Abstract

In longitudinal studies, the additive hazard model is often used to analyze covariate effects on the duration time, defined as the elapsed time between the first and second event. In this article, we consider the situation when the first event suffers partly interval-censoring and the second event suffers left-truncation and right-censoring. We proposed a two-step estimation procedure for estimating the regression coefficients of the additive hazards model. A simulation study is conducted to investigate the performance of the proposed estimator. The asymptotic properties of the proposed estimators are discussed.

Key Words: Interval-censoring, Left truncation, Additive hazards models.

Contents

Chapter 1

Introduction

In longitudinal studies, the interest often relies on the duration time, i.e. the elapsed time between the first and second events. In many situations, the first event suffers interval-censoring and the second event suffers right-censoring, the so-called doubly censored data (see Gómez and Calle (1999)). In some cases, the second event also suffers left-truncation. For example, in epidemiology, a prevalent cohort is defined as a group of diseased individuals who are recruited for a prospective study. Suppose that the disease population in a certain area is a representative sample from a large disease population. The target interest of a research project is to study the natural history of the disease for individuals who developed the disease during the calendar time period (τ_0, τ), $\tau_0 < \tau$. Let X denote the calendar time of the initial time of the first event. Let S denote the calendar time of the second event. One major interest is the estimation of the distribution function (denoted by F) of the duration time $T = S - X$, i.e. the elapsed time between the two events. Consider the sampling under which all of the individuals in the area who have experienced a first event (such as the origin of Alzheimer's disease or human immunodeficiency (HIV) infection) between τ_0 and τ and have not experienced a second event (such as death or the acquired immune deficiency syndrome (AIDS)) are recruited at the time τ for a prospective follow-up study. Hence, left truncation occurs since only subjects with $S \geq \tau$ can become part of the sample. In other words, only subjects with duration time $S - X = T \geq V = \tau - X$ can become part of the sample. Furthermore, suppose the follow-up study is terminated at τ^* ($\tau^* > \tau$). Then, the second event S is rightly censored by $C = \min(C_1, \tau^*)$, where C_1 denotes the calendar time of drop-out. Assume for each individual, data is available on a $(p+1) \times 1$ vector of covariates, $Z = [1, Z_1, \ldots, Z_p]^T$. It is important to investigate the association between Z and survival function of T. Suppose that the left and right endpoints of T are independent of Z. Let a_F and b_F denote the left and right endpoints of F, and similarly, define (a_G, b_G) and (a_Q, b_Q) as the left and right endpoint of G, and Q, respectively. For identifiabilities of F, we need the following condition (see Woodroofe (1985)):

$$a_G \leq \min(a_F, a_Q) \text{ and } b_F \leq b_Q. \tag{1.1}$$

Since $V = \tau - X$ and $T = Y - X$, we have $0 = a_G \leq a_F$. Furthermore, since $P(C_1 = \infty) > 0$, we have $b_Q = \tau^* - \tau_0$. Hence, if follow-up is sufficiently long then $b_F < b_Q$, we may assume that condition (1.1) holds. When X is always observable, the data is the so-called left-truncated and right-censored (LTRC) data. Assume that given Z, T and V are independent of each other but V and C are dependent with $P(C \geq V) = 1$. Cox's proportional hazards model (1972) has so far been the most popular model for the regression analysis of censored survival data. However, sometimes the proportional hazards model may not fit failure time data well. In contrast to the proportional hazard model, the additive risk model (Aalen (1980)) specifies that the hazard function associated with Z is the sum of the baseline hazard function and the regression function of Z, i.e. the effect of Z on T can be formulated through the following additive hazards model:

$$\lambda(t; \beta | Z) = \lambda_0(t) + Z^T \beta, \tag{1.2}$$

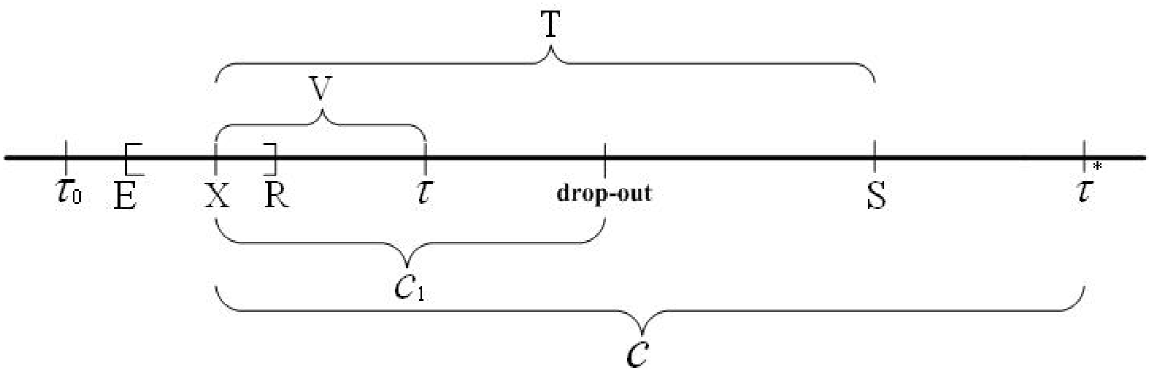


Figure 1. Schematic depiction of left-truncated and interval-censored data

where $\lambda_0(t)$ is an unknown baseline hazard function, β is a $(p+1)\times 1$ unknown vector of parameters, and $\lambda(\cdot)$ is an unspecified hazard function. In many situations, we only know that X belongs to an interval, say $[E, R]$, i.e. X is interval-censored, e.g. the infection of HIV (or the origin of Alzheimer's disease) can only be determined retrospectively to lie in some intervals. In this case, one can only observe $([E, R], \delta, Y, Z)$ if $S \geq \tau$ (i.e $T \geq V$), where $Y = \min(C, S)$, and $\delta = I\{Y = S\} = 1$ if $Y = S$ and $\delta = 0$ if $Y = C$. Figure 1 highlights all the different times for left-truncated and interval-censored data described above.

We briefly review the existing related literatures. When there is no truncation and covariate, De Gruttola and Lagakos (1989) proposed a method for analyzing discrete doubly censored survival data (i.e. both X and T are discrete) in the context of the study of the progression from HIV infection to AIDS. They estimated F by treating the data as a special type of bivariate survival data. Gómez and Lagakos (1994) pointed out that some practical problems with the De Gruttola and Lagakos (DGL) method related to the bivariate nature of the data were observed, which range from problems of convergence and speed of convergence to non-identifiability problems. To overcome these difficulties, they proposed an alternative methodology based on maximizing two univariate likelihood functions. The Gómez-Lagakos (GL) developed a two-step estimation procedure and provided an algorithm that is generally more stable and converges faster than does the DGL algorithm. When there is truncation and both X and T are discrete, Sun (1997) proposed a two-step nonparametric estimation procedure for the estimation of F. When there is no truncation, Gómez and Calle (1999) proposed a generalization of the Gómez and Lagakos two-step method for the case where both X and T are continuous. Zhao, Lim and Sun (2005) discussed statistical inference for the proportional hazards model when there exists interval-censoring on both survival time of interest and covariates.

In this article, we consider the situation when X suffers partly interval censoring, S suffers left-truncation and right-censoring, and both X and T are continuous. The data where X is partly interval censored arise often in follow-up studies. For example, for some individuals, time of infection of HIV, can be recorded exactly (e.g. infection due to blood transfusion) or Alzheimer's disease or vascular dementia onset is provided by the caregivers of those patients. But for others, time X is recorded only between two clinical examinations. In Section 2, we propose a two-step estimation procedure for estimating β of the proportional model (1.2) under partly interval-censored and truncated data. In Section 3, a simulation study is conducted to investigate the performance of the proposed estimator.

Chapter 2

The Proposed Estimator

Supposed there are n independent truncated subjects in a follow-up study and both Y and X are continuous. Let $([E_i, R_i], Y_i, \delta_i, Z_i)$ $(i = 1, \ldots, n)$ denote the truncated sample. Suppose that $E_i = R_i = X_i$, i.e. X is always observed. Then, we have LTRC data on the $T_i = Y_i - X_i$ and $V_i = \tau - X_i$. In this case, the estimate for β can be obtained by setting the following estimating equation equal to zero (see Lin and Ying 1993):

$$U(\beta|X_i's, Z_i's) = \sum_{i=1}^{n} \int_0^{\tau^*} [\{Z_i - \bar{Z}(x)]\{dN_i(x|X_i) - R_i(x|X_i) Z_i^T \beta dx\}, \tag{2.1}$$

where $\bar{Z}(x) = \sum_{i=1}^n R_i(x|X_i)Z_i / \sum_{i=1}^n R_i(x|X_i)$, $R_i(x|X_i) = I\{V_i < x \leq T_i\}$ and $N_i(x|X_i) = I\{T_i \leq x, \delta_i = 1\}$. Notice that $N_i(x|X_i)$ and $R_i(x|X_i)$ are written as function of X_i since both T_i and V_i depend on X_i. The resulting estimator takes the explicit form

$$\hat{\beta} = \left[\sum_{i=1}^{n} \int_0^{\tau^*} R_i(x|X_i)\{Z_i - \bar{Z}(x)\}\{Z_i - \bar{Z}(x)\}^T dx\right]^{-1} \left[\sum_{i=1}^{n} \int_0^{\tau^*} \{Z_i - \bar{Z}(x)\} dN_i(x|X_i)\right].$$

For partly interval-censored and truncated data, for some individuals, we only observe $X_i \in [E_i, R_i]$, where $E_i < R_i$. Without loss of generality, let $(E_i = R_i, Y_i, \delta_i, Z_i)$, $(i =$

$1, \ldots, n_1)$ denote the observations with $E_i = R_i$ and let $(E_i < R_i, Y_i, \delta_i, Z_i)$, $(i = n_1 + 1, \ldots, n)$ denote the partly interval-censored observation. Let $H(x) = P(X \leq x)$ denote the cumulative distribution functions of X.

Next, we shall propose an iterative algorithm for simultaneously estimating $F(t|Z) = P(T \leq t|Z)$ and H. The idea used in the following iterative algorithm is similar to that used in Gómez and Calle (1999), who considered the case when there is no truncation and covariates. In Step 0, using the observation with $E_i = R_i$, we first obtain an initial estimator of $F(t|Z_i)$, denoted by $F^{(0)}(t|Z_i)$. Based on $\hat{F}^{(0)}$, we obtain an estimated likelihood for H, denoted by $L(H, \hat{F}^{(0)})$. In Step 2, using the estimated likelihood $L(H, \hat{F}^{(0)})$, we obtain a first-step estimator of $H(x)$, denoted by $H^{(1)}(x)$. Given $H^{(1)}(t)$, in Step 2, we obtain a first-step estimator of $\hat{F}^{(1)}$ based on approximation of estimated equation. Iterate between Step 1 and Step 2 until convergence.

Step 0: Obtain an initial estimator of β

Based on the data $(V_i, T_i, \delta_i, Z_i)$ $(i = 1, \ldots, n_1)$ and (2.1), we can obtain an initial estimator of β, say $\hat{\beta}^{(0)}$. Note that the counting process $N_i(\cdot)$ can be uniquely decomposed by the relationship $N_i(x|X_i) = M_i(x|X_i) + \int_0^x R_i(u|X_i) d\Lambda(u; \beta|Z_i)$, where $M_i(\cdot|X_i)$ is a local square integrable martingale and $\Lambda(u; \beta|Z_i)$ is the cumulative hazard function. In view this relationship, using $\hat{\beta}^{(0)}$, we estimate $\Lambda_0(x) = \int_0^x \lambda_0(u) du$ by

$$\hat{\Lambda}_0^{(0)}(x) = \int_0^x \sum_{i=1}^n \frac{\{dN_i(u|X_i) - R_i(u|X_i) Z_i^T \hat{\beta}^{(0)} du\}}{\sum_{j=1}^n R_j(u|X_j)}.$$

Based on $\hat{S}_0^{(0)}(t) = e^{-\hat{\Lambda}_0^{(0)}(t)}$, we have an initial estimator of the distribution function of T given Z_i as $F^{(0)}(t|Z_i) = 1 - [\hat{S}_0^{(0)}(t|Z_i) e^{-\int_0^t Z_i^T \hat{\beta}^{(0)} du}]$.

Step 1: Obtain a first-step estimator of H

Given $\hat{F}^{(0)}$ and $[E_i, R_i]$, consider the following estimated conditional likelihood:

$$L(H, \hat{F}^{(0)}) = \prod_{i=1}^{n} P(X \in [E_i, R_i])/P(X \in [\tau - T, \infty), \tau - T \leq E_i)$$

$$= \prod_{i=1}^{n} \frac{\sum_{k=1}^{m} I\{\tau - u_k \leq E_i\}[H(R_i) - H(E_i)]\hat{F}^{(0)}(du_k|Z_i)}{\sum_{k=1}^{m} I\{\tau - u_k \leq E_i\}[1 - H(\tau - u_k)]\hat{F}^{(0)}(du_k|Z_i)}, \quad (2.2)$$

where $\hat{F}^{(0)}(du|Z_i) = \hat{F}^{(0)}(u|Z_i) - \hat{F}^{(0)}(u - |Z_i)$. We define the following two sets $E_S = \{E_i : i = 1, \ldots, n\}$ and $R_S = \{R_i : i = 1, \ldots, n\} \cup \{\tau - u_k : k = 1, \ldots, m\}$. Examination of (2.2) indicates that the estimated likelihood $L(H, \hat{F}^{(0)})$ will be maximized when the values $H(x)$ are as large as possible for $x \in R_S$ and as small as possible for $x \in E_S$. Accordingly, we construct a set Q as a union of disjoint closed intervals whose left and right end points lie in the set E_S and R_S, respectively, which contain no other members of E_S and R_S, and are covered by at least one censoring set. Let these Turnbull intervals (see Turnbull (1976) and Frydman (1994)) be written as $[q_1, p_1]$, $[q_2, p_2], \ldots, [q_J, p_J]$, where $q_1 \leq p_1 < q_2 \leq p_2 < \cdots < q_J \leq p_J$(See Appendix 1.). By Lemma 1 of Alioum and Commenges (1996), the likelihood function $L(H; \hat{F}^{(0)})$ can be written as

$$L(H; \hat{F}^{(0)}) = \prod_{i=1}^{n} \frac{\sum_{k=1}^{m} I\{\tau - u_k \leq E_i\} \sum_{j=1}^{J} s_j I\{[q_j, p_j] \subset [E_i, R_i]\}\hat{F}^{(0)}(du_k|Z_i)}{\sum_{k=1}^{m} I\{\tau - u_k \leq E_i\} \sum_{j=1}^{J} s_j I\{[q_j, p_j] \subset [\tau - u_k, \infty)\}\hat{F}^{(0)}(du_k|Z_i)}$$

$$= \prod_{i=1}^{n} \frac{\sum_{j=1}^{J} \hat{\alpha}_{ij} s_j}{\sum_{j=1}^{J} \hat{\beta}_{ij} s_j},$$

where $s_j = H(p_j) - H(q_j -)$, $\hat{\alpha}_{ij} = \sum_{k=1}^{m} I\{[\tau - u_k \leq E_i]\}I\{[q_j, p_j] \subset [E_i, R_i]\}\hat{F}^{(0)}(du_k|Z_i)$ and $\hat{\beta}_{ij} = \sum_{k=1}^{m} I\{[\tau - u_k \leq E_i]\}I\{[q_j, p_j] \subset [\tau - u_k, \infty)\}\hat{F}^{(0)}(du_k|Z_i)$.

Using (2.2), the logarithm of the likelihood can be expressed as

$$\log\{L(H)\} = \sum_{i=1}^{n}\left\{\log\left(\sum_{j=1}^{J} \hat{\alpha}_{ij} s_j\right) - \log\left(\sum_{j=1}^{J} \hat{\beta}_{ij} s_{ij}\right)\right\}. \quad (2.3)$$

By Lemmas 1 and 2 of Alioum and Commenges (1996), given $\hat{F}^{(0)}(du_k|Z_i)$, the problem of maximizing (2.2) is then equivalent to that of maximizing (2.3) subject to the constraints $\sum_{j=1}^{J} s_j = 1$ and $s_j \geq 0$ $(1 \leq j \leq J)$. To find the maximum likelihood estimate of the vector $s = (s_1, \ldots, s_J)^T$, one can use the self-consistency algorithm of Turnbull (1976) as follows:

$$s_j^{(b)} = \left\{ 1 + \frac{d_j(s^{(b-1)})}{M(s^{(b-1)})} \right\} s_j^{(b-1)} \ (1 \leq j \leq J), \tag{2.4}$$

where

$$d_j(s^{(b-1)}) = \sum_{i=1}^{n} \left\{ \left(\hat{\alpha}_{ij} \Big/ \sum_{k=1}^{J} \hat{\alpha}_{ik} s_k^{(b-1)} \right) - \left(\hat{\beta}_{ij} \Big/ \sum_{k=1}^{J} \hat{\beta}_{ik} s_k^{(b-1)} \right) \right\},$$

and

$$M(s^{(b-1)}) = \sum_{i=1}^{n} \frac{1}{\sum_{j=1}^{J} \hat{\beta}_{ij} s_j^{(b-1)}}.$$

Let $\hat{s}_j$ $(j = 1, \ldots, J)$ denote the first-step estimators obtained from (2.4). Based on $\hat{s}_j$, we have $\hat{H}^{(1)}(x) = 0$ if $x < q_1$; $\hat{H}^{(1)}(x) = \hat{s}_1 + \cdots + \hat{s}_j$ if $p_j < x < q_{j+1}$; $\hat{H}^{(1)}(x) = 1$ if $x > p_J$, and undefined for $x \in [q_j, p_j]$, for $1 \leq j \leq J$.

Step 2: Obtain an improved estimator of F

Given the first-step estimators $\hat{H}^{(1)}(t)$ and $[E_i, R_i]$, similar to the approach of Zhao et al. (2005), we propose the following simple Monte Carlo method:

Assume that $\hat{H}^{(1)}$ puts mass uniformly on $[q_j, p_j]$ $(j = 1, \ldots, J)$. For each $b = 1, \ldots, B$ and $i = 1, \ldots, n$, randomly sample $X_i^{(b)}$ from $\hat{H}^{(1)}(x)$ conditional on observed interval $[E_i, R_i]$. Let $X^{(b)} = (X_1^{(b)}, \ldots, X_n^{(b)})$(See Appendix 2.). Solve the equation

$$B^{-1} \sum_{b=1}^{B} U(\beta|Z_i's, X_i^{(b)'}s) = 0. \tag{2.5}$$

If B is large, we should expect that the left-hand side of equation (2.5) will give good approximation to

$$U(\beta, \hat{H}^{(1)}) = \left(\prod_{i=1}^{n} \frac{1}{\int_{E_i}^{R_i} \hat{H}^{(1)}(dx)}\right) \int_{E_1}^{R_1} \cdots \int_{E_n}^{R_n} U(\beta | Z_i's, X_i's) \prod_{i=1}^{n} \hat{H}^{(1)}(dX_i), \tag{2.6}$$

where X_i denote a random sample from $\hat{H}^{(1)}(x)$. Let $\hat{\beta}^{(1)}$ denote the solution of equation (2.5). Based on $\hat{\beta}^{(1)}$, we obtain the following estimator of the cumulative baseline hazard function:

$$\hat{\Lambda}_0^{(1)}(t) = B^{-1} \sum_{b=1}^{B} \int_0^t \frac{\sum_{i=1}^n dN_i(x|X_i^{(b)})}{\sum_{j=1}^n R_j(x|X_j^{(b)}) e^{\hat{\beta}^{(1)} Z_j}}.$$

When B is large, we should expect that $\hat{\Lambda}_0^{(1)}(t)$ will give good approximation to

$$\left(\prod_{i=1}^{n} \frac{1}{\int_{E_i}^{R_i} \hat{H}^{(1)}(dx)}\right) \int_{E_1}^{R_1} \cdots \int_{E_n}^{R_n} \int_0^t \frac{\sum_{i=1}^n dN_i(x|X_i)}{\sum_{j=1}^n R_j(x|X_j)) e^{\hat{\beta}^{(1)} Z_j}} \prod_{i=1}^{n} \hat{H}^{(1)}(dX_i).$$

Based on $\hat{S}_0^{(1)}(t) = e^{-\hat{\Lambda}_0^{(1)}(t)}$, we have a improved estimator of the distribution function of T given Z_i as $F^{(1)}(t|Z) = 1 - [\hat{S}_0^{(1)}(t|Z)]e^{\hat{\beta}^{(1)} Z}$. Replacing $\hat{F}^{(0)}$ in Step 1 with the improved estimator $\hat{F}^{(1)}$, we can obtain an improved estimator of H, say $\hat{H}^{(2)}$. Replacing $\hat{H}^{(1)}$ in Step 2 with $\hat{H}^{(2)}$, we can obtain an improved estimator $\hat{F}^{(2)}$. After Step 2, go back to Step 1 and repeat this cycle until convergence. Let $\hat{\beta}$ and $\hat{\Lambda}_0(t)$ denote the converged solutions.

Next, we briefly discuss the asymptotic properties of $\hat{\beta}$. To obtain the asymptotic normality of $\hat{\beta}$, we need to show the consistency of the first-step estimator $\hat{H}^{(1)}$, i.e. the consistency of the maximum likelihood estimate of equation (2.2). Hudgens (2005) (see Theorem 1, page 578) proposed a sufficient and necessary condition for the existence of maximum likelihood estimate of equation (2.3). However, large sample properties of the maximum likelihood estimate remain unknown. Although asymptotic properties of maximum likelihood estimate for interval-censored data have been derived (see Groeneboom and Wellner (1992), Shick and Yu (1998), Yu et al. (1998a,b)), much less is known about the large sample properties of the maximum likelihood estimate if both interval censoring and truncation are present.

Notice that (2.6) can be written as

$$n^{-1/2}U(\beta, \hat{H}^{(1)}) = \int_{E_1}^{R_1} \cdots \int_{E_n}^{R_n} \left\{ n^{-1/2} \sum_{i=1}^{n} \int_0^{\tau^*} [Z_i - \bar{Z}(x)] dM_i(x|X_i) \right\} \prod_{i=1}^{n} \frac{\hat{H}^{(1)}(dX_i)}{\hat{h}_i}, \quad (2.7)$$

where $\hat{h}_i = \int_{E_i}^{R_i} \hat{H}^{(1)}(dx)$ $(i = 1, \ldots, n)$. If the consistency of $\hat{H}^{(1)}$ holds, it follows that (2.7) converges in distribution to a $(p+1)$-variate normal with mean zero and with a covariance matrix which can be consistently estimated by

$$\hat{\Sigma} = \int_{E_1}^{R_1} \cdots \int_{E_n}^{R_n} \left\{ n^{-1} \sum_{i=1}^{n} \int_0^{\tau^*} [Z_i - \bar{Z}(x)][Z_i - \bar{Z}(x)]^T dN_i(x|X_i) \right\} \prod_{i=1}^{n} \frac{\hat{H}^{(1)}(dX_i)}{\hat{h}_i}.$$

Furthermore, the random vector $n^{1/2}(\hat{\beta} - \beta)$ converges in distribution to a $(p+1)$-variate normal with mean zero and with a covariance matrix which can be consistently estimated by $A(\hat{\beta})\hat{\Sigma}A^T(\hat{\beta})$, where $A(\beta) = \{-n^{-1}\partial U(\beta, \hat{H}^{(1)})/\partial\beta\}^{-1}$.

Chapter 3

A Simulation Study

A simulation study is conducted to investigate the performance of the proposed estimator. Suppose the target interest of a research project is to study the natural history of the disease for individuals who developed a certain disease before the Taiwan's Republican calendar (TRC) time 93 (i.e. calendar time 2004, TRC time =calendar time - 1911). Consider the sampling under which all of the individuals who were diagnosed with a certain disease and are still alive are recruited at the TRC time $\tau = 93$ for a prospective follow-up study. Furthermore, suppose the follow-up study is terminated at TRC time $\tau^* = 103$. The variable $V = 93 - X$ is exponential distribution: $G(x; \beta_g) = 1 - e^{-x/\beta_g}$, which implies that the incidence rate of the disease grows due to some reasons (e.g. eating habit change). The values of β_g are set at $\beta_g = 5, 10$. Given V, let $X = 93 - V$ be the TRC time of having the disease. To make X partly interval-censored, we generate two independent uniform random variables U_1 and U_2. If $U_1 > I_C$ then $X = E = R$. If $U_1 \leq I_C$ and $U_2 \leq 0.5$ then $E = X - c$ and $R = X + c + 2$. If $U_1 \leq I_C$ and $U_2 > 0.5$ then $E = X - (c + 2)$ and $R = X + c$. The values of I_C are set at $I_C = 0.4, 0.6$ and the values of c are set at $c = 0.5, 1.0$. Next, we consider a two-sample additive hazards model. The covariate $Z = 0$ represents sample 1, and $Z = 1$ represents sample 2. For each sample, the sample size is chosen as $n_1 = n_2 = 50; 100$.

Hence, the total sample size is set at $n = 100; 200$. For sample 1, T is generated with is generated with population hazard density $\lambda_1(t) = \lambda_0(t) = \alpha t$ for $t > 0$, and for sample 2, T is generated with hazard density $\lambda_2(t) = \lambda_0(t) + \beta$ for $t > 0$. The values of parameter α and β are set at 0.004 and 0.04, respectively. Given T, let $S = T + X$ be the TRC time of death and $C = \tau^* = 103$ be the TRC time of censoring. Let $Y = \min(C, S)$ and $\delta = I\{Y = S\}$ as defined in Section 1. Only subjects with duration time $T \geq V$ (or $S \geq 93$) can become part of the sample. The replication is 100 times. Table 1 shows the biases, standard deviations (std.) and root mean squared errors (rmse) of the estimator $\hat{\beta}$. For purpose of comparison, we also include the biases, std. and rmse of the estimator (denoted by $\tilde{\beta}$) based solely on uncensored data. Tables 2 through 4 show the biases, standard deviations (std.) and root mean squared errors (rmse) of the estimators $\hat{\Lambda}_0(t_p)$ with $p = 0.2, 0.5, 0.8$, where t_p denotes the p^{th} quantile of T. The values of $\Lambda_0(t_p)$ are equal to 0.22, 0.69 and 1.61 for $p = 0.2$, 0.5 and 0.8, respectively. For purpose of comparison, we also include the biases, std. and rmse of the estimator (denoted by $\tilde{\Lambda}_0(t_p)$) based solely on uncensored data. Tables 1 through 4 also list the proportion of truncation $P(T < V) = P(S < 93)$ (denoted by q_T), the proportion of right censoring $P(\delta_i = 0)$ (denoted by p_C) and the proportion of interval censoring $P(E_i < R_i)$ (denoted by p_I). Based on the results of Tables 1 through 4, we conclude that:

1. For the estimation of β:

(i) In terms of rmse, the estimator $\hat{\beta}$ outperforms the estimator $\tilde{\beta}$ for all the cases considered.

(ii) Given p_T and c, the rmse of both estimators $\hat{\beta}$ and $\tilde{\beta}$ increase as the proportion of interval censoring (i.e. p_I) increases.

(iii) Given I_C and c, the rmse of both estimators increase as the proportion of truncation (i.e. p_T) increases.

2. For the estimation of $\Lambda_0(t_p)$:

(i) In terms of rmse, the estimator $\hat{\Lambda}_0(t_p)$ outperforms the estimator $\tilde{\Lambda}_0(t_p)$ for all the cases considered.

(ii) Given p_T and c, the rmse of both estimators $\hat{\Lambda}_0(t_p)$ and $\tilde{\Lambda}_0(t_p)$ increase as the proportion of interval censoring (i.e. p_I) increases.

(iii) Given I_C and c, the rmse of both estimators increase as the proportion of truncation (i.e. p_T) increases.

Table 1. Simulation results for bias, standard deviation and

root mean squared error of $\tilde{\beta}$ and $\hat{\beta}$

							$\tilde{\beta}$			$\hat{\beta}$		
β_g	I_C	c	n	p_T	p_C	p_I	bias	std	rmse	bias	std	rmse
5	0.4	0.5	100	0.40	0.53	0.40	-0.001	0.016	0.016	-0.002	0.011	0.011
5	0.4	0.5	200	0.40	0.53	0.40	-0.002	0.011	0.011	-0.001	0.008	0.008
5	0.4	1.0	100	0.40	0.53	0.40	-0.002	0.018	0.018	-0.002	0.012	0.012
5	0.4	1.0	200	0.40	0.53	0.40	-0.005	0.012	0.013	-0.003	0.008	0.009
5	0.6	0.5	100	0.40	0.53	0.60	-0.002	0.019	0.019	-0.002	0.012	0.012
5	0.6	0.5	200	0.40	0.53	0.60	-0.003	0.014	0.014	-0.001	0.008	0.008
5	0.6	1.0	100	0.40	0.53	0.60	-0.005	0.019	0.019	-0.003	0.012	0.013
5	0.6	1.0	200	0.40	0.53	0.60	-0.004	0.014	0.014	-0.003	0.010	0.010
10	0.4	0.5	100	0.66	0.53	0.40	-0.006	0.017	0.018	-0.003	0.012	0.012
10	0.4	0.5	200	0.66	0.53	0.40	-0.001	0.012	0.012	-0.000	0.009	0.009
10	0.4	1.0	100	0.66	0.53	0.40	-0.002	0.021	0.021	-0.002	0.015	0.015
10	0.4	1.0	200	0.66	0.53	0.40	-0.001	0.017	0.017	-0.001	0.009	0.009
10	0.6	0.5	100	0.66	0.53	0.60	-0.004	0.022	0.023	-0.003	0.014	0.014
10	0.6	0.5	200	0.66	0.53	0.60	-0.001	0.013	0.013	-0.002	0.009	0.009
10	0.6	1.0	100	0.66	0.53	0.60	0.000	0.022	0.022	-0.001	0.016	0.016
10	0.6	1.0	200	0.66	0.53	0.60	-0.003	0.016	0.016	-0.002	0.010	0.010

Table 2. Simulation results for bias, standard deviation and root mean squared error of $\tilde{\Lambda}_0(t_{0.2})$ and $\hat{\Lambda}_0(t_{0.2})$

							$\tilde{\Lambda}_0(t_{0.2})$			$\hat{\Lambda}_0(t_{0.2})$		
β_g	I_C	c	n	p_T	p_C	p_I	bias	std	rmse	bias	std	rmse
5	0.4	0.5	100	0.40	0.53	0.40	0.037	0.114	0.120	0.017	0.076	0.077
5	0.4	0.5	200	0.40	0.53	0.40	0.016	0.074	0.076	0.012	0.063	0.064
5	0.4	1.0	100	0.40	0.53	0.40	-0.023	0.097	0.100	0.009	0.078	0.078
5	0.4	1.0	200	0.40	0.53	0.40	0.016	0.069	0.072	0.008	0.047	0.048
5	0.6	0.5	100	0.40	0.53	0.60	0.022	0.120	0.121	0.026	0.077	0.081
5	0.6	0.5	200	0.40	0.53	0.60	0.017	0.091	0.092	0.012	0.055	0.056
5	0.6	1.0	100	0.40	0.53	0.60	0.006	0.113	0.113	0.024	0.077	0.080
5	0.6	1.0	200	0.40	0.53	0.60	0.018	0.079	0.080	0.018	0.058	0.061
10	0.4	0.5	100	0.66	0.53	0.40	0.018	0.140	0.142	0.011	0.095	0.095
10	0.4	0.5	200	0.66	0.53	0.40	0.015	0.106	0.107	0.010	0.073	0.074
10	0.4	1.0	100	0.66	0.53	0.40	0.024	0.162	0.164	0.020	0.106	0.108
10	0.4	1.0	200	0.66	0.53	0.40	-0.014	0.097	0.098	-0.005	0.073	0.073
10	0.6	0.5	100	0.66	0.53	0.60	0.028	0.155	0.158	0.018	0.118	0.119
10	0.6	0.5	200	0.66	0.53	0.60	0.010	0.138	0.138	0.016	0.119	0.120
10	0.6	1.0	100	0.66	0.53	0.60	0.031	0.165	0.168	0.031	0.147	0.150
10	0.6	1.0	200	0.66	0.53	0.60	0.017	0.107	0.109	0.013	0.075	0.076

Table 3. Simulation results for bias, standard deviation and root mean squared error of $\tilde{\Lambda}_0(t_{0.5})$ and $\hat{\Lambda}_0(t_{0.5})$

							$\tilde{\Lambda}_0(t_{0.5})$			$\hat{\Lambda}_0(t_{0.5})$		
β_g	I_C	c	n	p_T	p_C	p_I	bias	std	rmse	bias	std	rmse
5	0.4	0.5	100	0.40	0.53	0.40	0.074	0.203	0.216	0.029	0.144	0.147
5	0.4	0.5	200	0.40	0.53	0.40	0.013	0.144	0.144	0.020	0.104	0.104
5	0.4	1.0	100	0.40	0.53	0.40	0.037	0.191	0.194	0.027	0.148	0.151
5	0.4	1.0	200	0.40	0.53	0.40	0.050	0.142	0.151	0.022	0.108	0.110
5	0.6	0.5	100	0.40	0.53	0.60	0.077	0.274	0.285	0.056	0.178	0.186
5	0.6	0.5	200	0.40	0.53	0.60	0.038	0.202	0.207	-0.025	0.117	0.119
5	0.6	1.0	100	0.40	0.53	0.60	0.041	0.254	0.258	0.040	0.168	0.173
5	0.6	1.0	200	0.40	0.53	0.60	0.060	0.181	0.190	0.032	0.112	0.117
10	0.4	0.5	100	0.66	0.53	0.40	0.060	0.256	0.263	0.045	0.178	0.184
10	0.4	0.5	200	0.66	0.53	0.40	0.024	0.173	0.175	0.008	0.115	0.115
10	0.4	1.0	100	0.66	0.53	0.40	0.030	0.250	0.251	0.037	0.158	0.162
10	0.4	1.0	200	0.66	0.53	0.40	0.029	0.135	0.138	0.018	0.112	0.113
10	0.6	0.5	100	0.66	0.53	0.60	0.087	0.293	0.306	0.044	0.205	0.209
10	0.6	0.5	200	0.66	0.53	0.60	0.031	0.175	0.178	0.026	0.140	0.142
10	0.6	1.0	100	0.66	0.53	0.60	0.055	0.258	0.264	0.028	0.199	0.201
10	0.6	1.0	200	0.66	0.53	0.60	0.003	0.192	0.192	0.015	0.123	0.124

Table 4. Simulation results for bias, standard deviation and root mean squared error of $\tilde{\Lambda}_0(t_{0.8})$ and $\hat{\Lambda}_0(t_{0.8})$

							$\tilde{\Lambda}_0(t_{0.8})$			$\hat{\Lambda}_0(t_{0.8})$		
β_g	I_C	c	n	p_T	p_C	p_I	bias	std	rmse	bias	std	rmse
5	0.4	0.5	100	0.40	0.53	0.40	-0.054	0.946	0.947	0.022	0.723	0.723
5	0.4	0.5	200	0.40	0.53	0.40	0.115	0.603	0.614	0.072	0.409	0.415
5	0.4	1.0	100	0.40	0.53	0.40	-0.129	0.793	0.804	0.220	0.660	0.695
5	0.4	1.0	200	0.40	0.53	0.40	0.167	0.592	0.615	0.157	0.450	0.476
5	0.6	0.5	100	0.40	0.53	0.60	-0.185	1.115	1.131	0.139	0.779	0.791
5	0.6	0.5	200	0.40	0.53	0.60	0.328	0.783	0.849	-0.153	0.495	0.518
5	0.6	1.0	100	0.40	0.53	0.60	0.115	1.034	1.040	0.110	0.713	0.722
5	0.6	1.0	200	0.40	0.53	0.60	0.231	0.759	0.794	0.122	0.502	0.516
10	0.4	0.5	100	0.66	0.53	0.40	0.151	0.534	0.555	0.061	0.331	0.337
10	0.4	0.5	200	0.66	0.53	0.40	0.054	0.318	0.323	0.008	0.210	0.210
10	0.4	1.0	100	0.66	0.53	0.40	0.148	0.398	0.414	0.108	0.316	0.334
10	0.4	1.0	200	0.66	0.53	0.40	0.080	0.265	0.277	0.036	0.216	0.219
10	0.6	0.5	100	0.66	0.53	0.60	0.137	0.558	0.574	0.089	0.378	0.389
10	0.6	0.5	200	0.66	0.53	0.60	0.057	0.356	0.360	0.040	0.265	0.268
10	0.6	1.0	100	0.66	0.53	0.60	0.111	0.608	0.626	0.096	0.341	0.354
10	0.6	1.0	200	0.66	0.53	0.60	0.114	0.342	0.360	0.045	0.213	0.218

Chapter 4

Concluding Remarks

This article discusses the estimation of the regression coefficients of the additive hazards model when the first event suffers partly interval-censoring and the second event suffers left-truncation and right-censoring. Simulation results indicate that the proposed estimator performs well. No formal discussion of asymptotic properties of the proposed estimator is undertaken here. A topic for future research is the rigorous investigation of the asymptotic properties of the estimators we presented. This would require the establishment of the consistency of $\hat{H}^{(1)}$.

References

Aalen, O.O. (1980). A model for non-parametric regression analysis of counting process. In:Klonecki, W., Kozek, A., Rosiski, J. (Eds). Lecture Notes on Mathematical Statistics and Probability, Vol. 2. Springer, New York, pp. 1-25.

Alioum A. and Commenges D. 1996. A proportional hazards model for arbitrarily censored and truncated data. *Biometrics*, **52**, 512-524.

Andersen P. K., Borgan, O., Gill, R. D. & Keiding, N. (1993). Statistical models based on counting processes. New York: Springer-Verlag.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of Royal Statistical Society, Ser.B*, **34**, 187-220.

DeGruttola, V. and Lagakos, S. W., 1989. Analysis of doubly-censored survival data, with application to AIDS. *Biometrics*, **45**, 1-12.

Frydman, H. 1994. A note on nonparametric estimation of the distribution function from interval-censored and truncated data. *Journal of the Royal Statistical Society, Series B*, **56**, 71-74.

Gómez, G. and Lagakos, S. W., 1994. Estimation of the infection time and latency distribution of AIDS with doubly censored data. *Biometrics*, **50**, 204-212.

Gómez, G. and Calle, M. L., 1999. Non-parametric estimation with doubly censored data. *Journal of Applied Statistics*, **26**, 45-58.

Groeneboom, P. and Wellner, J. A., 1992. *Information Bounds and Nonparametric Maximum Likelihood Estimation.* Basel: Birkhäuser.

Hudgens, M. G., 2005. On nonparametric maximum likelihood estimation with interval censoring and truncation. *Journal of the Royal Statistical Society, Series B*, **67**, part 4, 573-587.

Kalbfleisch, J. D., Prentice, R. L., 1980. The statistical analysis of failure time data, Wiley , New York

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association*, **53**, 457-481.

Lai, T. Z., Ying, Z., 1991. Estimating a distribution function with truncated and censored data. *Ann. Statist.*, **19**, 417-442.

Lin, D. Y. and Ying, Z., 1994. Semiparametric analysis of the additive risk model. *Biometrika*, **81**, 61-71.

Shick, A and Yu, Q. 2000. Consistency of the GMLE with mixed case interval-censored data. *Scandinavian Journal of Statistics*, **27**, 45-55.

Sun, J. 1997. Self-consistency estimation of distributions based on truncated and doubly censored survival data with applications to AIDS cohort studies.

vskip 2ptTsai, W.-Y., Jewell, N. P. and Wang, M.-C., 1987. A note on the product-limit estimate under right censoring and left truncation. *Biometrika*, **74**, 883-886.

Turnbull, B. W., 1976. The empirical distribution function with arbitrarily grouped censored and truncated data. *Journal of the Royal Statistical Society, Series B*, **38**, 290-295.

Wang, M.-C., 1987, Product-limit estimates: a generalized maximum likelihood study. *Communication in Statistics., Part A- Theory and Methods*, **6**, 3117-3132.

Woodroofe, M., 1985. Estimating a distribution function with truncated data. *Annals of Statistics*, **13**, 163-167.

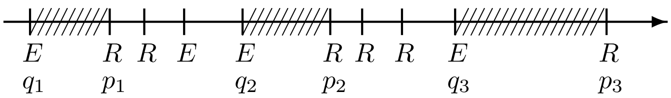
Yu, Q., Li, L. and Wong, G.Y.C., 1998a. Asymptotic variance of the GMLE of a survival function with interval censored data. *Sankhya*, **60**, 184-187.

Yu, Q., Shick, A., Li, L. and Wong, G.Y.C., 1998b. Asymptotic properties of the GMLE with case 2 interval-censored data. *Statistics* & *probability letters*, **37**, 223-228.

Zhao, X, Lim, H., Sun, J., 2005. Estimating equation approach for regression analysis of failure time data in the presence of interval-censoring. *Journal of statistical planning and inference*, **129**, 145-157.

Appendix

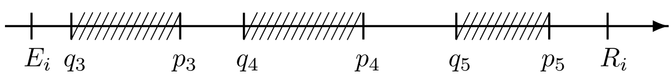
- 1. Example for Turnbull intervals:



- 2. To get $X^{(b)}$

  If $[E_i, R_i]$ contains k $[q_j, p_j]$ intervals$(j = m, \ldots, m+k-1)$, then we sample one of these intervals with each probability

$$\frac{s_j}{\sum_{j=m}^{m+k-1} s_j}$$



  After sampling, we generate a uniform random variable between the chosen interval, denoted by $X_i^{(b)}$