Statistica Sinica **13**(2003), 519-537

EFFICIENT ESTIMATION FOR THE PROPORTIONAL HAZARDS MODEL WITH LEFT-TRUNCATED AND "CASE 1" INTERVAL-CENSORED DATA

Jong S. Kim

Portland State University

Abstract: The maximum likelihood estimator (MLE) for the proportional hazards model with *left-truncated and "Case 1" interval-censored data* is studied. Under appropriate regularity conditions, the MLE of the regression parameter is shown to be asymptotically normal with a root-n convergence rate and achieves the information bound, even though the difference between left-truncation time and censoring time of the MLE of the baseline cumulative hazard function converges only at rate $n^{1/3}$. Two methods to estimate the variance-covariance matrix of the MLE of the regression parameter are considered. One is based on a generalized missing information principle and the other is based on the profile information procedure. Simulation studies show that both methods work well in terms of bias and variance for samples of moderate sizes. An example is provided to illustrate the methods.

Key words and phrases: Asymptotic distribution, left-truncated and "Case 1" interval-censored data, proportional hazards model, variance estimation.

1. Introduction

In many medical studies, we are interested in the relationship between a failure time and a covariate. However, failure times are subject to either truncation or censoring. According to Klein and Moeschberger (1997, Section 3.4), *truncation* is defined to be a condition which screens certain subjects so that the investigator will not be aware of their existence. For truncated data, only individuals who experience some event are observed by the investigator. The *event* may be some condition which must occur prior to the event of interest, such as exposure to a disease, entry into a retirement center, occurrence of an intermediate event prior to death (e.g., recurrence of leukemia prior to death), etc. In this case, the main event of interest is said to be *left-truncated*. The most common type of left-truncation occurs when subjects enter a study at random ages (not necessarily the origin for the event of interest) and are followed from this *delayed entry time* until the event of interest occurs or until the subject is right-censored. Andersen, Borgan, Gill and Keiding (1993) contains many examples of left-truncated data and statistical models based on them.

On the other hand, *censoring* occurs when we have some information about an individual lifetime, but we do not know the lifetime exactly. In particular, *interval censoring* occurs when the lifetime is known to occur only within an interval. Such interval censoring occurs when each patient in a clinical trial or longitudinal study has a periodic follow-up and the patient's event time is only known to lie in an interval $(L_i, R_i]$ (L for left endpoint and R for right endpoint of the censoring interval). Examples of interval-censored data can be found in animal carcinogenicity (Hoel and Walburg (1972)) and epidemiology studies (Finkelstein (1986)) among others. Huang and Wellner (1997) reviews recent progress in models based on interval-censored data.

In this paper, we suppose each failure time is only known to lie in an interval determined by an intermediate event time prior to the event of interest and an examination time. At the examination time, we only know if the event of interest has happened since the intermediate event time. We call this left-truncated and "Case 1" interval-censored (LTIC 1) data. Turnbull (1976) described a general scheme of incomplete failure time data, which includes LTIC 1 data as a special case.

The proportional hazards model (Cox (1972)) has been widely used for assessing the effects of covariates on survival time. For right-censored failure time data, inference can be made based on the partial likelihood of the combined ranks of the exact and the right-censored failure times. However, for LTIC 1 data, the observed intervals of failure overlap and vary in length. As a result, it may not be easy to identify the precise ranking of the failure times for study subjects.

We consider the maximum likelihood estimation approach for the proportional hazards model based on LTIC 1 data. A generalized Gauss-Seidel algorithm will compute the MLE. We show the consistency and the asymptotic normality of the MLE of the regression parameter. The asymptotic properties of the MLE with LTIC 1 data are different from those with "Case 1" interval-censored data alone. Notice the difference between LTIC 1 data and "Case 1" interval-censored data. In LTIC 1 data, we have a left-truncation time which is away from the start of a study. So we do not have any information near the start of the study. However, in "Case 1" interval-censored data, we have some information near the start of a study in case of a left-censored failure time because we know the failure time lies between the start of the study and the observed time. It is therefore expected that statistical inference with LTIC 1 data is more difficult. In particular, the consistency of the MLE of the baseline cumulative hazard function may not be an easy matter to prove. We were only able to prove the consistency of the difference between left-truncation time and censoring time of the MLE of the baseline cumulative hazard function because we do not have information near the start of the study. Nevertheless, we prove that it is a sufficient condition for the asymptotic normality of the MLE of the regression parameter.

In order to make a statistical inference it is essential to estimate the variance-covariance matrix of the estimator of the parameters of main interest. In the present model the number of parameters increases as the sample size does. In this case computation of the inverse of the high-dimensional observed information matrix may be numerically unstable. Therefore, we considered two methods that do not require inverting the large observed information matrix. One is to take advantage of sampling from the conditional distribution. Kim (1999) generalizes to a semiparametric setting the missing information principle for parametric models described in Louis (1982). The other method is to compute the inverse of the profile information matrix. Murphy and van der Vaart (1999) proved its appropriateness for i.i.d. samples. Based on the simulation results, it seems that the estimate from either method works well as an estimate of the Fisher information matrix.

The organization of the paper is as follows. In Section 2, we describe the model and consider estimation of the "true" regression parameter θ_0 and the baseline cumulative hazard function $\Lambda_0(\cdot)$. In Section 3, we compute the information matrix for θ_0. In Section 4, for the statistical inference for θ_0, we suggest two methods of estimating the variance-covariance matrix of the regression parameter estimator. In Section 5, we consider simulation studies. The cases with 15% and 36% of truncation proportion are treated. A brief summary of simulation procedures and results is provided. In Section 6, a data set illustrates the proposed methods. In addition, the effect of misspecifying the data as either "Case 1" or "Case 2" interval-censored data is considered. In Section 7, we show that under appropriate conditions, the MLE of θ_0 and the difference between the left-truncation time and the censoring time of the MLE of $\Lambda_0(\cdot)$ are consistent. We also show their rate of convergence. In Section 8, we show the MLE of θ_0 is asymptotically normal and efficient. Finally, we include concluding remarks in Section 9. Proofs are gathered together in the Appendix.

2. Model and Estimation

2.1. Model

In the proportional hazards model, the conditional hazard of a failure time T given a covariate $Z \in R^d$ is proportional to the baseline hazard, $\lambda(t|z) = \lambda_0(t)e^{\theta' z}$, where θ is a d-dimensional regression parameter and $\lambda_0(\cdot)$ is the baseline hazard function.

For each subject, there is a censoring indicator δ_i, a left-truncation time X_i, an examination time U_i, and a covariate Z_i. Consider an ith subject whose failure time T_i is left-truncated at X_i. If the ith subject tests positive at the examination time U_i, then T_i is left-truncated and left-censored. If the ith subject still tests negative at the examination time U_i, then T_i is left-truncated and right-censored. We assume that the true failure time is independent of the truncation time and the examination time given Z, and that the joint distribution of the truncation time, the examination time, and Z does not involve θ and $\Lambda_0(\cdot)$, the baseline cumulative hazard function. Suppose $Y_1, \ldots, Y_n$ is an i.i.d. sample from $Y = (\delta, X, U, Z)$, where $Y_i = (\delta_i, X_i, U_i, Z_i)$, $i = 1, \ldots, n$, with $\delta_i = 1$ if left-truncated and left-censored, and $\delta_i = 0$ if left-truncated and right-censored. Let $S_0(\cdot)$ be the baseline survival function of failure times. Then the joint likelihood function (up to a multiplicative constant) is

$$L = \prod_{i=1}^{n} [1 - \{S_0(U_i)/S_0(X_i)\}^{e^{\theta' Z_i}}]^{\delta_i} [\{S_0(U_i)/S_0(X_i)\}^{e^{\theta' Z_i}}]^{1-\delta_i}. \qquad (1)$$

Since $S_0(\cdot) = \exp\{-\Lambda_0(\cdot)\}$, the log-likelihood function can be written in terms of the cumulative hazard function as

$$\begin{aligned} l_n(\theta, \Lambda_0) = \sum_{i=1}^{n} & \delta_i \log[1 - \exp[-\{\Lambda_0(U_i) - \Lambda_0(X_i)\} e^{\theta' Z_i}]] \\ & -(1 - \delta_i)\{\Lambda_0(U_i) - \Lambda_0(X_i)\} e^{\theta' Z_i}. \end{aligned} \qquad (2)$$

2.2. Computation of MLE

Since the values of Λ_0 matter only at either left-truncated times or examination times in the log-likelihood function, we will take the MLE $\hat{\Lambda}_n$ of Λ_0 to be a right-continuous step function with possible jump points at X_i, U_i, $i = 1, \ldots, n$. Let $\Theta \subset \mathcal{R}^d$ be the finite-dimensional parameter space containing θ_0, the "true" regression parameter. The MLE $(\hat{\theta}_n, \hat{\Lambda}_n)$ maximizes $l_n(\theta, \Lambda)$ subject to $\theta \in \Theta$ and Λ being a right-continuous step function. We propose a generalized Gauss-Seidel algorithm to compute $(\hat{\theta}_n, \hat{\Lambda}_n)$. Let $\theta^{(0)}$ be a starting value and set the iteration counter $k = 0$. The algorithm then is the following.

(a) Maximize $l_n(\theta^{(k)}, \Lambda)$ with respect to Λ to obtain ${\Lambda_n}^{(k)}$.

(b) Maximize $l_n(\theta, {\Lambda_n}^{(k)})$ with respect to θ. Set $k \leftarrow k + 1$, and let ${\theta_n}^{(k)}$ be the maximizer.

(c) Repeat (a) and (b) until convergence.

In step (a), we first compute $\hat{\lambda}_n$ at X_i, U_i, $i = 1, \ldots, n$, the hazard size maximizing the log-likelihood as a function of θ and λ with θ being fixed. $\hat{\Lambda}_n$ is then the cumulative sum of the elements of $\hat{\lambda}_n$. It is clear that each iteration increases the likelihood. Hence the algorithm converges to at least a local maximum.

Step (a) is a constrained maximization problem since $\Lambda(\cdot)$ is restricted to be a right-continuous step function. In step (b), let $s_1(\theta) = (\partial/\partial\theta)l_n(\theta, \Lambda)$. By the strict concavity, the solution to $s_1(\theta) = 0$ is the unique maximizer of $l_n(\theta, \Lambda)$ for fixed Λ. In an iterated estimation procedure, it is common that the speed of convergence becomes slower as the estimates come closer to the limit point. In step (b), we generalize Louis' (1982) Newton-Raphson step for a parametric model to the current semiparametric setting. The following proposition implies that steps (a) and (b) are well-defined concave maximization problems.

Proposition 2.1. (1) *For any fixed θ, $l_n(\theta, \Lambda)$ is a strictly concave function of Λ; (2) for any fixed Λ, $l_n(\theta, \Lambda)$ is a strictly concave function of θ.*

3. Information Calculation

The following assumptions are needed for the information calculation, and for the proof of the asymptotic properties in later sections.

(A1) The finite-dimensional parameter space Θ is a bounded subset of R^d.

(A2) (a) There exists z_0 such that $|Z| \leq z_0$ with probability 1; (b) for any $\theta_1 \neq \theta_2 \in \Theta$, $P\{\theta_1' Z \neq \theta_2' Z\} > 0$.

(A3) There exists a positive number η such that $P(U - X \geq \eta) = 1$.

(A4) There exists $0 < \tau_0 < \tau_1$ and $0 < m_0 < M_0 < \infty$ such that $P(\tau_0 \leq X < U \leq \tau_1) = 1$ and $m_0 < \Lambda_0(\tau_0) < \Lambda_0(\tau_1) < M_0$.

(A5) Λ_0 is strictly increasing on $[\tau_0, \tau_1]$.

Assumptions (A1), (A2a), (A3), (A4) and (A5) are needed for the entropy calculation in Lemma 7.2, which is crucial for obtaining the rate of convergence and proving asymptotic normality of $\hat{\theta}_n$, the MLE of θ_0. Assumption (A2b) is imposed for the identifiability of θ_0. Note that if $\tau_0 = 0$ and $X = 0$ with probability 1, then (A4) reduces to "Case 1" interval-censored data considered in Huang (1996). Moreover, if $\tau_0 = 0$ and $P(X = 0)$ is between 0 and 1, then this reduces to a mixture of "Case 1" interval-censored data and LTIC 1 data. We focus only on those subjects who had an intermediate event prior to the event of interest, and that requires τ_0 should be strictly positive.

For the proportional hazards model with "Case 1" interval-censored data, Huang (1996) shows that the MLE of the regression parameter converges at root-n rate and is asymptotically efficient. A necessary condition is that we must have positive information. With LTIC 1 data, it is not clear that the information is, in fact, positive. Therefore, we first calculate the information for the regression parameter in the proportional hazards model with LTIC 1 data, and show that it is, indeed, positive under reasonable assumptions.

Define

$$Q(\delta, x, u, z) = \delta \frac{\exp[-\{\Lambda(u|z) - \Lambda(x|z)\}]}{1 - \exp[-\{\Lambda(u|z) - \Lambda(x|z)\}]} - (1 - \delta), \tag{3}$$

$$O(x,u|z) = E[Q^2(\delta,X,U,Z)|X=x,U=u,Z=z] = \frac{\exp[-\{\Lambda(u|z)-\Lambda(x|z)\}]}{1-\exp[-\{\Lambda(u|z)-\Lambda(x|z)\}]}, \tag{4}$$

where $\Lambda(\cdot|z) = \Lambda_0(\cdot)\exp(\theta' z)$ is the conditional cumulative hazard function given z. The following theorem closely follows Huang (1996).

Theorem 3.1. *Suppose that assumptions* (A2)–(A5) *are satisfied. Then*

(a) *The efficient score function for θ is*

$$\dot{l}^*_\theta(y) = r(\theta' z)Q(\delta,x,u,z)\{\Lambda_0(u)-\Lambda_0(x)\} \\ \times \left[z - \frac{E\{Zr(2\theta' Z)O(X,U|Z)|X=x,U=u\}}{E\{r(2\theta' Z)O(X,U|Z)|X=x,U=u\}}\right],$$

where $y=(\delta,x,u,z)$ and $r(\theta' z)=\exp(\theta' z)$.

(b) *The information for θ is*

$$I(\theta) = E[\dot{l}^*_\theta(Y)]^{\otimes 2} = E\left[R(X,U,Z)\left[Z - \frac{E\{ZR(X,U,Z)|X,U\}}{E\{R(X,U,Z)|X,U\}}\right]^{\otimes 2}\right],$$

where $a^{\otimes 2} = aa'$ for $a \in R^d$, and $R(X,U,Z) = \{\Lambda(U|Z)-\Lambda(X|Z)\}^2$ $O(X,U|Z)$.

4. Variance Estimation

Even though we have an explicit expression for the information matrix in Theorem 3.1, it is not an easy matter to directly estimate the information matrix. Having computed the MLE, $(\hat{\theta}_n, \hat{\Lambda}_n)$, one can potentially evaluate the observed information matrix

$$I = -\frac{\partial^2}{\partial\psi^2} l_n(\psi)|_{\hat{\psi}_n},$$

where $\psi = (\theta,\Lambda)$. However, computation of the inverse of the high-dimensional observed information matrix may be numerically unstable. Therefore, we considered two different approaches to estimating the variance-covariance matrix of $\hat{\theta}_n$.

4.1. Missing information principle and partial likelihood

Let $(X<U)$ be a two-dimensional vector with first component being a left-truncation time and the second being an examination time, and for $i=1,\ldots,n$, let

$$\delta_i = \begin{cases} 1 & \text{if } X_i < T_i \le U_i \\ 0 & \text{if } U_i < T_i. \end{cases}$$

Define $\underline{\mathrm{Y}} = \{(\delta_i, X_i, U_i, Z_i)\}_{i=1}^n$ to be the observed data. Note that the observed data have two parts, $\underline{\mathrm{Y}}^{LLC}$ and $\underline{\mathrm{Y}}^{LRC}$, where LLC stands for the left-truncated and left-censored part and LRC stands for the left-truncated and right-censored part. Let n_L be the number of left-truncated and left-censored observations. Define $\underline{\mathrm{W}} = \{(X_j, T_j, Z_j)\}_{j=1}^{n_L}$ to be the left-truncated data in the literature, where T_j, $j = 1, \ldots, n_L$, is the unknown failure time in our problem. We refer to $\underline{\mathrm{W}}$ as the missing (incomplete) data because the $T_j's$ are not known but will be estimated in the variance estimation procedure. Define $(\underline{\mathrm{W}}, \underline{\mathrm{Y}}^{LRC})$ to be the complete data. Note that the complete data in this definition are left-truncated and right-censored data in the literature (Turnbull (1976), Tsai, Jewell and Wang (1987)). We call it so because it is an augmented version of the observed data $(\underline{\mathrm{Y}}^{LLC}, \underline{\mathrm{Y}}^{LRC})$ and requires only the partial likelihood function of the regression parameter in the variance estimation procedure. Therefore, none of the missing data and the complete data in our definition involves inverting a high-dimensional observed information matrix. In order to take advantage of the partial likelihood function (Cox (1975)) and risk sets for the proportional hazards model with left-truncated and right-censored data (Tsai, Jewell and Wang (1987)), we impute for left-truncated and left-censored failure times from the conditional distribution and then compute the information for θ_0.

For partly interval-censored data, Kim (1999) developed a generalized missing information principle, which is similar to the one for a parametric setting explained by Louis (1982). Unlike the case with partly interval-censored data, for LTIC 1 data, we do not have the consistency of the MLE of the baseline cumulative hazard function (see Theorem 7.1). In the imputation step, we take advantage of the MLE of the regression parameter and the baseline cumulative hazard function. Therefore, we need to take the inconsistency of the MLE of the baseline cumulative hazard function into account when we develop a generalized missing information principle. Applying the same expression as in Kim (1999), we have experienced a severe underestimation of the standard error. Therefore, we propose for LTIC 1 data that the observed information matrix should be approximated based on two partial likelihoods with risk sets adjusted for the left-truncation times as functions of only the regression parameter.

The information for θ_0 from the observed data can be approximated by

$$\begin{aligned} I &= I_{cmp} - I_{mis} \\ &= -\int_{\underline{\mathrm{W}}} (\partial^2/\partial\theta^2) \log PLC(\theta) p(\underline{\mathrm{w}}|\hat{\psi}_n, \underline{\mathrm{Y}}) d\underline{\mathrm{w}} - \mathrm{Var}_{\,\underline{\mathrm{W}}|\underline{\mathrm{Y}}}\{(\partial/\partial\theta) \log PLM(\theta)|_{\hat{\theta}_n}\}, \end{aligned} \tag{5}$$

where PLC stands for partial likelihood of complete data and PLM stands for partial likelihood of missing data. So the first term, I_{cmp}, is the complete information, and the second term, I_{mis}, is the missing information. The second term in (5) takes into account an extra variability caused by the inconsistent MLE of the baseline cumulative hazard function when compared with "Case 1" interval-censored data for which the MLE of the baseline cumulative hazard function is consistent.

Now we describe how to compute I_{cmp} and I_{mis}. Let S be the set of ordered observed times. Suppose for the ith subject, the ranks of a left-truncated observation and a left-censored observation, (x_i, u_i), are m_1 and m_2. Then the conditional pdf of the subject's failure time is

$$p(T_i = s_{(j)}|\hat{\psi}_n, \delta_i\!=\!1, x_i, u_i, z_i) = \frac{\exp\{-e^{z_i'\hat{\theta}_n}\hat{\Lambda}_n(s_{(j-1)})\} - \exp\{-e^{z_i'\hat{\theta}_n}\hat{\Lambda}_n(s_{(j)}\}}{\exp\{-e^{z_i'\hat{\theta}_n}\hat{\Lambda}_n(s_{(m_1)})\}-\exp\{-e^{z_i'\hat{\theta}_n}\hat{\Lambda}_n(s_{(m_2)})\}},$$

where $s_{(j)}$ is the jth ordered observed time and $j = m_1 + 1, \ldots, m_2$.

Since we can sample from $p(\underline{\mathrm{W}}|\hat{\psi}_n, \underline{\mathrm{Y}})$, the first integral in (5) may be approximated by the sum

$$\frac{1}{B}\sum_{j=1}^{B}\{(\partial^2/\partial\theta^2)\log PLC(\theta, \underline{\mathrm{w}}_j, \underline{\mathrm{Y}}^{LRC})|_{\hat{\theta}_n}\},$$

where $\underline{\mathrm{w}}_1, \ldots, \underline{\mathrm{w}}_B \sim^{i.i.d.} p(\underline{\mathrm{W}}|\hat{\psi}_n, \underline{\mathrm{Y}})$ with B being a large enough sample size.

Likewise, the second term in (5) may be approximated by the sum

$$\frac{1}{B}\sum_{j=1}^{B}\{(\partial/\partial\theta)\log PLM(\theta, \underline{\mathrm{w}}_j)|_{\hat{\theta}_n}\}^2 - \left[(1/B)\sum_{j=1}^{B}\{(\partial/\partial\theta)\log PLM(\theta, \underline{\mathrm{w}}_j)|_{\hat{\theta}_n}\}\right]^2,$$

with the same imputed data, $\underline{\mathrm{w}}_1, \ldots, \underline{\mathrm{w}}_B$ and the same sample size B.

Let $\{T_i^{\dagger}\}_{i=1}^{n_L}$ be the set of the imputed times from the above conditional distribution. Then $[\{(X_i, T_i^{\dagger}, Z_i)\}_{i=1}^{n_L}, \{(\delta_i = 0, X_i, U_i, Z_i)\}_{i=n_L+1}^{n}]$ is the set of the complete data, where the first component is the set of the missing data. The partial likelihood for the complete data is given by

$$PLC(\theta) = \prod_{i=1}^{n_L}\left\{\frac{\exp(\theta' Z_i)}{\sum_{j\in\Re(T_i{}^{\dagger *})}\exp(\theta' Z_j)}\right\},$$

where $\Re(T_i{}^{\dagger *}) = \{j : X_j < T_i{}^{\dagger} \le T_j{}^{\triangle}\}$ with $T_j{}^{\triangle}$ being either an imputed time or a right-censored time. Then $\log PLC(\theta) = \sum_{i=1}^{n_L}[\theta' Z_i - \log\{\sum_{j\in\Re(T_i^{\dagger *})}\exp(\theta' Z_j)\}]$ and we find the second derivative

$$\sum_{i=1}^{n_L}\frac{\sum_{j\in\Re(T_i^{\dagger *})}\exp(\theta' Z_j)Z_j^2}{\sum_{j\in\Re(T_i^{\dagger *})}\exp(\theta' Z_j)} - \sum_{i=1}^{n_L}\left\{\frac{\sum_{j\in\Re(T_i^{\dagger *})}\exp(\theta' Z_j)Z_j}{\sum_{j\in\Re(T_i^{\dagger *})}\exp(\theta' Z_j)}\right\}^2.$$

The partial likelihood for the missing data, $\{(X_i, T_i{}^{\dagger}, Z_i)\}_{i=1}^{n_L}$, is given by

$$PLM(\theta) = \prod_{i=1}^{n_L} \left\{ \frac{\exp(\theta' Z_i)}{\sum_{j \in \Re(T_i{}^{\dagger **})} \exp(\theta' Z_j)} \right\},$$

where $\Re(T_i{}^{\dagger **}) = \{j : X_j < T_i{}^{\dagger} \le T_j{}^{\dagger}\}$. The score function for the missing data is then given by

$$(\partial/\partial\theta)\, \log PLM(\theta) = \sum_{i=1}^{n_L} \left\{ Z_i - \frac{\sum_{j \in \Re(T_i^{\dagger **})} \exp(\theta' Z_j) Z_j}{\sum_{j \in \Re(T_i^{\dagger **})} \exp(\theta' Z_j)} \right\}.$$

Consequently, the two integrals in (5) are approximated by the following expressions, where we call the first one complete information and the second one missing information:

$$I_{cmp} = \frac{1}{B} \sum_{k=1}^{B} \left[\sum_{i=1}^{n_L} \frac{\sum_{j \in \Re(T_{ik}^{\dagger *})} \exp(\theta' Z_{jk}) Z_{jk}^2}{\sum_{j \in \Re(T_{ik}^{\dagger *})} \exp(\theta' Z_{jk})} - \sum_{i=1}^{n_L} \left\{ \frac{\sum_{j \in \Re(T_{ik}^{\dagger *})} \exp(\theta' Z_{jk}) Z_{jk}}{\sum_{j \in \Re(T_{ik}^{\dagger *})} \exp(\theta' Z_{jk})} \right\}^2 \right],$$

where for each i and k, $\Re(T_{ik}^{\dagger *}) = \{j : X_{jk} < T_{ik}{}^{\dagger} \le T_{jk}{}^{\triangle}\}$;

$$\begin{aligned} I_{mis} = {} & \frac{1}{B} \sum_{k=1}^{B} \left[\sum_{i=1}^{n_L} \left\{ Z_{ik} - \frac{\sum_{j \in \Re(T_{ik}^{\dagger **})} \exp(\theta' Z_{jk}) Z_{jk}}{\sum_{j \in \Re(T_{ik}^{\dagger **})} \exp(\theta' Z_{jk})} \right\} \right]^2 \\ & - \left[\frac{1}{B} \sum_{k=1}^{B} \sum_{i=1}^{n_L} \left\{ Z_{ik} - \frac{\sum_{j \in \Re(T_{ik}^{\dagger **})} \exp(\theta' Z_{jk}) Z_{jk}}{\sum_{j \in \Re(T_{ik}^{\dagger **})} \exp(\theta' Z_{jk})} \right\} \right]^2, \end{aligned}$$

where for each i and k, $\Re(T_{ik}^{\dagger **}) = \{j : X_{jk} < T_{ik}{}^{\dagger} \le T_{jk}{}^{\dagger}\}$.

4.2. Profile information

Murphy and van der Vaart (1999) proved that the profile information matrix is a consistent estimator of the efficient information matrix for i.i.d. samples. For any fixed θ in a neighborhood of $\hat{\theta}_n$, let $\hat{\Lambda}_n(\cdot, \theta)$ maximize $l_n(\theta, \Lambda)$ over Φ, where Φ is the collection of all bounded nonnegative nondecreasing functions on the support of the truncation variable and the censoring variable. The profile log-likelihood is defined to be $l_n(\theta, \hat{\Lambda}_n(\cdot, \theta))$. The standard error of $\hat{\theta}_n$ is estimated based on the second derivative of $l_n(\theta, \hat{\Lambda}_n(\cdot, \theta))$, treating it as the standard log-likelihood. Since there is no closed form expression for $\hat{\Lambda}_n(\cdot, \theta)$, the second derivative of $l_n(\theta, \hat{\Lambda}_n(\cdot, \theta))$ is computed numerically. In particular, the likelihood is computed for a small grid of θ-values near $\hat{\theta}_n$, a quadratic function is fit by the least squares, and then the inverse of the second derivative matrix is used to estimate the variance-covariance matrix of $\hat{\theta}_n$. This method was coded with S-plus functions, "nlminb" and "lm". In simulation studies, a grid of 21 points (the middle one being $\hat{\theta}_n$) spaced at intervals of 0.01 was used to estimate the second derivative of the profile log-likelihood.

5. Simulation Studies

In order to illustrate our methods, we performed a simulation study. Under the proportional hazards model $\lambda(t|z) = \lambda_0(t)e^{\theta' z}$, we generated true failure times from two distributions: exponential$\{\lambda_0(t) = 1\}$ and Weibull$\{\lambda_0(t) = 2t\}$. We considered a covariate, $Z \sim$ Bernoulli(1/2). The true regression parameter θ_0 is 1. Left-truncation times and examination times were generated to make the proportions of left-truncated and left-censored observations, and left-truncated and right-censored observations about equal. We considered samples of size 50, 100 and 150. The implementation was done in S-plus using a non-linear programming routine for optimization, "nlminb". The main programming task is writing the likelihood function. One may save the computation time substantially provided a hessian matrix is supplied to the routine. Let δ, T, X, U and Z be a vector of censoring indicators, true failure times, left-truncation times, examination times, and covariates, respectively. Then an i.i.d. sample of size n is generated as follows.

(a) Set the counter $k = 0$.

(b) (T_i, X_i, U_i, Z_i) is generated independently from their specified distributions.

(c) If $T_i < X_i$ then we ignore this observation and go back to (b). If $X_i < T_i \leq U_i$ then we obtain a left-truncated and left-censored observation, $(\delta_i = 1, X_i, U_i, Z_i)$. If $U_i < T_i$ then we obtain a left-truncated and right-censored observation, $(\delta_i = 0, X_i, U_i, Z_i)$. Set $k \leftarrow k + 1$.

(d) Repeat (b) and (c) until k becomes n.

Let B be the number of replications. Repeat (a) through (d) B times to obtain B replicated data sets. In step (c), the percentage of left-truncation depends on how we choose distributions of X_i and U_i. A referee suggested different percentages of left-truncation in the simulation. Here we consider two cases. On average, 36% and 15% of original observations were discarded due to left-truncation, respectively. Table 1 summarizes the results of the simulation study for both cases. For each sample size, bias and mean standard error estimate are computed from $B = 500$ replications. The percentage of left-truncation seems to play an important role in the small sample properties of the MLE of the regression parameter. In particular, for a given sample size, the smaller the percentage of left-truncation the smaller the bias. Another interesting result is that for a sample of size 50, the asymptotics do not work well. The simulation study suggests that a sample size for the asymptotics to kick in should be at least 150. Mean standard error estimates computed by the proposed methods are close to the sample standard deviation. However, there seems to be a slight hint of bias in the regression parameter. This is not surprising if we refer to Alioum and Commenges (1996). Under a similar condition ($n = 200$, $B = 200$ and $\theta_0 = 1$), their bias in the regression parameter is 16.7%. In summary, the simulation study supports the generalized missing information principle and the profile information.

Table 1. Simulation results for MLEs.

		36% discard				15% discard			
n	distribution	$\hat{b}$ [a]	$\hat{\sigma}_M$ [b]	$\hat{\sigma}_P$ [c]	s [d]	$\hat{b}$ [a]	$\hat{\sigma}_M$ [b]	$\hat{\sigma}_P$ [c]	s [d]
50	exponential	0.301	0.731	0.753	0.777	0.281	0.624	0.650	0.670
50	Weibull	0.307	0.751	0.788	0.810	0.288	0.645	0.672	0.698
100	exponential	0.231	0.398	0.383	0.422	0.208	0.391	0.386	0.426
100	Weibull	0.221	0.388	0.378	0.434	0.196	0.383	0.374	0.432
150	exponential	0.157	0.359	0.330	0.372	0.128	0.355	0.344	0.352
150	Weibull	0.152	0.348	0.343	0.388	0.125	0.344	0.341	0.358

[a] estimated bias

[b] mean standard error estimate by missing information principle

[c] mean standard error estimate by profile information

[d] sample standard deviation of regression parameter estimates.

6. Example

Consider the AIDS cohort study of hemophiliacs discussed in Kim, DeGruttola and Lagakos (1993). Their time of interest was defined to be the AIDS-related-symptom induction time, i.e., the difference between the AIDS-related-symptom diagnosis time and the HIV-1 infection time, but ours is defined to be the time until AIDS-related-symptom diagnosis since the time of the receipt of the contaminated blood factor. The study population consisted of 257 individuals with Type A or B hemophilia who had been treated at Hôpital Kremlin Bicetre and Hôpital Coeur des Yvelines in France since 1978. These hemophiliacs were at risk for HIV-1 infection through the contaminated blood factor they received for their treatment. By the time of analysis, 188 were found to be infected with the virus, 41 of whom subsequently progressed to AIDS-related symptoms. In this case, there is 27% left-truncation.

The primary goal of this example is to apply the procedures described in the previous sections to assess the effects of level of treatment received for hemophilia on the risk of developing AIDS-related symptoms. The subjects are classified into two groups, lightly and heavily treated groups, according to the amount of blood they received. In the original data set, there are HIV-1 infection time intervals and AIDS-related-symptom diagnosis time intervals. The time unit is 6-months. We slightly modify this data set into LTIC 1 format, and then illustrate the proposed methods for the modified data. For each individual, we replace one's HIV-1 infection time interval and AIDS-related-symptom diagnosis time interval by the midpoint and the right end point of the interval, respectively. Define $Z_i = 0$ if the ith individual belongs to the lightly treated group and $Z_i = 1$ otherwise. Figure 1 shows the estimated survival functions. The solid graph is the estimator of the baseline survival function for lightly treated subjects ($Z = 0$) and the dotted graph is the estimator of the survival function for heavily treated subjects ($Z = 1$). Applying the procedures described in Sections 2.1, 2.2, 4.1 and 4.2, we obtain $\hat{\theta} = 0.765$ with $\hat{\sigma}_M$ and $\hat{\sigma}_P$ being 0.367 and 0.353, respectively. The test of $\theta_0 = 0$ results in a p-value of 0.038. The results here suggest that the subjects in the heavily treated group had significantly greater risk of developing AIDS-related symptoms. This result is similar to that obtained by Kim, DeGruttola and Lagakos (1993), who had $\hat{\theta} = 0.69$ with the estimated standard error of 0.34, using a discrete analogue of the proportional hazards model.

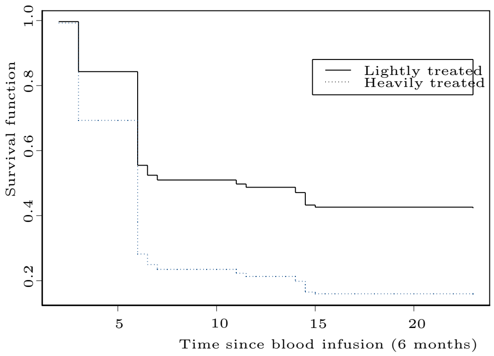

Figure 1. Estimated survival functions for heavily treated and lightly treated group.

Since LTIC 1 data are somewhat similar to "Case 1" and "Case 2" interval-censored data, methods to deal with them are not sufficiently known and practitioners would treat them as either "Case 1" or "Case 2" interval-censored data. Therefore we considered the effect of misspecifying the data as either "Case 1" interval-censored data by ignoring the left-truncation time, or a special case of "Case 2" interval-censored data (Huang and Wellner (1997)) by treating the left-truncation time as one of the examination times for the event of interest. As we notice in Table 2 both of the cases resulted in severe overestimation of the standard error. This confirms that left-truncation time is not only nonnegligible but also should not be considered as one of the examination times for the event of interest.

Table 2. MLEs from three types.

Data type	$\hat{\theta}$	standard error
LTIC 1	0.765	0.367
"Case 1" interval-censored	0.692	1.031
"Case 2" interval-censored	0.732	0.556

7. Consistency and Rate of Convergence

This section closely follows Huang (1996).

7.1. Consistency of $\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ and $\hat{\boldsymbol{\Gamma}}_{\boldsymbol{n}}$

For simplicity, define $V = (X, U)$ and $\Gamma(v) = \Lambda(u) - \Lambda(x)$. Likewise, define $\hat{\Gamma}_n(v) = \hat{\Lambda}_n(u) - \hat{\Lambda}_n(x)$. Let $G(v)$ be the joint distribution function of V. Define

$$\Psi = \{\Gamma : [\tau_0, \tau_1]^2 \to [m_1, M_1], \quad \text{for some} \quad 0 < m_1 < M_1\}. \tag{6}$$

Define a distance d on $\Theta \times \Psi$ by

$$d((\theta_1, \Gamma_1), (\theta_2, \Gamma_2)) = |\theta_1 - \theta_2| + \left[\int (\Gamma_1(v) - \Gamma_2(v))^2 dG(v)\right]^{1/2}. \tag{7}$$

Theorem 7.1. (*Consistency*) *Suppose that conditions* (A1)–(A5) *of Section* 3 *hold. Then* $d((\hat{\theta}_n, \hat{\Gamma}_n), (\theta_0, \Gamma_0)) \overset{a.s.}{\to} 0$.

The proof of Theorem 7.1 is given in the Appendix.

7.2. Rate of convergence

We now consider the convergence rate for $(\hat{\theta}_n, \hat{\Gamma}_n)$ under the norm defined in (7). Convergence rates of nonparametric maximum likelihood estimators have been considered by Wong and Severini (1991), Birgé and Massart (1993), van de Geer (1993), and van der Vaart and Wellner (1996). These authors demonstrated that the convergence rate is closely related to the entropy numbers of the parameter space. In particular, Theorem 3.4.1 of van der Vaart and Wellner (1996) shows that, once consistency is established, the convergence rate is determined by the smoothness of the model and the continuity modulus of the objective function (which is the log-likelihood function in the case of maximum likelihood estimation) over the parameter space. In the following, we apply the approach developed in van der Vaart and Wellner (1996) to study the convergence rate of $(\hat{\theta}_n, \hat{\Gamma}_n)$.

After consistency of $\hat{\theta}_n$ is established, we can focus our attention to a neighborhood of θ_0. For any $\eta > 0$, let $B(\theta_0, \eta)$ be a ball centered at θ_0 with radius η. If θ_0 is on the boundary of Θ, then take $B(\theta_0, \eta)$ to be $B(\theta_0, \eta) \cap \Theta$. In this way, we always have $B(\theta_0, \eta) \subset \Theta$. Restrict the class of log-likelihood functions $l(\theta, \Lambda)$ defined by (2) to

$$\mathcal{H} = \{l(\theta, \Gamma) : \theta \in B(\theta_0, \eta), \Gamma \in \Psi\}. \tag{8}$$

For any probability measure Q, let $L_2(Q) = \{f : \int f^2 dQ < \infty\}$ with norm $\|\cdot\|_2$. For any subclass $\mathcal{F}$ of $L_2(Q)$, define the bracketing number

$$N_{[\ \ ]}(\epsilon, \mathcal{F}, L_2(Q)) = \min\{m : \text{there exist } \ f_1^L, f_1^U, \ldots, f_m^L, f_m^U \ \text{ such that for each } \ f \in \mathcal{F}, \\ f_i^L \le f \le f_i^U \ \text{ for some } i, \ \text{ and } \ \|f_i^U - f_i^L\|_2 \le \epsilon\}.$$

Following van der Vaart and Wellner (1996), let

$$J_{[\ \ ]}(\eta, \mathcal{F}, \|.\|_2) = \int_o^\eta [1 + \log N\{\epsilon, \mathcal{F}, L_2(Q)\}]^{1/2} d\epsilon \tag{9}$$

be the bracketing integral of the class of functions $\mathcal{F}$.

Lemma 7.2. *Let $\mathcal{H}$ be defined by* (8) *and suppose that Z has bounded support. Then there exists a constant $C > 0$ such that $\sup_Q N_{[\ ]}\{\epsilon, \mathcal{H}, L_2(Q)\} \le C(1/\epsilon^d)e^{1/\epsilon}$, for all $\epsilon > 0$, where d is the dimension of θ_0. Hence, for ϵ small enough and for some $C_0 > 0$, we have $\sup_Q \log N_{[\ ]}\{\epsilon, \mathcal{H}, L_2(Q)\} \le C_0(1/\epsilon)$. Here Q runs through the class of all probability measures.*

Remark 7.1. From this lemma, the bracketing integral for the class $\mathcal{H}$ is

$$J_{[\ ]}\{\eta, \mathcal{H}, L_2(Q)\} = O(1) \int_0^\eta \sqrt{1/\epsilon}\, d\epsilon = O(\eta^{1/2}) \ \text{ for } \ \eta \ \text{ close to zero.}$$

Applying Lemma 7.2 above, and Theorem 3.4.1 and Lemma 3.4.2 in van der Vaart and Wellner (1996), we can prove the following result.

Theorem 7.3. (*Rate of Convergence*) *Assume that* (A1)–(A5) *in Section* 3 *are satisfied. Then $d((\hat{\theta}_n, \hat{\Gamma}_n), (\theta_0, \Gamma_0)) = O_p(n^{-1/3})$.*

We remark here that the overall rate of convergence is dominated by $\hat{\Gamma}_n$. This rate agrees with the convergence rate of the NPMLE of a distribution function studied by Groeneboom and Wellner (1992). In the next section we show that the convergence rate of $\hat{\theta}_n$ can be refined to achieve root-n convergence.

8. Asymptotic normality

This section closely follows Huang (1996, 1998) as well. We give sufficient conditions for the MLE of the finite dimensional parameter in a semiparametric model to be asymptotically normal and efficient. The results are applied to prove Theorem 8.1.

Let $Y_1, \ldots, Y_n$ be independent random variables with a common probability measure $P_{\theta,\phi}$, where $(\theta, \phi) \in \Theta \times \Phi$. Here Θ is a subset of R^d and Φ is a general space. Assume that $P_{\theta,\phi}$ has a density $p(\cdot, \theta, \phi)$ with respect to a σ-finite measure. Let $(\theta_0, \phi_0) \in \Theta \times \Phi$

be the true parameter value under which the observations are generated. The MLE of (θ_0, ϕ_0) is the value $(\hat{\theta}_n, \hat{\phi}_n)$ that maximizes the log-likelihood function

$$l_n(\underline{Y}, \theta, \phi) = \sum_{i=1}^{n} \log p(Y_i, \theta, \phi)$$

over the parameter space $\Theta \times \Phi$. Let $l(\cdot, \theta, \phi) = \log p(\cdot, \theta, \phi)$, and define

$$\begin{aligned}
&\dot{l}_1(Y, \theta, \phi) = \frac{\partial}{\partial \theta} l(Y, \theta, \phi), \quad \dot{l}_2(Y, \theta, \phi)[h] = \frac{\partial}{\partial \epsilon} l(Y, \theta, \phi + \epsilon h)|_{\epsilon=0}, \\
&\dot{l}^*_\theta(Y, \theta, \phi) = \dot{l}_1(Y, \theta, \phi) - \dot{l}_2(Y, \theta, \phi)[\mathbf{h}^*],
\end{aligned}$$

where h is an element of $\mathcal{H}$, a class of bounded functions on the support of ϕ, and $\mathbf{h}^*$ is an element of $\mathcal{H}$ that minimizes $\rho(h) \equiv E\|\dot{l}_1(Y, \theta, \phi) - \dot{l}_2(Y, \theta, \phi)[h]\|^2$ over $\mathcal{H}$. The minimizer $\mathbf{h}^*$ is called the least favorable direction. Then $\dot{l}_{1n}(\underline{Y}, \theta, \phi)$ and $\dot{l}_{2n}(\underline{Y}, \theta, \phi)[h]$ are similarly defined. Proposition 8.1 shows that the following conditions are sufficient for the MLE of the finite dimensional parameter in a semiparametric model to be asymptotically normal and efficient:

$$\dot{l}_{1n}(\underline{Y}, \hat{\theta}_n, \hat{\phi}_n) = 0, \quad \text{and} \quad \dot{l}_{2n}(\underline{Y}, \hat{\theta}_n, \hat{\phi}_n)[\mathbf{h}^*] = o_p(n^{-1/2}); \tag{10}$$

$$(P_n - P)\{\dot{l}^*_\theta(Y, \hat{\theta}_n, \hat{\phi}_n) - \dot{l}^*_\theta(Y, \theta_0, \phi_0)\} = o_p(n^{-1/2}); \tag{11}$$

$$P\{\dot{l}^*_\theta(Y, \hat{\theta}_n, \hat{\phi}_n) - \dot{l}^*_\theta(Y, \theta_0, \phi_0)\} = I(\theta_0)(\hat{\theta}_n - \theta_0) + o_p(\|\hat{\theta}_n - \theta_0\|) + o_p(n^{-1/2}), \tag{12}$$

where $I(\theta_0) = E[\dot{l}^*_\theta(Y, \theta_0, \phi_0)]^{\otimes 2}$ is the information matrix.

Proposition 8.1. *Suppose conditions* (10) − (12) *are satisfied, and* $I(\theta_0)$ *is nonsingular. Then* $\sqrt{n}(\hat{\theta}_n - \theta_0) = -I(\theta_0)^{-1}\sqrt{n}P_n\dot{l}^*_\theta(Y, \theta_0, \phi_0) + o_p(1) \xrightarrow{d} N(0, I(\theta_0)^{-1})$.

Proof. Combining (11) and (12), we have

$$P_n\{\dot{l}^*_\theta(Y, \hat{\theta}_n, \hat{\phi}_n) - \dot{l}^*_\theta(Y, \theta_0, \phi_0)\} = I(\theta_0)(\hat{\theta}_n - \theta_0) + o_p(\|\hat{\theta}_n - \theta_0\|) + o_p(n^{-1/2}).$$

By (10), it follows that

$$P_n\dot{l}^*_\theta(Y, \theta_0, \phi_0) = -I(\theta_0)(\hat{\theta}_n - \theta_0) + o_p(\|\hat{\theta}_n - \theta_0\|) + o_p(n^{-1/2}).$$

Since $I(\theta_0)$ is nonsingular, and $P_n\dot{l}^*_\theta(Y, \theta_0, \phi_0) = O_p(n^{-1/2})$, this implies $\|\hat{\theta}_n - \theta_0\| = O_p(n^{-1/2})$. Thus, $o_p(\|\hat{\theta}_n - \theta_0\|) = o_p(n^{-1/2})$. Therefore, $P_n\dot{l}^*_\theta(Y, \theta_0, \phi_0) = -I(\theta_0)(\hat{\theta}_n - \theta_0) + o_p(n^{-1/2})$ and the result follows.

We now state the main theorem that under appropriate regularity conditions, the MLE $\hat{\theta}_n$ satisfies a central limit theorem and is asymptotically efficient.

Theorem 8.1. (*Asymptotic normality and efficiency*) *Suppose that* θ_0 *is an interior point of* Θ *and that assumptions* (A1)–(A5) *in section* 3 *are satisfied. Then*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -I(\theta_0)^{-1}\sqrt{n}P_n\dot{l}^*_{\theta_0}(Y) + o_p(1) \xrightarrow{d} N(0, I(\theta_0)^{-1}),$$

where P_n is the empirical measure of $Y_i = (\delta_i, X_i, U_i, Z_i)$, $i = 1, \ldots, n$, $\dot{l}^*_{\theta_0}(Y)$ is the efficient score defined in Theorem 3.1*, and $I(\theta_0)$ is the information matrix.*

9. Concluding Remarks

We considered the maximum likelihood estimation approach for the proportional hazards model with LTIC 1 data. Treating the difference between left-truncation time and censoring time of the baseline cumulative hazard function as a nuisance parameter, we were able to prove the asymptotic properties of the MLE of the regression parameter. They are similar to those for "Case 1" interval-censored data (Huang (1996)). It is clear that left-truncation is responsible for the inconsistency of the MLE of the cumulative hazard function, and without additional properties of the baseline cumulative hazard functions, the consistency of the MLE of the cumulative hazard function will not be easy to prove.

Left-truncation also plays an important role in the small sample properties of the MLE of the regression parameter. First, misspecifying the data as either "Case 1" interval-censored data or a special case of "Case 2" interval-censored data can result in a severe overestimation of the standard error. Second, for the same sample size, the smaller the percentage of truncation the smaller the bias. Lastly, for our asymptotics to apply, the sample size should be at least 150.

Regarding variance estimation, the profile information procedure is very simple to apply for one or two dimensional covariates because we can compute the profile likelihood surface. However, it becomes very difficult for higher dimensional covariates because we may not be able to compute the profile likelihood surface unless we develop better computational methods. On the other hand, the generalized missing information principle can be applied for any finite dimensional covariates because it takes advantage of imputed data and the partial likelihood function. It seems that the approach in this paper can be extended to analyze left-truncated and general interval-censored data. As a further study, it would be interesting to consider LTIC 1 data with time-dependent covariates as well.

Acknowledgements

This paper is based on part of the author's Ph.D. dissertation written under the supervision of Professor Jian Huang. The author is grateful to him for his guidance and for many helpful discussions on the subject of this paper. He is also grateful to a co-editor, an associate editor, and a referee for their valuable comments.

Appendix. Proofs

Proof of Proposition 2.1. We prove only (1). For any function ϕ,

$$\frac{\partial^2}{\partial s^2} l_n(\theta, \Lambda + s\phi)|_{s=0} = -\sum_{i=1}^{n} \delta_i \frac{\exp(-a_i)\{2 - \exp(-a_i)\}{b_i}^2}{\{1 - \exp(-a_i)\}^2} < 0,$$

where $a_i = \{\Lambda(U_i) - \Lambda(X_i)\}e^{\theta' Z_i}$ and $b_i = \{\phi(U_i) - \phi(X_i)\}e^{\theta' Z_i}$, for $i = 1, \ldots, n$.

Proof of Theorem 3.1. See Kim (1999).

Proof of Theorem 7.1. Let $\hat{p}_n(y) = p(y; \hat{\theta}_n, \hat{\Lambda}_n)$ and $p_0(Y) = p(Y; \theta_0, \Lambda_0)$, where p is defined by (2). Since $(\hat{\theta}_n, \hat{\Lambda}_n)$ maximizes the likelihood function over $\Theta \times \Phi$, and $(\theta_0, \Lambda_0) \in \Theta \times \Phi$,

$$\sum_{i=1}^{n} \log p_n(Y_i) \geq \sum_{i=1}^{n} \log p_0(Y_i),$$

$$\sum_{i=1}^{n} \log \frac{p_n(Y_i)}{p_0(Y_i)} \geq 0.$$

By concavity of the function $y \rightarrow \log y$, for any $0 < \alpha < 1$,

$$\frac{1}{n} \sum_{i=1}^{n} \log \left\{ 1 - \alpha + \alpha \frac{p_n(Y_i)}{p_0(Y_i)} \right\} \geq 0. \tag{13}$$

The left hand side can be written as

$$\int \log \left\{ 1 - \alpha + \alpha \frac{p_n(y)}{p_0(y)} \right\} d(P_n - P)(y) + \int \log \left\{ 1 - \alpha + \alpha \frac{p_n(y)}{p_0(y)} \right\} dP(y), \tag{14}$$

where P_n is the empirical measure of $(\delta_i, X_i, U_i, Z_i)$, $i = 1, \ldots, n$; here P is the joint probability measure of (δ, X, U, Z).

Let the sample space Ω be the space of all infinite sequences $(\delta_1, X_1, U_1, Z_1)$, $(\delta_2, X_2, U_2, Z_2), \ldots$, endowed with the usual σ-algebra generated by the product topology on $\prod_{i=1}^{n} \{0, 1\} \times R^{d+2}$ and the product measure $\mathbf{P}$.

The class of functions of the first term of (14), $\mathcal{H} = \{\log(1 - \alpha + \alpha p/p_0) : p \in \mathcal{P}\}$, where $p_0(y) = p(y; \theta_0, \Lambda_0)$, is uniformly bounded and uniformly Lipschitz of order 1 and hence is Donsker. The generalized Glivenko-Cantelli theorem together with Donsker guarantees that there exists a set $\Omega_0 \in \Omega$ with $\mathbf{P}(\Omega_0) = 1$ such that for every $\omega \in \Omega_0$, the first term of (14) converges to zero.

Now fix $\omega \in \Omega_0$. For this ω, write $\hat{\theta}_n = \hat{\theta}_n(\omega)$ and $\hat{\Lambda}_n(\cdot) = \hat{\Lambda}_n(\cdot, \omega)$. Since Θ is bounded, for any subsequence of $\hat{\theta}_n$ we can find a further subsequence converging to $\theta_* \in \bar{\Theta}$, the closure of Θ. Moreover, by Helly's Selection Theorem, for any subsequence of $\hat{\Lambda}_n$ we can find a further subsequence converging to some increasing function Λ_*. Choose the convergent subsequence of $\hat{\theta}_n$ and the convergent subsequence of $\hat{\Lambda}_n$ so that they have the same indices and, without loss of generality, assume that $\hat{\theta}_n$ converges to θ_* and that $\hat{\Lambda}_n$ converges to Λ_*. Let $p_*(y) = p(y; \theta_*, \Lambda_*)$. By the Bounded Convergence Theorem, the second term of (14) converges to

$$\int \log \left\{ 1 - \alpha + \alpha \frac{p_*(y)}{p_0(y)} \right\} dP(y).$$

By (13), this is nonnegative. However, by Jensen's Inequality, it must be nonpositive, therefore it must be zero. It follows that $p_*(y) = p_0(y)$ P – almost surely. This implies $\Gamma_*(v) e^{\theta_*{}' Z} = \Gamma_0(v) e^{\theta_0{}' Z}$ P – almost surely. This and condition (b) of (A2) imply that there exists $z_1 \neq z_2$ such that for some $v^* \in [\tau_0, \tau_1]^2$, $\Gamma_*(v^*) e^{\theta_*{}' z_1} = \Gamma_0(v^*) e^{\theta_0{}' z_1}$ and

$\Gamma_*(v^*)e^{\theta_*{}' z_2} = \Gamma_0(v^*)e^{\theta_0{}' z_2}$. Since by (A3)–(A5), $\Gamma_*(v^*) > 0$ and $\Gamma_0(v^*) > 0$, this implies $(\theta_* - \theta_0)'(z_1 - z_2) = 0$. Again, by condition (b) of (A2), the collection of such z_1 and z_2 has positive probability and there exists at least d such pairs that constitute a full rank $d \times d$ matrix. It follows that $\theta_* = \theta_0$. This in turn implies $\Gamma_*(v) = \Gamma_0(v)$ G – almost surely. By the Bounded Convergence Theorem,

$$\int \{\hat{\Gamma}_n(v) - \Gamma_0(v)\}^2 dG(v) \to 0. \tag{15}$$

Since (15) and $\theta_* = \theta_0$ hold for any $\omega \in \Omega_0$ with $\mathbf{P}(\Omega_0) = 1$, the proof is complete.

Proof of Lemma 7.2. See Huang (Lemma 3.1, 1996).

Proof of Theorem 7.3. See Huang (Theorem 3.3, 1996).

Proof of Theorem 8.1. See Huang (Theorem 3.4 and Lemma 7.1, 1996).

References

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

Andersen, P. K., Borgan, Ø., Gill, R. D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes.* Springer-Verlag, New York.

Bickel, P., Klaassen, C., Ritov, Y. and Wellner, J. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* Springer Verlag, New York.

Birgé, L. and Massart, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields* **97**, 113-150.

Birman, M. S. and Solomjak, M. Z. (1967). Piecewise-polynomial approximation of functions of the classes W_p. *Math. USSR Sbornik* **73**, 295-317.

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

Cox, D. R. (1975). Partial likelihood. *Biometrika* **62**, 269-276.

Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics* **42**, 845-854.

Groeneboom, P. and Wellner, J. A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation.* DMV Seminar Band 19, Birkhauser, Basel.

Hoel, D. G. and Walburg, H. E. (1972). Statistical analysis of survival experiments. *J. Nat. Cancer Inst.* **49**, 361-372.

Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. *Ann. Statist.* **24**, 540-568.

Huang, J. (1998). A least squares approach to consistant information estimation in semiparametric models. Preprint, Department of Statistics and Actuarial Science, University of Iowa.

Huang, J. and Wellner, J. A. (1997). Interval censored survival data: a review of recent progress. In *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis* (Edited by D. Lin and T. Fleming). Springer-Verlag, New York.

Kim, J. S. (1999). The proportional Hazards model with arbitrarily interval-censored data. Ph.D Dissertation, Department of Statistics and Actuarial Science, University of Iowa.

Kim, M. Y., DeGruttola, V. G. and Lagakos, S. W. (1993). Analyzing doubly censored data with covariates, with application to AIDS. *Biometrics* **49**, 13-22.

Klein, J. P. and Moeschberger, M. L. (1997). *Survival Analysis: Techniques for Censored and Truncated Data.* Springer-Verlag, New York, Inc.

Louis, T. A. (1982). Finding observed information using the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **44**, 98-130.

Murphy, S. A. and van der Vaart, A. W. (1999). Observed information in semiparametric models. *Bernoulli* **5**, 381-412.

Tsai, W. Y., Jewell, N. P. and Wang, M. C. (1987). A note on the product-limit estimator 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. *J. Roy. Statist. Soc. Ser. B* **38**, 290-295.

van de Geer, S. (1993). Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.* **21**, 14-44.

van der Vaart, A. W. (1991). On differentiable functionals. *Ann. Statist.* **19**, 178-204.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes with Applications to Statistics.* Springer, New York.

Wong, W. H. and Severini, T. A. (1991). On maximum likelihood estimation in infinite dimensional parameter space. *Ann. Statist.* **16**, 603-632.

Department of Mathematics and Statistics, Portland State University, Portland, OR 97207, U.S.A.

E-mail: jkim@mth.pdx.edu

(Received July 2001; accepted October 2002)