© Board of the Foundation of the Scandinavian Journal of Statistics 1998. Published by Blackwell Publishers Ltd., 108 Cowley Road, Oxford OX4 1JF, UK and 238 Main Street, Cambridge, MA 02142, USA Vol 25: 589–601, 1998

Product-limit Estimators and Cox Regression with Missing Censoring Information

IAN W. McKEAGUE

Florida State University

SUNDARRAMAN SUBRAMANIAN

University of Maine, Orono

ABSTRACT. The Kaplan–Meier estimator of a survival function requires that the censoring indicator is always observed. A method of survival function estimation is developed when the censoring indicators are missing completely at random (MCAR). The resulting estimator is a smooth functional of the Nelson–Aalen estimators of certain cumulative transition intensities. The asymptotic properties of this estimator are derived. A simulation study shows that the proposed estimator has greater efficiency than competing MCAR-based estimators. The approach is extended to the Cox model setting for the estimation of a conditional survival function given a covariate.

Key words: counting processes, incomplete data, Nelson–Aalen estimators, product integral, right censorship

1. Introduction

In this article we study inference from right-censored survival data in which the censoring indicator can be missing for some individuals. Let T denote the failure time of interest, Z a covariate vector, and let C be a censoring time that is conditionally independent of T given Z. We are given data on $X = T \wedge C$ and also on any covariates Z that are to be included in the analysis, but the indicator $\delta = I(T \leqslant C)$ is allowed to be missing. The problem is to estimate the survival function of T and to account for the covariate effect. Apart from the possibility that δ is missing, this is the classical survival analysis framework.

The censoring indicator δ may be missing for a variety of reasons, e.g. in a bioassay experiment some subjects might not be autopsied to save expense, or the results of an autopsy may be inconclusive; in population mortality studies relevant death certificate information can be missing due to emigration. We shall assume that the censoring indicators are *missing completely at random* (MCAR) in the sense that the mechanism for missingness is independent of everything else. MCAR is a special case of missing at random (MAR), which was introduced by Rubin (1976), see also Little & Rubin (1987), Heitjan & Rubin (1991), Jacobsen & Keiding (1995), and Gill *et al.* (1997).

We first consider the problem in the absence of covariates. The survival function of T, denoted S, can be consistently estimated under MCAR by simply ignoring the missing data and applying the usual Kaplan–Meier estimator to the complete data. However, such a procedure (called the *complete case estimator*) would be highly inefficient if there is a significant degree of missingness.

The first attempt to improve upon the complete case estimator was made by Dinse (1982) who used the EM algorithm to obtain a non-parametric maximum likelihood estimator (NPMLE). Lo (1991) showed that there are infinitely many NPMLEs, and each of them is self-consistent in the sense of Efron (1967). However, Lo noted that some of the NPMLEs may be inconsistent. He constructed two alternative estimators, one of which is consistent and asymptotically normal. Gijbels *et al.* (1993), further improved upon this estimator by taking a product-integral of a convex combination of two estimates of the cumulative hazard function of T.

Goetghebeur & Ryan (1995) considered a competing risks model in which cause-of-failure is missing at random. Their model is more general than ours in allowing two failure types having *dependent* latent failure times, and in allowing MAR instead of MCAR, but the two cause-specific hazard rates (or baseline hazards in the case of a Cox model) are assumed to be proportional. Instead, we consider the classical Kaplan–Meier model with independent censoring, but without any restriction on the hazard rates.

We shall construct a new estimator for S. The observations are considered to be i.i.d. replicates of (X, ξ, σ) where ξ is the indicator that δ is not missing and $\sigma = \xi\delta$. It is assumed that ξ is independent of T and C, and $\rho = P(\xi = 1) > 0$. We will find it convenient to view this set-up as a four state non-time-homogeneous Markov process having state space $\{0, 11, 10, 00\}$, where "0" represents the initial state in which no failure has occurred, "11" and "10" represent $(\xi, \delta) = (1, 1)$ and $(1, 0)$, respectively, and "00" represents a missing censoring indicator. The process jumps at time X from state 0 into one of the absorbing states 11, 10, or 00.

The cumulative transition intensities Λ_{11}, Λ_{10}, Λ_{00} for the transitions $0 \to 11$, $0 \to 10$, $0 \to 00$ can be expressed in terms of the cumulative hazard function of T (denoted by Λ_T), ρ and the survival function of C. The cumulative hazard function Λ_T can in turn be expressed as the functional (see section 2):

$$\Lambda_T(t) = \phi(\Lambda_{10}, \Lambda_{11}, \Lambda_{00})(t) = \Lambda_{11}(t) + \pi(t)\Lambda_{00}(t), \tag{1.1}$$

where $\pi(\cdot) = \Lambda_{11}/(\Lambda_{10} + \Lambda_{11})$. We shall plug into (1.1) the Nelson–Aalen estimators $\hat{\Lambda}_{10}$, $\hat{\Lambda}_{11}$, and $\hat{\Lambda}_{00}$ of Λ_{10}, Λ_{11}, and Λ_{00}. Finally, by taking a product integral we will obtain our estimator $\hat{S}$. This estimator is easier to compute than the Gijbels *et al.* estimator, which involves a further step of estimating an optimal convex combination of two 'sub-optimal' estimators, and has superior small sample performance, see section 2. Moreover, its asymptotic properties are relatively straightforward to derive, being a consequence of (1.1) and standard results for Nelson–Aalen estimators, and a specially tailored analysis is not required.

Note that $\hat{\Lambda}_{11}$ underestimates Λ_T and we need to compensate by adding a suitably scaled estimate of Λ_{00}. The factor $\pi(t)$ represents the proportion of transitions to 00 that are uncensored failures. In the 'full data' case ($\rho = 1$) we have $\hat{\Lambda}_{00} = 0$, and $\rho = 1$, so $\hat{\Lambda}_T$ reduces to the Nelson–Aalen estimator, and $\hat{S}$ reduces to the Kaplan–Meier estimator.

We next consider an extension of our approach to the setting of Cox's (1972) proportional hazards model in order to account for the effect of the covariate Z. Here the conditional hazard function of T given $Z = z$ takes the form $\lambda(t|z) = \lambda_0(t) \exp(\beta' z)$, where β is a vector of regression parameters and λ_0 is a baseline hazard function. We assume that ξ is independent of (T, C, Z). If only the complete case data are used, then Cox's maximum partial likelihood estimator of β is consistent and asymptotically normal, see Andersen & Gill (1982). Gijbels *et al.* (1993) introduced a more efficient estimator of β based on an optimal linear combination of estimating equations corresponding to the $0 \to 11$ and $0 \to 00$ transitions, and they showed that considerable improvements over the complete case estimator are possible. We propose an alternative estimating equation that treats the $0 \to 00$ transitions in a way that is similar to our derivation of $\hat{S}$. Our estimator improves upon the overall performance of Gijbel *et al.* estimator and it is computationally simpler. We also show how our basic estimator $\hat{S}$ can be extended to the Cox model setting for estimation of the conditional survival function of T given Z.

Van der Laan & McKeague (1998) recently obtained an asymptotically efficient estimator of S in the more general setting in which δ is missing at random (MAR), i.e. ρ is an unknown function of time. Although their estimator is efficient, it requires an artificial binning of the data which can be unappealing in practice, especially for small samples. Our estimator $\hat{S}$ is inconsistent under general MAR, but it does not require binning of the data or any other smoothing technique, and is expected to have a better small sample performance under MCAR.

The paper is organized as follows. Section 2 deals with estimation of the survival function in the absence of covariates. We give a brief review of existing estimators in section 2.1. The proposed estimator and its asymptotic properties are presented in section 2.2, and some numerical results are discussed in section 2.3. We treat estimation for the Cox model in section 3. Existing estimators of β are reviewed in section 3.1. We propose a new estimator of β and discuss its asymptotic properties in section 3.2. An estimator of the conditional survival function of T given the covariate is introduced in section 3.3. Some numerical results are given in section 3.4. An important direction for further research is discussed in section 3.5. All proofs are placed in the appendix.

2. Estimation of a survival function

2.1. Review of existing estimators

The observations consist of n i.i.d. replicates (X_i, ξ_i, σ_i) of the generic triple (X, ξ, σ) that was defined in the introduction. Lo's (1991) estimator of the survival function S is given by

$$\hat{S}_L(t) = \prod_{X_i \leqslant t} \left(1 - \frac{\sigma_i}{r(X_i)}\right)^{1/\hat{\rho}},$$

where $\hat{\rho} = \sum_{i=1}^n \xi_i / n$ is the proportion of observed censoring indicators, and $r(t) = \#\{j\colon X_j \geqslant t\}$ is the "size of the risk set" at time t. Like the complete case estimator, $\hat{S}_L$ jumps only at the uncensored failure times with known failure status. Note, however, that $\hat{S}_L$ uses the full size of the risk set that would be used by the Kaplan–Meier estimator if all the δ_i's were observed, so it makes more efficient use of the available data than the complete case estimator.

Gijbels *et al.* (1993) observe that $\hat{S}_L$ does not use all of the information from individuals with $\xi_i = 0$, and they find a way to make better use of such information. They first estimate Λ_T and then estimate S by taking a product integral (as in (2.3)). Their estimator of Λ_T can be expressed as

$$\hat{\Lambda}_{\mathrm{GLY}}(t) = \alpha(t)\hat{\Lambda}_1(t) + (1 - \alpha(t))\hat{\Lambda}_2(t), \tag{2.1}$$

where $0 \leqslant \alpha(t) \leqslant 1$ is specified,

$$\hat{\Lambda}_1(t) = \hat{\Lambda}_{11}(t)/\hat{\rho},$$

$$\hat{\Lambda}_2(t) = \hat{\Lambda}_{00}(t)/(1 - \hat{\rho}) - \hat{\Lambda}_{10}(t)/\hat{\rho}.$$

They estimate the function $\alpha(t)$ that minimizes the asymptotic variance of estimators of the form (2.1), and plug that back into (2.1). They point out that the choice $\alpha(t) = 1$ is equivalent to Lo's estimator.

2.2. Proposed estimator

The distinction between our approach and those of Lo (1991) and Gijbels *et al.* (1993) is that their estimators involve $\hat{\rho}$, whereas the estimator $\hat{S}$ to be proposed in this section does not. We believe that our approach is preferable because $\hat{\rho}$ is an ancillary statistic under the MCAR model (its distribution is independent of the parameter of interest), so it does not contribute any information about S.

The distribution functions of T and C are denoted F and G, respectively, and the corresponding survival functions by $S = 1 - F$ and $R = 1 - G$. The subdistribution functions $H_{kl}(t) = P(X \leqslant t, \xi = k, \sigma = l)$ for $(k, l) \in \Gamma = \{(1, 0), (1, 1), (0, 0)\}$, can be written as

$$
\begin{aligned}
H_{10}(t) &= \rho \int_0^t S_- dG, \\
H_{11}(t) &= \rho \int_0^t R_- dF, \qquad (2.5)\\
H_{00}(t) &= (1-\rho) \int_0^t (S_- dG + R_- dF),
\end{aligned}
$$

where $S_-(t) = S(t-)$ is the left-continuous version of S. The range of integration in (2.2) is $[0, t]$. The survival function $\bar{H}$ of X may be expressed in terms of these subdistribution functions as $\bar{H} = 1 - \sum_{(k,l)\in\Gamma} H_{kl}$.

We shall consider estimation of S over an interval $[0, \tau]$, where $\bar{H}(\tau) > 0$. The three cumulative transition intensities for the non-time-homogeneous Markov process discussed in the introduction are defined on $[0, \tau]$ by $d\Lambda_{kl} = dH_{kl}/\bar{H}_-$, for $(k, l) \in \Gamma$. Let Λ_X denote the cumulative hazard function of X. Some basic equations relating these functions on $[0, \tau]$ are

$$
\begin{aligned}
\Lambda_X &= \Lambda_{10} + \Lambda_{11} + \Lambda_{00}, \\
\rho\Lambda_X &= \Lambda_{10} + \Lambda_{11}, \\
\rho\Lambda_T &= \Lambda_{11}.
\end{aligned}
$$

These identities can be checked directly from (2.2), or more readily using the interpretation of the Λ_{kl} as cumulative transition intensities. The fundamental identity (1.1) is obtained by eliminating ρ from the last equation using the first two.

The Nelson–Aalen estimators of the cumulative transition intensities Λ_{kl} are given by

$$
\hat{\Lambda}_{kl}(t) = \int_0^t \frac{d\hat{H}_{kl}(s)}{\hat{\bar{H}}(s-)},
$$

where $\hat{\bar{H}}$ and $\hat{H}_{kl}$ are the empirical analogues of $\bar{H}$ and H_{kl}. The estimator proposed by Lo and the $\hat{\Lambda}_1$ and $\hat{\Lambda}_2$ estimators proposed by Gijbels *et al.* (1993) can be obtained from our basic equations by plugging-in the relevant Nelson–Aalen estimators and $\hat{\rho}$. Specifically, Lo's estimator and $\hat{\Lambda}_1$ arise from our third basic equation $\rho\Lambda_T = \Lambda_{11}$, and $\hat{\Lambda}_2$ arises from $\Lambda_T = \Lambda_{00}/(1-\rho) - \Lambda_{10}/\rho$.

Our approach is to estimate Λ_T by plugging the Nelson–Aalen estimators $\hat{\Lambda}_{kl}$ into (1.1) to obtain

$$
\hat{\Lambda}_T(t) = \hat{\Lambda}_{11}(t) + \hat{\pi}(t)\hat{\Lambda}_{00}(t),
$$

where

$$
\hat{\pi}(t) = \frac{\hat{\Lambda}_{11}(t)}{\hat{\Lambda}_{10}(t) + \hat{\Lambda}_{11}(t)}.
$$

Here $\hat{\pi}(t)$ is defined to be 0 when the denominator vanishes. Finally, our estimator of the survival function S is the product-limit estimator

$$
\hat{S}(t) = \Pi_{(0,t]}(1 - d\hat{\Lambda}_T(s)), \qquad (2.6)
$$

where Π denotes the product integral as defined by Gill & Johansen (1990).

The following result gives the asymptotic distribution of $\hat{S}$.

Theorem 1

Suppose that F and G are absolutely continuous. Then $\sqrt{n}(\hat{S} - S)$ *converges weakly in* $D[0, \tau]$ *to* $S \cdot L$, *where L is a continuous zero-mean Gaussian process with covariance function*

$$\begin{aligned}\operatorname{cov}(L(u), L(v)) &= D_{11}(u)D_{11}(v)C_{00}(u) + D_{00}(u)D_{00}(v)D_{11}(u)D_{11}(v)C_{10}(u) \\ &\quad + (1 + D_{00}(u)D_{10}(u))(1 + D_{00}(v)D_{10}(v))C_{11}(u) \qquad (2.7)\end{aligned}$$

for $u \leqslant v$, *and where*

$$C_{kl}(u) = \int_0^u \frac{d\Lambda_{kl}(s)}{\bar{H}(s-)}, \qquad D_{kl}(u) = \frac{\Lambda_{kl}(u)}{\Lambda_{10}(u) + \Lambda_{11}(u)}.$$

An asymptotic $100(1 - \alpha)\%$ pointwise confidence interval for $S(t)$ is given by

$$\hat{S}(t) \pm z_{\alpha/2} n^{-1/2} \hat{V}(t),$$

where z_α is the upper α-quantile of the standard normal distribution and $\hat{V}(t)$ is a consistent estimate of the asymptotic variance of $\hat{S}(t)$ obtained by replacing the unknown functions in $V(t) = S(t)^2 \operatorname{var}(L(t))$ by their estimates. Simultaneous confidence bands for S can not readily be obtained from the above result, however, because of the complexity of the limiting covariance function.

The proof of theorem 1, given in the appendix, uses the functional delta method of Gill (1989) in a similar way to Gill & Johansen's (1990) derivation of the asymptotic distribution of the Kaplan–Meier estimator. This approach works because $\hat{S}$ is a smooth (compactly differentiable) functional of the Nelson–Aalen estimators $\hat{\Lambda}_{kl}$. Moreover, due to the interpretation of the missing censoring indicator model as a non-time-homogeneous Markov process, we can appeal to the standard weak convergence result for Nelson–Aalen estimators in that context. This makes the proof relatively straightforward, and avoids the lengthy covariance calculations made by Gijbels *et al.* (1993).

2.3. Numerical results

In this section we report the results of a simulation study comparing the performance of our estimator with that of the Gijbels *et al.* estimator. We based the comparison on mean integrated square error (MISE) over the follow-up period. The failure time and censoring distributions are taken to be exponential with parameters 1 and λ, respectively, with λ adjusted to give prescribed censoring rates.

Figure 1 gives plots of the MISE as a function of the probability of non-missingness ρ for censoring rates 90%, 75%, 50%, sample size 100, and follow-up period [0, 1]. The MISE was estimated over a fine grid of values of ρ using 10,000 samples at each point on the grid.

In all cases the proposed estimator is found to be more efficient than the Gijbels *et al.* estimator. Moreover, the relative efficiency of $\hat{S}$ with respect to $\hat{S}_{\mathrm{GLY}}$, where $\hat{S}_{\mathrm{GLY}}$ denotes the survival function estimator of Gijbels *et al.* (1993), increases as ρ decreases and as the censoring rate increases. The proposed estimator makes significant improvements over $\hat{S}_{\mathrm{GLY}}$ when the censoring rate is at least 70% and at least 70% of the censoring indicators are missing.

3. Adjustment for covariates using the Cox model

In many applications it is necessary to adjust estimates of the survival distribution for the presence of risk factors, and to assess the influence of those risk factors. The standard way of

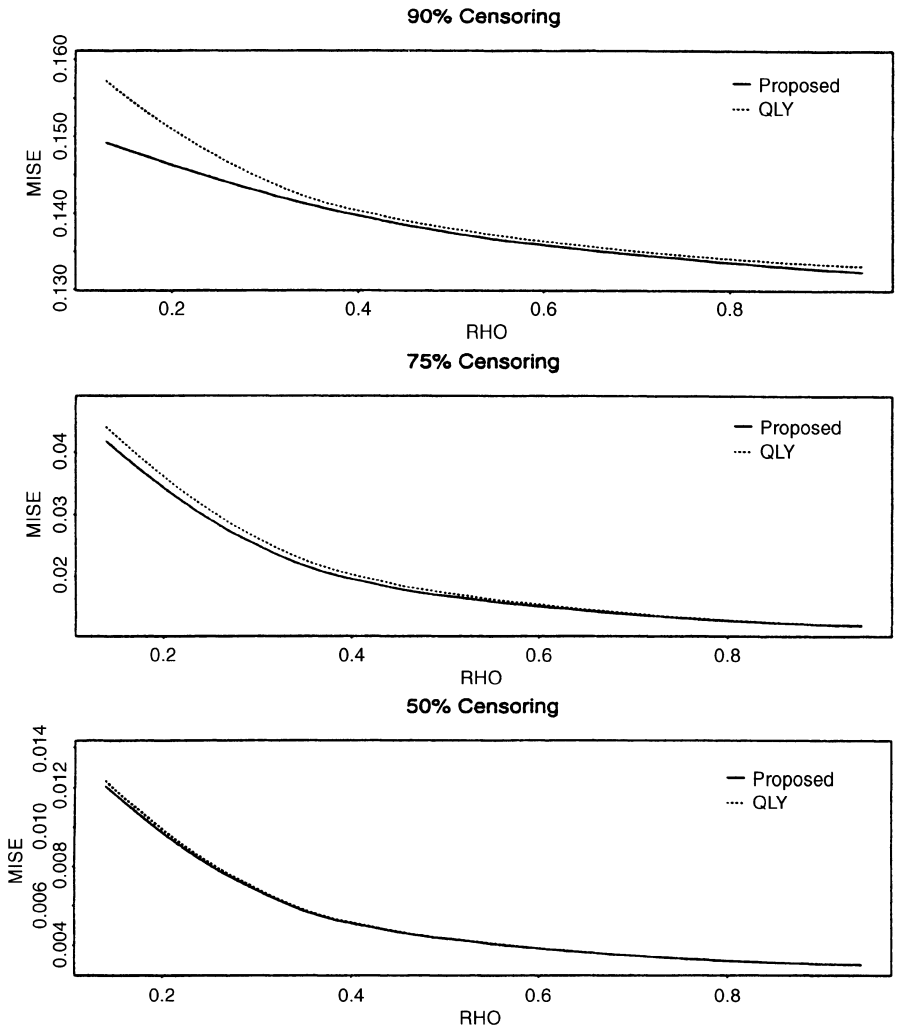

Fig. 1. The MISE of the proposed estimator $\hat{S}$ (solid line) and of $\hat{S}_{\mathrm{GLY}}$ (dotted line) are plotted against the probability of non-missingness

doing this is through a Cox proportional hazards model analysis, and in this section we consider an extension of our approach to deal with missing censoring indicators in this model.

The observations now consist of n i.i.d. replicates $(X_i, \xi_i, \sigma_i, Z_i)$ of the generic (X, ξ, σ, Z), where Z is a $p \times 1$ vector of bounded covariates and the conditional hazard function of T given Z is specified by the Cox model, as defined in the introduction. For simplicity, we have assumed that the covariates are non-time-dependent, but it is straightforward to extend our results to the time-dependent case. We begin with a review of existing estimators of the regression parameter β, the true value of which is denoted β_0.

3.1. Review of existing estimators

When none of the censoring indicators are missing, the standard estimator of β is the maximum partial likelihood estimator $\hat{\beta}_P$ which solves the estimating equation $U(\beta) = 0$, where

$$U(\beta) = \sum_{i=1}^{n} \int_0^{\infty} (Z_i - \bar{Z}(\beta, t))\, dN_i^u(t).$$

Here $N_i^u(t) = \delta_i N_i(t)$ with $N_i(t) = 1_{\{X_i \leqslant t\}}$,

$$\bar{Z}(\beta, t) = \frac{\sum_{i=1}^{n} Y_i(t) \exp(\beta' Z_i) Z_i}{\sum_{i=1}^{n} Y_i(t) \exp(\beta' Z_i)},$$

and $Y_i(t) = 1_{\{X_i \geqslant t\}}$ denotes the at-risk indicator for the ith individual. Andersen & Gill (1982) showed that

$$\sqrt{n}(\hat{\beta}_P - \beta_0) \xrightarrow{D} N(0, \Sigma^{-1}),$$

where $\Sigma = -\lim_{n\to\infty} n^{-1} \partial U(\beta_0)/\partial\beta$ is assumed to be positive definite.

In the case of missing censoring indicators, Gijbels *et al.* (1993) proposed the estimating equation

$$U_{11}(\beta, \infty) + DU_*(\beta, \infty) = 0,$$

where D is a certain $p \times p$ matrix,

$$U_{11}(\beta, t) = \sum_{i=1}^{n} \int_0^{t} (Z_i - \bar{Z}(\beta, s)) \xi_i \, dN_i^u(s),$$

$$U_*(\beta, t) = \sum_{i=1}^{n} \int_0^{t} (Z_i - \bar{Z}(\beta, s)) \left((1 - \xi_i)\, dN_i(s) - \frac{1 - \hat{\rho}}{\hat{\rho}} \xi_i \, dN_i^c(s) \right),$$

and $N_i^c(t) = (1 - \delta_i) N_i(t)$. They show that solutions of this estimating equation are asymptotically normal, and they find the asymptotically optimal choice of D.

3.2. Proposed estimating function

The idea is to exploit the information in the $0 \to 00$ transitions in a way that is similar to our approach in the absence of covariates. Define the estimating function corresponding to the $0 \to 00$ transitions by

$$U_{00}(\beta, t) = \sum_{i=1}^{n} \int_0^{t} (Z_i - \bar{Z}(\beta, s))(1 - \xi_i)\, dN_i(s), \tag{3.5}$$

and the estimating function U_{10} corresponding to the $0 \to 10$ transitions by integrating with respect to $\xi_i \, dN_i^c$ instead of $(1 - \xi_i)\, dN_i$. We propose the estimating function

$$U(\beta, t) = U_{11}(\beta, t) + \hat{\pi}_P(\beta, t) U_{00}(\beta, t),$$

where

$$\hat{\pi}_P(\beta, t) = P(\beta, t)(P(\beta, t) + Q(\beta, t))^{-1},$$

$P(\beta, t) = \mathrm{diag}(U_{11}(\beta, t))$, $Q(\beta, t) = \mathrm{diag}(U_{10}(\beta, t))$ and $\mathrm{diag}(v)$ is the diagonal matrix with diagonal v.

Our estimator $\hat{\beta}$ is a solution to the estimating equation $U(\beta, \infty) = 0$. It can be shown that within any compact set containing β_0, a unique solution to this estimating equation exists with probability tending to 1 as the sample size increases, see the proof of consistency part of theorem 2. To show that $\hat{\beta}$ is asymptotically normal we need the following notation:

$$W_i(t) = Z_i - \bar{z}(\beta_0, t),$$

$$N_i^{CZ}(t) = \int_0^t W_i(s)\, dN_i^c(s),$$

$$\bar{z}(\beta, t) = \frac{E(Y_1(t)\exp(\beta' Z_1)Z_1)}{E(Y_1(t)\exp(\beta' Z_1))},$$

$$N_i^Z(t) = \int_0^t W_i(s)\, dN_i(s),$$

$$B(t) = \operatorname{diag}\{E(N_{1j}^{CZ}(t))/E(N_{1j}^Z(t)),\ j = 1, \ldots, p\}.$$

Also, let $v^{\otimes 2}$ denote the cross product vv' for any vector v.

Theorem 2
The estimator $\hat{\beta}$ is consistent and $\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{D} N(0, V)$, where

$$V = \Sigma^{-1} + (\rho^{-1} - 1)\Sigma^{-1} E(N_1^{CZ}(\infty) - B(\infty)N_1^Z(\infty))^{\otimes 2}\Sigma^{-1}.$$

The first term in V is the asymptotic variance of Cox's maximum partial likelihood estimator based on the full data, and the second term represents the effect of the missing censoring indicators.

A consistent estimator of V can be obtained as follows. Note that if we suppose that the processes N_i^u and N_i^c are observed, each term in V can be consistently estimated by plugging $\hat{\beta}$ and $\hat{\rho}$ into the corresponding empirical estimator in place of the unknown β_0 and ρ. Replacing those (unobserved) processes by $\hat{\rho}^{-1}\xi_i N_i^u$ and $\hat{\rho}^{-1}\xi_i N_i^c$, respectively, results in a consistent estimator of V.

3.3. Estimation of the conditional survival function

In this section we briefly discuss estimation of the conditional survival function $S(t|z) = P(T > t|Z = z)$ under the Cox model with missing censoring indicators. This will be done by plugging $\hat{\beta}$ and an estimate of the cumulative baseline hazard function $\Lambda_0(t) = \int_0^t \lambda_0(s)\, ds$ into the Cox model based expression for $S(t|z)$.

The baseline hazard Λ_0 can be expressed in the same form as the basic equation (1.1) for Λ_T, except that the cumulative transition intensities Λ_{kl} are now replaced by the functions

$$\Lambda_{kl}^0(t) = \int_0^t \frac{dH_{kl}(s)}{E(Y_1(s)\exp(\beta_0' Z_1))}.$$

This leads to the following consistent estimator of the baseline hazard function:

$$\hat{\Lambda}_0(t) = \hat{\Lambda}_{11}^0(t) + \hat{\pi}^0(t)\hat{\Lambda}_{00}^0(t),$$

where

$$\hat{\pi}^0(t) \equiv \frac{\hat{\Lambda}^0_{11}(t)}{\hat{\Lambda}^0_{11}(t) + \hat{\Lambda}^0_{10}(t)}$$

is defined to be zero when the denominator vanishes, and $\hat{\Lambda}^0_{kl}$ is the empirical version of Λ^0_{kl}. The estimator $\hat{\Lambda}_0$ reduces to Breslow's estimator when there are no missing censoring indicators.

The resulting estimator of the conditional survival function is the product integral

$$\hat{S}(t|z) = \Pi_{(0,t]}(1 - \exp(\hat{\beta}'z)\,d\hat{\Lambda}_0(s)),$$

cf. Andersen *et al.* (1993, p. 509). It is straightforward to show that $\hat{S}(t|z)$ consistently estimates $S(t|z)$. It is also possible to prove asymptotic normality of $\hat{S}(t|z)$ using the techniques of theorems 1 and 2, but the asymptotic variance is complicated.

3.4. Numerical results

We carried out a simulation study to compare the performance of the proposed estimator with that of the Gijbels *et al.* and complete case estimators of β.

The underlying Cox model was taken to be $\lambda(t|z) = \exp(\beta_0 z)$, for $\beta_0 = 0.$, 0.5, 1.0, and the covariate Z was standard normal. The censoring was exponential with the parameter adjusted to give prescribed censoring rates of 30% and 80%. In each case the mean square errors (MSE) of the various estimators were computed from 10,000 simulated samples of size $n = 100$, see Tables 1 and 2. The "full data" estimator is included for comparison. The results given in Tables 1 and 2 are classified according to the value of ρ (0.8 or 0.5).

The proposed estimator improves upon the Gijbels *et al.* (1993) estimator in terms of MSE when the censoring is heavy (80%) and there is a low proportion of missing censoring indicators

Table 1. $Z \sim N(0, 1)$, $\rho = 0.8$

Censoring rate	β_0	Full data Mean	Full data MSE	Proposed Mean	Proposed MSE	Gijbels *et al.* Mean	Gijbels *et al.* MSE	Complete case Mean	Complete case MSE
	0.0	0.003	0.060	0.003	0.067	0.005	0.073	0.005	0.078
80%	0.5	0.514	0.071	0.514	0.083	0.514	0.087	0.516	0.093
	1.0	1.014	0.101	1.017	0.118	1.016	0.125	1.018	0.134
	0.0	−0.003	0.016	−0.002	0.023	−0.003	0.023	−0.002	0.037
30%	0.5	0.506	0.019	0.507	0.021	0.506	0.021	0.507	0.025
	1.0	1.011	0.027	1.102	0.031	1.102	0.030	1.014	0.035

Table 2. $Z \sim N(0, 1)$, $\rho = 0.5$

Censoring rate	β_0	Full data Mean	Full data MSE	Proposed Mean	Proposed MSE	Gijbels *et al.* Mean	Gijbels *et al.* MSE	Complete case Mean	Complete case MSE
	0.0	0.003	0.060	0.004	0.106	0.002	0.124	0.004	0.161
80%	0.5	0.510	0.073	0.513	0.153	0.512	0.150	0.521	0.190
	1.0	1.025	0.101	1.046	0.202	1.042	0.203	1.057	0.254
	0.0	−0.003	0.016	−0.002	0.023	−0.003	0.021	−0.002	0.037
30%	0.5	0.508	0.020	0.512	0.029	0.512	0.026	0.519	0.044
	1.0	1.012	0.0277	1.014	0.041	1.016	0.036	1.024	0.060

($\rho = 0.8$). The degree of the improvement increases as the covariate effect (β_0) diminishes. Although $\hat{\beta}$ can do slightly worse than the Gijbels *et al.* estimator (under light censoring and $\rho = 0.5$; see the last three rows of Table 2), it has the best overall performance. In terms of bias, the two estimators have very similar performance. Note that the censoring rate and ρ can be estimated under the MCAR assumption, and this information can be used to choose between the proposed method and the Gijbels *et al.* method.

3.5. *Further research*

One of the referees asked whether the MCAR assumption could be relaxed to allow the missingness to depend on the covariate. Our approach does not extend without modification to this case; in particular, the above simulation example shows that $\hat{\beta}$ can be quite biased when ρ depends on Z. Such an extension would be non-trivial because the "curse-of-dimensionality" implies that $\rho(Z)$ is difficult to estimate for high-dimensional covariates unless it is assumed to be sufficiently smooth, as in logistic regression model on Z (cf. Robins & Ritov, 1997). Such an extension is of considerable interest in applications, however, and would be a worthwhile direction for further research.

Acknowledgements

The authors thank the associate editor and the referees for their helpful comments.

References

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

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

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

Dinse, G. E. (1982). Nonparametric estimation for partially-complete time and type of failure data. *Biometrics* **38**, 417–431.

Efron, B. (1967). The two sample problem with censored data. *Proc. 5th Berkeley Symp. Math. Statist. Probab.* **4**, 831–853.

Gijbels, I., Lin, D. Y. & Ying, Z. (1993). Non- and semi-parametric analysis of failure time data with missing failure indicators. Tech. Report 039-93, Mathematical Sciences Research Institute, Berkeley.

Gill, R. D. (1989). Non- and semi- parametric maximum likelihood estimators and the von Mises method (part 1). *Scand. J. Statist.* **16**, 97–124.

Gill, R. D. & Johansen, S. (1990). A survey of product-integration with a view toward application in survival analysis. *Ann. Statist.* **18**, 1501–1555.

Gill, R. D., van der Laan, M. J. & Robins, J. M. (1997). Coarsening at random. In *1st Seattle Symp. Biostatist.* (ed. D.-Y. Lin). Springer, New York.

Goetghebeur, E. J. & Ryan, L. (1995). Analysis of competing risks survival data when some failure types are missing. *Biometrika* **82**, 821–833.

Heitjan D. F. & Rubin D. B. (1991). Ignorability and coarse data. *Ann. Statist.* **19**, 2244–2253.

Jacobsen, M. & Keiding, N. (1995). Coarsening at random in general sample spaces and random censoring in continuous time. *Ann. Statist.* **23**, 774–786.

van der Laan, M. J. & McKeague, I. W. (1998). Efficient estimation from right-censored data when failure indicators are missing at random. *Ann. Statist.* **26**, No.1.

Little, R. J. A. & Rubin, D. B. (1987). *Statistical analysis with missing data*. Wiley, New York.

Lo, S.-H. (1991). Estimating a survival function with incomplete cause-of-death data. *J. Multivar. Anal.* **39**, 217–235.

Robins, J. M. & Ritov, Y. (1997). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statist. Med.* **16**, 285–319.
Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–590.

Received September 1996, in final form October 1997

Sundarraman Subramanian, Department of Mathematics and Statistics, 5752 Neville Hall, University of Maine, Orono, ME 04469-5752, USA.

Appendix

Proof of theorem 1. The product-limit functional $\phi(\Lambda_T) = \Pi(1 - d\Lambda_T)$ is compactly differentiable with derivative $[d\phi(\Lambda_T)(l)](t) = -S(t)l(t)$, for $l \in D[0, \tau]$, see Gill & Johansen (1990). Here we have used the continuity of S. We will show below that $\sqrt{n}(\hat{\Lambda}_T - \Lambda_T)$ converges in distribution to the Gaussian process L in theorem 1, so the functional delta-method implies that $\sqrt{n}(\hat{S} - S)$ converges in distribution to $-S \cdot L \overset{\mathscr{D}}{=} S \cdot L$. This will complete the proof.

Using (1.1) and the definition of $\hat{\Lambda}_T$ we may write

$$\sqrt{n}(\hat{\Lambda}_T(t) - \Lambda_T(t)) = \left(1 + \frac{\Lambda_{00}(t)D_{10}(t)}{\hat{\Lambda}_{11}(t) + \hat{\Lambda}_{10}(t)}\right)\sqrt{n}(\hat{\Lambda}_{11}(t) - \Lambda_{11}(t))$$

$$+ \hat{D}_{11}(t)\sqrt{n}(\hat{\Lambda}_{00}(t) - \Lambda_{00}(t)) - \frac{\Lambda_{00}(t)D_{11}(t)}{\hat{\Lambda}_{11}(t) + \hat{\Lambda}_{10}(t)}\sqrt{n}(\hat{\Lambda}_{10}(t) - \Lambda_{10}(t)), \tag{3.6}$$

where $\hat{D}_{11}$ is the empirical version of D_{11}. Now we apply the standard weak convergence result for Nelson–Aalen estimators of cumulative transition functions: $(\sqrt{n}(\hat{\Lambda}_{kl} - \Lambda_{kl}(t));\ (k, l) \in \Gamma) \xrightarrow{\mathscr{D}} (W_{kl};\ (k, l) \in \Gamma)$ on $[0, \tau]$, where the W_{kl} are independent zero-mean continuous Gaussian martingales with covariance function $\operatorname{cov}(W_{kl}(u), W_{kl}(v)) = C_{kl}(u \wedge v)$, see Andersen *et al.* (1993, p. 198). The assumed absolute continuity of F and G allows us to use the latter result, which requires the existence of transition intensities. The filtration used in this connection is the natural filtration of the four-state Markov process that underlies the missing censoring indicator model. It follows immediately from (3.6) that $\sqrt{n}(\hat{\Lambda}_T - \Lambda_T)$ converges weakly on $[0, \tau]$ to the Gaussian process

$$L(t) = D_{11}(t)W_{00}(t) - D_{00}(t)D_{11}(t)W_{10}(t) + (1 + D_{00}(t)D_{10}(t))W_{11}(t)$$

which has the covariance function given in theorem 1. □

Proof of theorem 2. This proof follows the basic approach of Andersen & Gill (1982) in deriving the asymptotic distribution of the maximum partial likelihood estimator $\hat{\beta}_P$, and it also uses some of the steps taken by Gijbels *et al.* (1993).

Taylor expanding $U(\beta, \infty)$ about β_0 and using $U(\hat{\beta}, \infty) = 0$, it is seen that $n^{-1/2}U(\beta_0, \infty) = n^{-1}J(\beta^*, \infty)n^{1/2}(\hat{\beta} - \beta)$, where β^* is on the line segment between $\hat{\beta}$ and β_0, and $J(\beta, \infty)$ is the partial derivative of $U(\beta, \infty)$ with respect to β. Thus, once we have shown consistency of $\hat{\beta}$, asymptotic normality will follow if we show that $n^{-1/2}U(\beta_0, \cdot)$ converges weakly to a Gaussian process and $n^{-1}J(\beta^*, \infty)$ converges in probability to the non-singular matrix Σ for any $\beta^* \xrightarrow{P} \beta_0$. We only consider the case of a one-dimensional covariate, as the general case is similar.

The following expression for $U(\beta, t)$ will be useful: $U(\beta, t) = U_{11}(\beta, t) + U_{11}^D(\beta, t)$, where $U_{11}^D(\beta, t) = U_{00}(\beta, t) - \tilde{P}(\beta, t)U_{10}(\beta, t)$, and $\tilde{P}(\beta, t) = \operatorname{diag}(U_{00}(\beta, t))((P(\beta, t) + Q(\beta, t))^{-1}$.

We begin by establishing the weak convergence of $n^{-1/2}U(\beta_0, \cdot)$. To simplify the notation, when an expression is evaluated at $\beta = \beta_0$ we shall suppress its dependence on β_0. First recall that the counting process $N_i^u(t)$ has intensity $\lambda_i(t) = \lambda(t|Z_i)Y_i(t)$ with respect to the filtration generated by the processes N_j^u, Y_j and Z_j, $j = 1, \ldots, n$. The corresponding martingale is $M_i(t) = N_i^u(t) - \int_0^t \lambda_i(s)\,ds$.

Note that $n^{-1/2}U_{11}(t)$ is asymptotically equivalent to $n^{-1/2}\tilde{U}_{11}(t)$, where

$$\tilde{U}_{11}(t) = \rho \sum_{i=1}^n \int_0^t W_i(s)\,dM_i(s) + \sum_{i=1}^n (\xi_i - \rho) \int_0^t W_i(s)\,dN_i^u(s).$$

Next we obtain a suitable representation for $U_{11}^D(t)$. Note that

$$\sum_{i=1}^n \int_0^t (Z_i - \bar{Z}(s))\lambda_i(s)\,ds = 0. \tag{3.7}$$

Using $N_i = N_i^u + N_i^c$ we may write

$$U_{00}(t) = \sum_{i=1}^n \int_0^t (Z_i - \bar{Z}(s))(1 - \xi_i)\,dM_i(s) + \sum_{i=1}^n \int_0^t (Z_i - \bar{Z}(s))(1 - \xi_i)\lambda_i(s)\,ds$$
$$+ \sum_{i=1}^n \int_0^t (Z_i - \bar{Z}(s))(1 - \xi_i)\,dN_i^c(s).$$

Then, since $\tilde{P}(t) = 1/\hat{\rho}_1(t) - 1$ where

$$\hat{\rho}_1(t) = \frac{\sum_{i=1}^n \int_0^t (Z_i - \bar{Z}(s))\xi_i\,dN_i(s)}{\sum_{i=1}^n \int_0^t (Z_i - \bar{Z}(s))\,dN_i(s)},$$

we have

$$\tilde{P}(t)U_{10}(t) = \sum_{i=1}^n \int_0^t (Z_i - \bar{Z}(s))\left(\frac{1}{\hat{\rho}_1(t)} - 1\right)\xi_i\,dN_i^c(s).$$

Using (3.7), this gives $U_{11}^D(t) = A_1(t) + A_2(t)$ where

$$A_1(t) = \sum_{i=1}^n \int_0^t (Z_i - \bar{Z}(s))(1 - \xi_i)\,dM_i(s) + \sum_{i=1}^n \int_0^t (Z_i - \bar{Z}(s))(1 - \xi_i - (1 - \rho))\lambda_i(s)\,ds,$$

and

$$A_2(t) = \sum_{i=1}^n \int_0^t (Z_i - \bar{Z}(s))\left(1 - \frac{\xi_i}{\hat{\rho}_1(t)}\right)dN_i^c(s).$$

Gijbels *et al.* (1993) (proof of th. 2) show that $n^{-1/2}A_1(t)$ is asymptotically equivalent to $n^{-1/2}\tilde{A}_1(t)$, where

$$\tilde{A}_1(t) = (1 - \rho)\sum_{i=1}^n \int_0^t W_i(s)\,dM_i(s) - \sum_{i=1}^n (\xi_i - \rho)\int_0^t W_i(s)\,dN_i^u(s).$$

Next consider $A_2(t)$. Using cond. B and D of Andersen & Gill (1982) it can be shown that $\hat{\rho}_1(t) - \rho = O_p(n^{-1/2})$. In addition, applying the delta method gives

$$A_2(t) = -\sum_{i=1}^{n} \rho^{-1}(\xi_i - \rho) \int_0^t (Z_i - \bar{Z}(s))\, dN_i^c(s)$$
$$+ \left(\sum_{i=1}^{n} \rho^{-1}(\xi_i - \rho) \int_0^t (Z_i - \bar{Z}(s)) dN_i(s) \right) \frac{E\big(N_i^{CZ}(t)\big)}{E(N_i^Z(t))} + r_n(t),$$

where $\sup_t |r_n(t)| = o_p(n^{1/2})$. Using lem. 1 of Gijbels *et al.* (1993) it can then be shown that $n^{-1/2}A_2(t)$ is asymptotically equivalent to $n^{-1/2}\tilde{A}_2(t)$, where

$$\tilde{A}_2(t) = -\sum_{i=1}^{n} \rho^{-1}(\xi_i - \rho)\big(N_i^{CZ}(t) - B(t)N_i^Z(t)\big).$$

We have shown that $n^{-1/2}U(t)$ is asymptotically equivalent to $n^{-1/2}\tilde{U}(t)$, where $\tilde{U}(t) = \tilde{U}_{11}(t) + \tilde{A}_1(t) + \tilde{A}_2(t)$, which is a sum of n i.i.d. processes. Convergence of the finite dimensional distributions of $n^{-1/2}U(\cdot)$ follows. Tightness can be established as in Gijbels *et al.* (1993) using their lem. 1.

We next prove consistency. Using the independence assumption of the MCAR model, it is easily shown (cf. Andersen & Gill (1982)) that $n^{-1}U_{11}(\beta, \infty)$ converges uniformly in any compact set to $\rho m(\beta)$ almost surely, where $m(\beta) = E\int_0^\infty (Z_1 - \bar{z}(\beta, t))\, dN_1^u(t)$. Similarly, using the decomposition

$$\begin{aligned} U_{11}^D(\beta, \infty) &= \sum_{i=1}^{n} \int_0^\infty (Z_i - \bar{Z}(\beta, s))(1 - \xi_i)\, dN_i^u(s) \\ &+ \sum_{i=1}^{n} \int_0^\infty (Z_i - \bar{Z}(\beta, s))(1 - \xi_i)\, dN_i^c(s) \\ &- \frac{\sum_{i=1}^{n} \int_0^\infty (Z_i - \bar{Z}(\beta, s))(1 - \xi_i)\, dN_i(s)}{\sum_{i=1}^{n} \int_0^\infty (Z_i - \bar{Z}(\beta, s))\xi_i\, dN_i(s)} \sum_{i=1}^{n} \int_0^\infty (Z_i - \bar{Z}(\beta, s))\xi_i\, dN_i^c(s), \end{aligned} \tag{3.8}$$

it can be shown that $n^{-1}U_{11}^D(\beta, \infty)$ converges uniformly in any compact set to $(1 - \rho)m(\beta)$, since the contribution from the sum of the last two terms in the decomposition is asymptotically negligible. Thus $n^{-1}U(\beta, \infty)$ is uniformly approximated by the continuous function $m(\beta)$, which has unique root β_0, and we conclude that $\hat{\beta}$ is consistent.

It remains to show that $n^{-1}J(\beta^*, \infty)$ converges in probability to Σ. Along the lines of Andersen & Gill (1982, p. 1108), but also using the independence assumption of the MCAR model, we have that $n^{-1}\partial U_{11}(\beta^*, \infty)/\partial\beta$ converges in probability to $\rho\Sigma$. Similarly, using the decomposition (3.8), in which only the first term makes a contribution asymptotically, it can be shown that $n^{-1}\partial U_{11}^D(\beta^*, \infty)/\partial\beta$ converges in probability to $(1 - \rho)\Sigma$. This completes the proof.