson and Gill (1982), we can easily show that $A^{-1}(\beta^*) = -n^{-1}\, \partial U(\beta^*, \hat{H}, \hat{G})/\partial\beta$ converges in probability as $n \to \infty$. Thus, for the proof, it is sufficient to show that $n^{-1/2} U(\beta_0, \hat{H}, \hat{G})$ is asymptotically normally distributed with mean zero and covariance matrix that can be estimated by $\Gamma(\hat{\beta})$ given in Section 3.

To see the asymptotic distribution of $n^{-1/2} U(\beta_0, \hat{H}, \hat{G})$, note that

$$\begin{aligned} n^{-\frac{1}{2}} U(\beta_0|X, Z) = & n^{-\frac{1}{2}} \int_0^\tau \sum_{i=1}^{n} \left\{ Z_i - \frac{S^{(1)}(\beta_0, t|X, Z)}{S^{(0)}(\beta_0, t|X, Z)} \right\} \\ & \times \{ \mathrm{d}M_i(t|X_i, Z_i) + \lambda_0(t) Y_i(t|X_i) \mathrm{e}^{\beta_0' Z_i}\, \mathrm{d}t \} \\ = & n^{-\frac{1}{2}} \int_0^\tau \sum_{i=1}^{n} \left\{ Z_i - \frac{S^{(1)}(\beta_0, t|X, Z)}{S^{(0)}(\beta_0, t|X, Z)} \right\} \mathrm{d}M_i(t|X_i, Z_i), \end{aligned}$$

which is asymptotically equivalent to

$$n^{-\frac{1}{2}} \sum_{i=1}^{n} \int_0^{\tau} u_i(\beta_0, t|Z_i) \, \mathrm{d}M_i(t|X_i, Z_i),$$

where

$$u_i(\beta_0, t|Z_i) = Z_i - \frac{s^{(1)}(\beta_0, t)}{s^{(0)}(\beta_0, t)} \ .$$

Hence asymptotically,

$$\begin{aligned} n^{-\frac{1}{2}} U(\beta_0, \hat{H}, \hat{G}) = & \int_{U_1}^{V_1} \cdots \int_{U_n}^{V_n} \int_{L_1}^{R_1} \cdots \int_{L_n}^{R_1} \\ & \times \left\{ n^{-\frac{1}{2}} \sum_{i=1}^{n} \int_0^{\tau} u_i(\beta_0, t|Z_i) \, \mathrm{d}M_i(t|X_i, Z_i) \right\} \\ & \times \prod_{l=1}^{n} \hat{a}_j^{-1} \, \mathrm{d}\hat{H}(X_j) \prod_{l=1}^{n} \hat{b}_l^{-1} \, \mathrm{d}\hat{G}(Z_l) \\ = & n^{-\frac{1}{2}} \sum_{i=1}^{n} \hat{b}_i^{-1} \int_{U_i}^{V_i} \hat{a}_i^{-1} \int_{L_i}^{R_i} \int_0^{\tau} u_i(\beta_0, t|Z_i) \\ & \times \mathrm{d}M_i(t|X_i, Z_i) \, \mathrm{d}\hat{H}(X_i) \, \mathrm{d}\hat{G}(Z_i) \\ = & n^{-\frac{1}{2}} \sum_{i=1}^{n} b_i^{-1} \int_{U_i}^{V_i} a_i^{-1} \int_{L_i}^{R_i} \int_0^{\tau} u_i(\beta_0, t|Z_i) \\ & \times \mathrm{d}M_i(t|X_, Z_i) \, \mathrm{d}H(X_i) \, \mathrm{d}G(Z_i), \end{aligned}$$

which can be easily shown to converge in distribution to the normal distribution with mean zero and covariance matrix that can be consistently estimated by $\Gamma(\hat{\beta})$, where $a_i = \int_{L_i}^{R_i} \mathrm{d}H(x)$ and $b_i = \int_{U_i}^{V_i} \mathrm{d}G(z)$. This completes the proof. □

References

Andersen, P.K., Gill, R.D., 1982. Cox's regression model for counting processes: a large sample study. Ann. Statist. 10, 1100–1120.

Cox, D.R., 1972. Regression models and life-tables (with discussion). J. Roy. Statist. Soc. B 34, 187–220.

De Gruttola, V., Lagakos, S.W., 1989. Analysis of doubly-censored survival data, with application to AIDS. Biometrics 45, 1–12.

Gentleman, R.T., Geyer, C.J., 1994. Maximum likelihood for interval-censored data: consistency and computation. Biometrika 81, 618–623.

Goggins, W.B., Finkelstein, D.M., Zaslavsky, A.M., 1999. Applying the Cox proportional hazards model when the change time of a binary time-varying covariate is interval-censored. Biometrics 55, 445–451.

Gómez, G., Espinal, A., Lagakos, S.W., 2000. Inference for a linear regression model with an interval censored covariate. *Technical Report*, Department d'Estadística i I.O., Universitat Politècnica de Catalunya.

Kalbfleisch, J.D., Prentice, R.L., 1980. The Statistical Analysis of Failure Time Data, Wiley, New York.

Kim, M.Y., De Gruttola, V.G., Lagakos, S.W., 1993. Analyzing doubly censored data with covariates, with application to AIDS. Biometrics 49, 13–22.
Sun, J., 1998. Interval censoring. Encyclopedia of Biostatistics. Wiley, New York, pp. 2090–2095.
Sun, J., Liao, Q., Pagano, M., 1999. Regression analysis of doubly censored failure time data with applications to AIDS studies. Biometrics 55, 909–914.
Turnbull, B.W., 1976. The empirical distribution function with arbitrarily grouped censored and truncated data. J. Roy. Stat. Soc. B 38, 290–295.
Yu, Q., Schick, A., Li, L., Wong, G.Y.C., 1998. Asymptotic properties of the GMLE with case 2 interval-censored data. Statist. Probab. Lett. 44, 223–228.