Statistica Neerlandica (2001) Vol. 55, nr. 3, pp. 367–376

Efficiency considerations in the additive hazards model with current status data

D. Ghosh[1]

Department of Biostatics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109-2029

For current status data, LIN, OAKES and YING (1998) proposed a procedure for estimation of the regression parameters in the additive hazards model that makes clever use of martingale theory. However, one of the outstanding problems posed in the paper was the issue of efficient estimation, as their estimators do not attain the semiparametric information bound. In this paper, we explore this issue and provide a characterization of the NPMLE. We conduct efficiency comparisons between the NPMLE and the procedure of LIN *et al.* (1998) analytically and numerically through analysis of a dataset from a tumorigenicity experiment.

Key Words and Phrases: additive risk, empirical process, interval censoring, semiparametric efficiency, survival analysis.

1 Introduction

The analysis of failure-time data is an area of substantial statistical research, having started with the seminal papers of KAPLAN and MEIER (1958) and COX (1972). The focus of much of the work in this field has dealt with the situation where the time to event is subject to right censoring. In this case, if T denotes the survival time, C the censoring time and Z a p-vector of covariates, one observes $(\min(T, C), I(T \leqslant C), Z)$, where $\min(a, b)$ is the minimum of two numbers a and b, and $I(A)$ is the indicator function for the set A. However, other censoring schemes are possible. For example, in certain situations, one can only observe $(C, I\ (T \leqslant C), Z)$; this is known as 'case 1' interval censored or current status data. With this type of censoring, the exact survival time is never observed. Current status data arise in a variety of settings. DIAMOND and MCDONALD (1991) considered current status data in demography studies, while HOEL and WALBURG (1972) examined such data arising from animal tumorigenicity experiments. Current status data also occur in HIV studies, such as in SHIBOSKI and JEWELL (1992) and JEWELL, MALANI and VITTINGHOFF (1994). The situation of 'case 1' interval censored data can be extended to a more general interval censoring setting, as discussed in HUANG and WELLNER (1997).

* E-mail address: ghoshd@umich.edu

[1] Research supported in part by the National Science Foundation and National Institutes of Health.

Several authors, including AYER *et al.* (1955), PETO (1973), TURNBULL (1976) and GRONENBOOM and WELLNER (1992) have proposed procedures for nonparametric estimation of the distribution function of T, F, with current status data. In particular, GRONENBOOM and WELLNER (1992) showed that the nonparametric maximum likelihood estimate (NPMLE) of F, $\hat{F}$, is consistent for F and converges at an $n^{1/3}$ rate to a complicated limiting distribution.

Semiparametric regression models for current status data have also been examined. FINKELSTEIN (1986) proposed a method of estimation for the proportional hazards model, but the asymptotic properties of her estimator remain unknown. Efficient estimation in this model based on NPMLE methods was developed by HUANG (1996), who showed that the regression parameters converge at an $n^{1/2}$ rate to a normal distribution.

An alternative to the proportional hazards model is the additive hazards model, proposed in LIN and YING (1994). This model postulates that covariates have an additive effect on the baseline risk:

$$\lambda(t|Z) = \lambda_0(t) + \theta' Z, \tag{1}$$

where $\lambda_0(t)$ is the baseline hazard of T and θ denotes the regression parameters. We leave $\lambda_0(t)$ unspecified so that (1) specifies a semiparametric model. For the analysis of current status data using this model, LIN, OAKES and YING (1998) derived a novel estimation procedure for θ that uses martingale theory in a clever manner. One of the outstanding questions in their paper was the investigation of efficiency in the estimation of the regression parameters. The estimators proposed by LIN *et al.* (1998) do not achieve the semiparametric information bound. In this article, we investigate the issue of efficiency and propose a method of estimation based on the NPMLE. While much of the asymptotic results will be developed along the lines of HUANG (1996), the characterization of the NPMLE is quite different from that for the proportional hazards model and will make use of interior point methods. A further exposition of this topic can be found in WRIGHT (1997). We analytically compare the relative efficiency of the procedure proposed by LIN *et al.* (1998) to the NPMLE. In addition, we perform a numerical comparison between the two methods based on data from a tumorigenicity experiment.

2 Computation of the NPMLE

In this section, we characterize the maximum likelihood estimator (MLE) $(\hat{\theta}, \hat{\Lambda}) \equiv (\hat{\theta}_n, \hat{\Lambda}_n)$ of (θ_0, Λ_0) for a fixed sample size n, where $\Lambda(t) = \int_0^t \lambda(s)\,\mathrm{d}s$ is the integrated or cumulative hazard. Here, θ_0 and Λ_0 are the true values of the regression parameters and baseline cumulative hazard. We will demonstrate that a proper characterization of the relevant optimization problem cannot make use of the results from HUANG (1996) so that a different approach is needed. We are able to provide necessary and sufficient conditions for the existence of the MLE and subsequently outline an algorithm for computing $(\hat{\theta}_n, \hat{\Lambda}_n)$.

In the rest of the paper, we assume that T and C are independent, conditional on Z and that the joint distribution (C, Z) is noninformative for θ and Λ. For a single observation $X = (C, \delta, Z)$, where $\delta = I(T \leqslant C)$, the pdf is proportional to

$$\begin{aligned} p_{\theta,F}(x) &= F(c|z)^{\delta}\bar{F}(c|z)^{1-\delta} \\ &= [1 - \bar{F}_0(c)\exp(-c\theta' Z)]^{\delta}[\bar{F}_0(c)\exp(-c\theta' Z)]^{1-\delta}, \end{aligned}$$

where $\bar{F} = 1 - F$. Thus, the log-likelihood for a single observation, up to a constant, is given by

$$l(\theta, F) = \delta \log[1 - \bar{F}_0(C)\exp(-C\theta' Z)] + (1 - \delta)[\log \bar{F}_0(C) - C\theta' Z].$$

Let (C_i, δ_i, Z_i), $i = 1, \ldots, n$, be an i.i.d. sample from (C, δ, Z). Then the log-likelihood for the sample can be written as

$$\begin{aligned} l_n(\theta, Z) = \sum_{i=1}^{n} (\delta_i \log\{1 - \exp[-\Lambda_0(C_i) - C_i\theta' Z_i]\} \\ - [1 - \delta_i][\Lambda_0(C_i) + C_i\theta' Z_i]), \end{aligned} \tag{2}$$

where we have used the relation $\Lambda_0 = -\log(1 - F_0)$.

Since the regression parameters in the additive hazards model act additively on the baseline hazard, there is an an inherent positivity constraint, i.e. the right hand side in (1) must be nonnegative. Let $C_{(1)} \leqslant C_{(2)} \leqslant \cdots \leqslant C_{(n)}$ denote the ordered censoring times, and let $\delta_{(i)}$, $Z_{(i)}$ correspond to $C_{(i)}$, $i = 1, \ldots, n$. In the likelihood function (2), only the values of Λ_0 at the $C_{(i)}$'s matter. Thus, the MLE $\hat{\Lambda}_n$ of Λ_0 will be a right-continuous increasing step function with jump points at $C_{(i)}$ and corresponding values $\hat{\Lambda}_n(C_{(i)})$, $i = 1, \ldots, n$, where $C_{(0)} = 0$ and $\hat{\Lambda}_n(0) = 0$. We have that

$$\hat{\Lambda}_n(y) = \begin{cases} 0 & 0 \leqslant y < C_{(1)} \\ \hat{\Lambda}_n(C_{(i)}) & C_{(i)} \leqslant y < C_{(i+1)}, \quad i = 1, \ldots, n - 1. \end{cases}$$

However, $\hat{\Lambda}_n(y)$ for $y > C_{(n)}$ is left unspecified.

Now define $\Lambda_{(i)} = \Lambda_0(C_{(i)})$. Then the proper optimization problem is to maximize

$$\begin{aligned} l(\theta, \Lambda_{(\cdot)}) = \sum_{i=1}^{n} \{\delta_{(i)} \log[1 - \exp(-\Lambda_{(i)} - C_{(i)}\theta' Z_{(i)})] \\ - (1 - \delta_{(i)})(\Lambda_{(i)} + C_{(i)}\theta' Z_{(i)})\} \end{aligned}$$

subject to

(C1) $\theta \in \Theta \subset R^p$;

(C2) $0 \leqslant \Lambda_{(1)} \leqslant \cdots \leqslant \Lambda_{(n)}$;

(C3) $\Lambda_{(i)} + C_{(i)}\theta' Z_{(i)} \geqslant 0$, $i = 1, \ldots, n$

in order to obtain MLE's of θ_0 and Λ_0. In the absence of the positivity constraint (C3), the optimization problem would be the same one as that considered by HUANG

(1996). However, due to the presence of (C3), we are not able to use Fenchel duality results directly as Huang did.

Nevertheless, we are able to derive a characterization of the NPMLE using results based on convex analysis, which can be found in any standard text, such as ROCKAFELLAR (1970). To simplify the presentation, for the rest of the paper, we will assume that $p = 1$. The generalization to $p > 1$ is straightforward. Let $v = (\Lambda_{(1)}, \Lambda_{(2)}, \ldots, \Lambda_{(n)}, \theta)$. Define g: $R^{n+1} \rightarrow R^{2n}$ by $g(v) = (g_1(v), \ldots, g_{2n}(v))'$, where

$$g_1(v) = -\Lambda_{(1)},$$

$$g_i(v) = \Lambda_{(i-1)} - \Lambda_{(i)},\ i = 2, \ldots, n,$$

$$g_{n+\mathrm{i}}(v) = -(\Lambda_{(i)} + C_{(i)}\theta Z_{(i)}),\ i = 1, \ldots, n.$$

Furthermore, define $G = [\partial g_i(v)/\partial v_j]$, $i = 1, \ldots, 2n$; $j = 1, \ldots, n+1$. Note that G does not depend on v due to the linearity in the side conditions. Now the optimization problem above can be re-expressed as trying to minimize $\phi(v)$, where

$$\phi(v) = -\sum_{i=1}^{n}\{\delta_{(i)}\log[1 - \exp(-\Lambda_{(i)} - C_{(i)}\theta Z_{(i)})] - (1 - \delta_{(i)})[\Lambda_{(i)} + C_{(i)}\theta Z_{(i)}]\}$$

subject to $g(v) \leqslant 0$, $v \in R^{n+1}$. Define $\nabla\phi$ to be the gradient of ϕ and $\langle u, w\rangle$ to be the dot product vector between two vectors u and w. Based on the optimization problem, we have the following characterization theorem.

THEOREM 1. *Let $\hat{v} = (\hat{\Lambda}_{(1)}, \hat{\Lambda}_{(2)}, \ldots, \hat{\Lambda}_{(n)}, \hat{\theta})$ be a vector in R^{n+1} such that $\phi(\hat{v}) < \infty$. Then $\hat{v}$ minimizes $\phi(v)$ over the set $\{v: v \in R^{n+1}, g(v) \leqslant 0\}$ iff the following conditions are satisfied:*

$$\nabla\phi(\hat{v}) + G'\omega = 0$$

$$g(\hat{v}) + s = 0$$

$$\langle\omega, s\rangle = 0$$

for vectors ω and s in R_+^{2n}.

The proof of this theorem is not stated here, as it is a simple modification of that found in GRONENBOOM (1998). Theorem 1 leads to an algorithm for computing $(\hat{\theta}_n, \hat{\Lambda}_n)$ based on primal-dual interior point methods (WRIGHT, 1997). We now briefly outline the computational procedure. It is discussed in greater detail in another setting by GRONENBOOM (1998).

We start with $v_0 \in R^{n+1}$ such that $g(v_0) < 0$. We take $\omega_0 = s_0 = e$, where e is the vector in R_+^{2n} with each component equal to 1. Define ϕ_ω: $R^{n+1} \rightarrow R$ by $\phi_\omega(v) = \phi(v) + \langle\omega, g(v)\rangle$. Then the first iteration of the algorithm solves

$$\begin{pmatrix} \nabla_{v_0 v_0}\phi_{\omega_0}(v_0) & \mathrm{G}' & 0 \\ G & 0 & I \\ 0 & S_0 & \Omega_0 \end{pmatrix} \begin{pmatrix} v - v_0 \\ \omega - \omega_0 \\ s - s_0 \end{pmatrix} = - \begin{pmatrix} \nabla\phi_{\omega_0}(v_0) \\ g(v_0) + s_0 \\ (\Omega_0 S_0 - \sigma\mu_0)e \end{pmatrix}, \tag{3}$$

where S_0 and Ω_0 are diagonal matrices with diagonal elements s_0 and ω_0, respectively, I represents the identity matrix, $\nabla_{vv}\phi_\omega(v)$ is the Hessian of ϕ_ω with respect to v, and μ_0 and σ are tuning parameters whose values are set to 1 and $1/2$. For a fixed $\beta > 0$, define $N(\mu)$ as

$$N(\mu) = \{(v, \omega, s)\colon \|\nabla_v \phi_\omega(v)\| \leqslant \beta\mu,\ \|g(v) + s\|$$

$$\leqslant \beta\mu,\ \omega \geqslant 0,\ s \geqslant 0,\ \omega_i s_i \geqslant \mu/2,\ 1 \leqslant i \leqslant 2n\},$$

where $\|\cdot\|$ denotes the Euclidean norm, and μ is the duality measure, defined by

$$\mu \equiv \frac{1}{2n}\langle \omega, \mathrm{s}\rangle.$$

Taking $\omega_0 = s_0 = e$ gives a value of $\mu = 1$ for the first iteration.

We then choose a number $\gamma \in (0, 1)$ and take α as the first number in the sequence $\{1, \gamma, \gamma^2, \gamma^3, \ldots\}$ such that

$$(v(\alpha), \omega(\alpha), s(\alpha)) \equiv (v_0, \omega_0, s_0) + \alpha(v - v_0, \omega - \omega_0, s - s_0) \in N(\mu_0),$$

where (v, ω, s) solves (3) and

$$\mu(\alpha) = \frac{1}{2n}\langle \omega(\alpha), s(\alpha)\rangle \leqslant (1 - .01\alpha)\mu_0.$$

One then takes $(v_1, \omega_1, s_1) = (v(\alpha), \omega(\alpha), s(\alpha))$ and $\mu_1 = \mu(\alpha)$; based on these values, we solve the following system of equations:

$$\begin{pmatrix} \nabla_{v_1 v_1}\phi_{\omega_1}(v_1) & G' & 0 \\ G & 0 & I \\ 0 & S_1 & \Omega_1 \end{pmatrix} \begin{pmatrix} v - v_1 \\ \omega - \omega_1 \\ s - s_1 \end{pmatrix} = - \begin{pmatrix} \nabla\phi_{\omega_1}(v_1) \\ g(v_1) + s_1 \\ (\Omega_1 S_1 - \sigma\mu_1)e \end{pmatrix}$$

and find the new $(v(\alpha), \omega(\alpha), \mathrm{s}(\alpha))$ required to lie in $N(\mu_1)$ for this system, (v_2, ω_2, s_2) and the new μ_2. This procedure is repeated until μ_k is below a certain tolerance criterion, such as 10^{-10} or 10^{-15}.

3 Main Results

Having characterized the NPMLE, we are now in a position to investigate the asymptotic properties of the estimator. In this section, we state the main theoretical results. Since most of the proofs of the results involve simple modifications of the arguments given by HUANG (1996), they will not be given here. The results are stated for the sake of completeness. It is quite intuitive to expect that similar asymptotic results would hold for the NPMLE in the additive and proportional hazards models with current status data.

Again, for notational convenience, we will stick to the case where $p = 1$; the extension of the results to general p is straightforward. First, we make the following assumptions, the same as those in HUANG (1996):

(A1) Θ is a bounded subset of R.
(A2) Z has bounded support.
(A3) For any $\theta \neq \theta_0$, $P(\theta Z \neq \theta_0 Z) > 0$.
(A4) $F_0(0) = 0$. Let $\tau_{F_0} \equiv \inf\{t\colon F_0(t) = 1\}$. Define the support of C to be an interval $S[C] = [l_C, u_C]$, where $0 \leqslant l_C \leqslant u_C < \tau_{F_0}$.
(A4′) Everything is the same as in (A4), except that $0 < l_C \leqslant u_C < \tau_{\mathrm{F}_0}$.
(A5) Λ_0 has strictly positive derivative on $S[C]$ and the joint distribution function $G(c, z)$ of (C, Z) has bounded second-order (partial) derivative with respect to c.

Before developing the asymptotics of the NPMLE, it is useful to compute the information bound for θ. With interval-censored data, since the exact time to the event is never observed, it is not intuitively clear that there will be positive information for the regression parameter. As stated in the next theorem, however, there is indeed positive information. Before stating the theorem, define $a^{\otimes 2} = aa'$,

$$Q(c, \delta, z) = \frac{\delta \overline{F}(c|z)}{1 - \overline{F}(c|z)} - (1 - \delta)$$

and $O(c|z) = E[Q^2(C, \delta, Z)|C = c, Z = z] = \overline{F}(c|z)/[1 - \overline{F}(c|z)]$.

THEOREM 2. *Suppose Assumptions (A3) and (A4) are satisfied. Then*

(a) *the efficient score for θ is given by*

$$\dot{l}^*_\theta(x) = Q(c, \delta, z)c\left(z - \frac{E[ZO(C|Z)|C = c]}{E[O(C|Z)|C = c]}\right).$$

(b) *The information for θ is given by*

$$I(\theta) = E[\dot{l}^*_\theta(X)]^{\otimes 2} = E\left\{O(C|Z)\left[C\left(Z - \frac{E[ZO(C|Z)|C]}{E[O(C|Z)|C]}\right)\right]^{\otimes 2}\right\}.$$

Having computed the semiparametric information bound for θ, we can now determine the asymptotic behavior of the $\hat{\theta}_n$ and $\hat{\Lambda}_n$. The next theorem states the consistency of $\hat{\theta}_n$ and $\hat{\Lambda}_n$ on the support of C.

THEOREM 3. *Suppose that assumptions (A1)–(A4) hold. Then*

$$\hat{\theta}_n \to \theta_0 \quad \text{a.s.,}$$

and if $y \in S[C]$ is a continuity point of F_0, then

$$\hat{\Lambda}_n(y) \to \Lambda_0(y) \quad \text{a.s.}$$

Furthermore, if F_0 is continuous, then

$$\sup_{y \in S[C]} |\hat{\Lambda}_n(y) - \Lambda_0(y)| \to 0 \quad \text{a.s.}$$

Consistency of the NPMLEs is a useful first step in determining the asymptotic behavior of the estimators. The next step is to determine their rate of convergence. In order to do this, it is necessary to define an appropriate metric. One can define a metric d on $R \times \Phi$, where

$$\Phi =$$

$$\{\Lambda\colon \Lambda \text{ is increasing and } 0 < 1/M \leqslant \Lambda(y) \leqslant M < \infty \text{ for all } y \in S[C]\},$$

as follows; d is defined as

$$d((\theta_1, \Lambda_1), (\theta_2, \Lambda_2)) = |\theta_1 - \theta_2| + \|\Lambda_1 - \Lambda_2\|_2,$$

where $|a|$ denotes the usual Euclidean distance in R and

$$\|\Lambda_1 - \Lambda_2\|_2 = \left\{ \int [\Lambda_1(y) - \Lambda_2(y)]^2 \, \mathrm{d}Q_C(y) \right\}^{1/2}$$

is the L_2 distance between Λ_1 and Λ_2 with respect to Q_C, the probability measure of C. We now have the following result.

THEOREM 4. *Suppose that assumptions (A1)–(A3) and (A4′) are satisfied. Then*

$$d((\hat{\theta}_n, \hat{\Lambda}_n), (\theta_0, \Lambda_0)) = O_P(n^{-1/3}).$$

Roughly, what Theorem 4 states is that the rate of the convergence of the NPMLE is dominated by that of the infinite-dimensional parameter. However, the hope is that the NPMLE of the regression parameter converges at a $n^{1/2}$ rate. By performing some calculations using empirical process methods, such as those in VAN DER VAART and WELLNER (1996), one can apply Theorem 6.1 in HUANG (1996) to demonstrate asymptotically normality of $\hat{\theta}_n$.

THEOREM 5. *Suppose that θ_0 is an interior point of Θ and that assumptions (A1)–(A3), (A4′) and (A5) are satisfied. Then*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = I(\theta_0)^{-1}\sqrt{n}P_n l^*_{\theta_0}(x) + o_P(1) \to {}_d N(0, I(\theta_0)^{-1}),$$

where P_n is the empirical measure of (C_i, δ_i, Z_i), $i = 1, \ldots, n$ and $l^*_{\theta_0}(x)$ and $I(\theta_0)^{-1}$ are defined as in Theorem 2.

Thus, just as for the proportional hazards model, the NPMLE of the regression parameters converges at a regular rate and is fully efficient for the semiparametric variance bound in the additive hazards model.

4 Analytical comparison with LIN *et al.* (1998)

One of the main results from the previous section was the calculation of the semiparametric information bound for θ. This result allows us to examine the loss in efficiency from other procedures. As was mentioned in the introduction, LIN *et al.* (1998) proposed an estimating function approach to estimation of the regression parameters in (1) with current status data. While the general formula for the relative efficiency is not very informative, we consider a special case.

Suppose that Z is scalar, C is independent of Z, $\theta = 0$, and $P(T \geqslant C) \in (0, 1)$. Let $I(\theta)$ be the information bound derived in Section 3. Let $I_{LOY}(\theta)$ be the information for θ, based on equation (2.2) in LIN *et al.* (1998). The efficiency of the LIN *et al.* procedure can be shown to be

$$\frac{I_{LOY}\theta)}{I(\theta)} = \frac{\{E[C^2]\,\mathrm{Var}(Z) + E[Z]^2\,\mathrm{Var}(C)\}P(T \leqslant C)}{E[C^2]\,\mathrm{Var}(Z)}. \tag{4}$$

If $E[Z] = 0$, (4) reduces to $P(T \leqslant C)$. Given a distribution for T and C, we can calculate this probability quiteeasily. For example, if T and C are independent exponential random variables, each with rate 1, then the efficiency of the LIN *et al.* (1998) procedure is 0.5.

5 Numerical Example

We now compare the results given by the LIN *et al.* (1998) and NPMLE procedures on a real dataset. We consider data from a tumorigenicity study, first analyzed by HOEL and WALBURG (1972). The objective of the study was to compare the time to lung tumor development between two environments, germ-free and conventional. The study involved 144 RFM mice; 96 were assigned to the conventional environment, while 48 were assigned to the germ-free environment. The mice were followed until sacrifice or death. At this time, it was determined whether or not a lung tumor had developed. Since these tumors are nonlethal, it is reasonable to treat the times to lung tumor development as current status data. Of the 96 mice in the conventional environment group, 27 had lung tumors at the time of sacrifice or death, while 35 of the 48 mice in the germ-free environment arm had lung tumors. Here, we focus on the effect of environment on time to tumor devevelopment.

As was noted by LIN *et al.* (1998), there was a significant difference in monitoring times between mice in the two environments. Using their dependent monitoring procedure, their estimate of the treatment effect was -0.00071 with a standard error of 0.00041. Using the interior point algorithm described in Section 2, we get an estimated treatment effect of -0.00065.

The issue of estimating the standard error for the treatment effect now arises. To estimate $I(\theta_0)$, we follow the development in Section 4 of HUANG (1996). Having computed estimators $\hat{\theta}_n$ and $\hat{F}_n$ of θ_0 and F_0, we estimate $O(c|z)$ by

$$\hat{O}_n(c|z) = [1 - \hat{F}_n(c|z)]/\hat{F}_n(c|z),$$

where $\hat{F}_n(c|z) = \hat{F}_n(c)\exp(-c\hat{\theta}_n z)$. Let $P(Z = 1) = p$. Since the covariate encoding environment is binary, we note that

$$E[O(C|Z)|C = c] = O(c|1)P(Z = 1|C = c) + O(c|0)P(Z = 0|C = c)$$

$$= \frac{O(c|1)f_1(c)p}{f(c)} + \frac{O(c|0)f_0(c)(1 - p)}{f(c)},$$

where $f_k(c)$ is the conditional density of C given $Z = k(k = 0, 1)$ and $f(y)$ is the marginal density of C. It can be shown similarly that $E[ZO(C|Z)|C = c] = O(c|1)f_1(c)p$. If we let $\hat{p}_n$ denote the empirical estimator of $P(Z = 1)$ and $\hat{f}_{kn}(c)$ represent a kernel density estimate of $f_k(c)$ $(k = 0, 1)$, a natural estimator of $\mu \equiv E[ZO(C|Z)|C = c]/E[O(C|Z)|C = c]$ is given by

$$\hat{\mu}_n(c) = \frac{\hat{O}_n(c|1)\hat{f}_{1n}(c)\hat{p}_n}{\hat{O}_n(c|1)\hat{f}_{1n}(c)\hat{p}_n + \hat{O}_n(c|0)\hat{f}_{0n}(c)(1 - \hat{p}_n)}.$$

It can be shown that for an appropriate choice of bandwidth and kernel, $\hat{\mu}_n$ is a consistent estimator of μ. This leads to the following estimator of $I(\theta_0)$:

$$\hat{I}_n(\hat{\theta}_n) = n^{-1}\sum_{i=1}^{n}\{\hat{O}_n(C_i|Z_i)[C_i(Z_i - \hat{\mu}_n(C_i))]^2\}. \tag{5}$$

Using equation (5) with the tumorigenicity data yields an estimated standard error of 0.00037.

6 Conclusion

In this paper, we have explored the issue of efficient estimation in the additive hazards model with current status data. Theoretically, results analogous to those derived by HUANG (1996) hold here as well. However, the characterization of the NPMLE in this problem is quite different from that for the proportional hazards model due to the presence of the positivity constraint (C3).

Similar methods can be applied to the analysis of interval censored data under alternative observation schemes. For example, ZHANG (1998) studied the problem of analysis of panel count data, which is a generalization of the problem considered here. He proposed nonparametric maximum pseudolikelihood estimation (NPMPLE) in both the one-sample and regression problems. The method proposed here can be generalized similarly.

While we made an analytic comparison between the procedure of LIN *et al.* (1998) with the NPMLE, the behavior in small samples has yet to be investigated. This is definitely of interest in practice and would be a worthwhile avenue to explore.

Acknowledgements

The author would like to thank Danyu Lin for suggesting this problem. He would also like to thank Jon Wellner for revision of an earlier version of the manuscript and Piet Groeneboom for helpful discussions about the topics covered in the paper.

References

AYER, M., H. D. BRUNK, G. M. EWING, W. T. REID and E. SILVERMAN (1955), An empirical distribution function for sampling with incomplete information, *Annals of Mathematical Statistics* **26**, 641–647.

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

DIAMOND, I. D. and J. W. MCDONALD (1991), Analysis of current status data, in: J. TRUSSELL R. HANKINSON and J. TILTON (eds.), *Demographic applications of event history analysis*, Oxford University Press, Oxford 231–252.

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

GRONENBOOM, P. (1998), Lecture notes on inverse problems, Technical Report, Department of Statistics, University of Washington.

GRONENBOOM, P. and J. A. WELLNER (1992), *Information bounds and nonparametric maximum likelihood estimation*, DMV Seminar Band 19, Birkhäuser, Basel.

HOEL, D. G. and H. E. WALBURG (1972), Statistical analysis of survival experiments, *Journal of the National Cancer Institute* **49**, 361–372.

HUANG, J. (1996), Efficient estimation for the propotional hazards with interval censoring, *Annals of Statistics* **24** 540–568.

HUANG, J. and J. A. WELLNER (1997), Interval censored survival data: a review of recent progress, in D. Y. LIN and T. R. FLEMING (eds.), *Proceedings of the first Seattle symposium in biostatistics: survival analysis*, Springer-Verlag, New York, 123–169.

JEWELL, N. P., H. M. MALANI and E. VITTINGHOFF (1994), Nonparametric estimation for a form of doubly censored data, with application to two problems in AIDS, *Journal of the American Statistical Association* **89** 7–18.

KAPLAN, E. L. and P. MEIER (1958), Non-parametric estimation from incomplete observations, *Journal of the American Statistical Association* **53**, 457–481.

LIN, D. Y., D. OAKES and Z. YING (1998), Additive hazards regression with current status data, *Biometrika* **85**, 289–298.

LIN, D. Y., and Z. YING (1994), Semiparametric analysis of the additive risk model, *Biometrika* **81**, 61–71.

PETO, R. (1973), Experimental survival curves for interval-censored data, *Applied Statistics* **22**, 86–91.

ROCKAFELLAR, R. T. (1970), *Convex analysis*, Princeton University Press.

SHIBOSKI, S. C. and N. P. JEWELL (1992), Statistical analysis of the time dependence of HIV infectivity based on partner study data, *Journal of the American Statistical Association*, **87**, 360–372.

TURNBULL, B. W. (1976), The empirical distribution function with arbitrarily grouped, censored and truncated data, *Journal of the Royal Statististical Society* B **38**, 290–5.

VAN DER VAART, A. W. and J. A. WELLNER (1996), *Weak convergence and empirical processes*, Springer Verlag, New York.

WRIGHT, S. J. (1997), *Primal-dual interior-point methods*, SIAM, Philadelphia.

ZHANG, Y. (1998), Estimation for counting processes with incomplete data, Ph.D. dissertation, Department of Statistics, University of Washington.

Received: May 1999. Revised: September 2000