Inference Based on Imputed Failure Times for the Proportional Hazards Model With Interval-Censored Data

Glen A. SATTEN, Somnath DATTA, and John M. WILLIAMSON

We propose an approach to the proportional hazards model for interval-censored data in which parameter estimates are obtained by solving estimating equations that are the partial likelihood score equations for the full-data proportional hazards model, averaged over all rankings of imputed failure times consistent with the observed censoring intervals. Imputed failure times are generated using a parametric estimate of the baseline distribution; the parameters of the baseline distribution are estimated simultaneously with the proportional hazards regression parameters. Although a parametric form for the baseline distribution must be specified, simulation studies show that the method performs well even when the baseline distribution is misspecified. The estimating equations are solved using Monte Carlo techniques. We present a recursive stochastic approximation scheme that converges to the zero of the estimating equations; the solution has a random error that is asymptotically normally distributed with a variance–covariance matrix that can itself be estimated recursively.

KEY WORDS: Cox model; Current-status data; Monte Carlo method; Robbins–Monro process; Stochastic approximation; Survival analysis.

1. INTRODUCTION

The proportional hazards, or Cox, model is commonly used for assessing the effects of covariates on survival time (Cox 1972); inference can be based on the partial likelihood when a distinct failure or censoring time is observed for each individual or experimental unit. However, in many instances, one may know only a time interval in which the failure occurred, in which case we say that the data are interval censored.

Interval censoring arises naturally whenever individuals or experimental units are observed only occasionally and where the failure event of interest does not preclude continued follow-up. Examples of interval-censored data abound in research on human immunodeficiency virus (HIV) and acquired immunodeficiency syndrome (AIDS), because important events such as infection, seroconversion (first appearance of detectable antibodies), and extent of disease progression (as measured by the steadily decreasing concentration of CD4 cell counts) are ascertainable only by laboratory analysis and do not produce uniquely identifiable clinical symptoms. A concrete example is given by a follow-up study of HIV-negative persons who are at high risk for becoming HIV infected; such studies are being conducted in preparation for testing HIV vaccines (Hoff 1994). Here we are interested in cofactors that affect the risk of becoming infected with HIV. Individuals are tested periodically for HIV; however, the time of seroconversion is known only to be between the last negative test and the first positive test. Other concrete examples have been provided by Finkelstein (1986) and Kooperberg and Clarkson (1997). A special case of interval-censored data is current-status data, where individuals are each seen only once after enrollment. Current-status data arises naturally in cross-sectional surveys, where the goal is calculation of the distribution of age of onset for a disease or life event. Grummer-Strawn (1993) analyzed cofactors affecting duration of breast-feeding using current-status data from a demographic survey. Grouped data (or tied data) are another special case of interval-censored data.

Several approaches are currently available for fitting the proportional hazards model to interval-censored or grouped data. Finkelstein (1986) considered a parametric method in which the baseline distribution is fit simultaneously along with regression coefficients, by maximizing the full likelihood of the observed data. Diamond, McDonald, and Shah (1986) proposed a similar model for current-status data. Huang (1996) considered simultaneous estimation of regression parameters and a nonparametric maximum likelihood estimator of the baseline survival distribution for current-status data. Many authors have considered use of the marginal likelihood of all possible rankings for grouped data (DeLong, Guirguis, and So 1994; Kalbfleisch and Prentice 1972, 1973; Peto 1972; Sinha, Tanner, and Hall 1994). Satten (1996) developed a Monte Carlo method for maximizing the marginal likelihood for the general interval censored case, allowing estimation of regression coefficients without specifying the baseline distribution.

In this article we consider a method that is intermediate between Finkelstein's parametric method and Satten's rank-based marginal likelihood approach. A parametric model for the baseline hazard is assumed, but only as a method for generating imputed failure times; then a rank-based procedure using the imputed failure times is used to estimate the regression coefficients. This proposed approach is easier than the marginal likelihood approach to generalize to more complicated settings where interval-censored data are encountered and allows for a simpler treatment of observations that are right censored (see, e.g., Kalbfleisch and

Glen A. Satten and John M. Williamson are Mathematical Statisticians, Division of HIV/AIDS Prevention: Surveillance and Epidemiology, MS E-48, National Center for HIV, STD and TB Prevention, Centers for Disease Control and Prevention, Atlanta, GA 30333. Somnath Datta is Associate Professor, Department of Statistics, University of Georgia, Athens, GA 30602. This work was supported by National Institutes of Health grant T32-AI07442 (Williamson) and partially supported by National Security Agency grant MDA904-96-1-0049 (Datta).

Journal of the American Statistical Association
March 1998, Vol. 93, No. 441, Theory and Methods

Prentice 1980). In addition, the proposed approach is easier to implement than Satten's method for the maximizing the marginal likelihood. Like the marginal likelihood approach, the new method reduces to the usual Cox model when the failure times are known or when censoring intervals are nonoverlapping. The disadvantage of the proposed method is its dependence on specification of a parametric model for the baseline hazard, although this dependence appears to be slight in practice.

In Section 2 we present our proposed approach, and in Section 3 we outline our theoretical results on consistency and asymptotic normality. In Section 4 we show how parameter estimates may be obtained using Monte Carlo methods. We provide simulation results in Section 5 to demonstrate the effectiveness of the proposed method for differing censoring schemes. We conclude with a short discussion in Section 6.

2. MODEL

If data are interval censored, then for each individual, instead of a failure time, we observe a censoring interval $[l_i, u_i)$ that is known to contain the actual failure time. A rank-based approach to inference on regression parameters $\boldsymbol{\beta}$ in the proportional hazards model that does not require knowledge of the baseline distribution F_0 is available (Satten 1996). In this methodology the data are considered to be the set of all possible rankings $\mathcal{R}$ of the unobserved failure times that are consistent with the observed censoring intervals. Maximizing the marginal likelihood of $\mathcal{R}$ leads to a score function that is the average of the score function for the full-data proportional hazards model, with respect to a probability distribution on the set $\mathcal{R}$ that is independent of the baseline distribution. The score equations are solved by a Monte Carlo procedure that requires a Gibbs sampler to generate rankings from the set $\mathcal{R}$. Because the set $\mathcal{R}$ is considered to be the data in this methodology, we call this methodology the missing-rank approach.

We propose here a simpler procedure that is motivated by the rank-based method just described. We consider an estimating function that is also the expected value of the score function for the full-data proportional hazards model, but averaged with respect to the distribution of the rank order of the imputed failure times. We refer to this methodology as the missing-failure time approach.

Let $\mathbf{t} = (t_1, t_2, \ldots, t_n)$ be the vector of (possibly right-censored) failure times; let $\boldsymbol{\delta}$ be a vector of right-censoring indicators with components $\delta_i = 0$ if the ith individual is right censored and $\delta_i = 1$ otherwise; and let $\mathbf{l}$ and $\mathbf{u}$ denote vectors with ith component l_i and u_i. Let $\mathbf{x}_i$ be a vector of covariates for the ith observation, and let $\mathbf{x}$ denote the matrix whose ith row is $\mathbf{x}_i$. Let $\mathbf{d}_i$ denote the vector $(\mathbf{x}_i, l_i, u_i, \delta_i)$, and let $\mathbf{d}$ denote the matrix whose ith row is $\mathbf{d}_i$. We assume that the observed data comprise iid samples of $\mathbf{d}_i$. Let $F(\mathbf{t}|\mathbf{d}; \boldsymbol{\beta}, \boldsymbol{\theta})$ be a parametric family of conditional distributions of failure times $\mathbf{t}$ given the observed censoring intervals, covariates, and right-censoring indicator $\boldsymbol{\delta}$, and assume that this family of distributions contains the true distribution of $\mathbf{t}$ and that the distribution of t_i conditional on $\mathbf{x}_i$ is in the proportional hazards family. For this family of distributions, the conditional expectation of $\mathbf{S}_\beta(\mathbf{t}|\boldsymbol{\delta}, \mathbf{x}; \boldsymbol{\beta})$, the partial likelihood score function for the full-data proportional hazards model, given the observed censoring intervals, is

$$\begin{aligned}\mathbb{S}_\beta(\boldsymbol{\beta}, \boldsymbol{\theta}) &= E_{F(\cdot|\mathbf{d};\boldsymbol{\theta})}[\mathbf{S}(\mathbf{t}|\boldsymbol{\delta}, \mathbf{x}; \boldsymbol{\beta})] \\ &= \int \mathbf{S}_\beta(\mathbf{t}|\boldsymbol{\delta}, \mathbf{x}; \boldsymbol{\beta})\, dF(\mathbf{t}|\mathbf{d}; \boldsymbol{\beta}, \boldsymbol{\theta}). \qquad (1)\end{aligned}$$

We propose estimating $\boldsymbol{\beta}$ by the solution to

$$\mathbb{S}_\beta(\boldsymbol{\beta}, \hat{\boldsymbol{\theta}}(\boldsymbol{\beta})) = \mathbf{0}, \qquad (2)$$

where $\hat{\boldsymbol{\theta}}(\boldsymbol{\beta})$ is an estimate of the parameters $\boldsymbol{\theta}$, as a function of $\boldsymbol{\beta}$, required to specify F.

Why do we propose using (2) instead of the full-likelihood score equation when we have already made parametric assumptions on the failure time distribution? The reason is that because the role of F in (1) is only to assign a relative weighting to the possible rank orderings of imputed failure times, we may expect that (1) is less sensitive to misspecification of the form of the failure time distribution than a fully parametric estimator of the regression parameters $\boldsymbol{\beta}$. This intuition is verified in our simulation studies presented in Section 4. The arguments that we present later require that the parametric form of F be correctly specified; however, the variance–covariance matrix that we propose is a robust (sandwich) estimator, which may further protect against misspecification of F.

The distribution F has the product form

$$F(\mathbf{t}|\mathbf{d}; \boldsymbol{\beta}, \boldsymbol{\theta}) = \prod_{i=1}^{n} F(t_i|l_i, u_i, \mathbf{x}_i; \boldsymbol{\beta}, \boldsymbol{\theta})^{\delta_i} I(t_i \geq l_i)^{1-\delta_i}, \quad (3)$$

where $I(\mathcal{C}) = 1$ if $\mathcal{C}$ is true and 0 otherwise. We assume that the censoring is independent of failure times, so that

$$F(t_i|l_i, u_i, \mathbf{x}_i; \boldsymbol{\beta}, \boldsymbol{\theta}) = \frac{F(t_i|\mathbf{x}_i; \boldsymbol{\beta}, \boldsymbol{\theta}) - F(l_i|\mathbf{x}_i; \boldsymbol{\beta}, \boldsymbol{\theta})}{F(u_i|\mathbf{x}_i; \boldsymbol{\beta}, \boldsymbol{\theta}) - F(l_i|\mathbf{x}_i; \boldsymbol{\beta}, \boldsymbol{\theta})} I(l_i \leq t_i \leq u_i).$$

As $F(t_i|\mathbf{x}_i; \boldsymbol{\beta}, \boldsymbol{\theta})$ is in the proportional hazards family, we have

$$F(t_i|\mathbf{x}_i; \boldsymbol{\beta}, \boldsymbol{\theta}) = 1 - [1 - F_0(t_i|\boldsymbol{\theta})]^{e^{\boldsymbol{\beta}\cdot\mathbf{x}_i}},$$

where $F_0(t|\boldsymbol{\theta})$ is the cumulative distribution function (cdf) of failure times in the baseline group; we refer to F_0 as the baseline distribution. We denote the hazard function for the distribution F_0 by $\lambda_0(t)$.

We propose estimating $\boldsymbol{\theta}$ by maximizing the observed log-likelihood $\mathcal{L}^0$, given by

$$\begin{aligned}\mathcal{L}^0 = \sum_{i=1}^{n} \delta_i \ln[F(u_i|\mathbf{x}_i; \boldsymbol{\beta}, \boldsymbol{\theta}) - F(l_i|\mathbf{x}_i; \boldsymbol{\beta}, \boldsymbol{\theta})] \\ + (1 - \delta_i) \ln[1 - F(l_i|\mathbf{x}_i; \boldsymbol{\beta}, \boldsymbol{\theta})], \quad (4)\end{aligned}$$

with respect to $\boldsymbol{\theta}$. Because we could use (5) to implicitly define $\hat{\boldsymbol{\beta}}$ as a function of $\hat{\boldsymbol{\theta}}$, estimating $\boldsymbol{\beta}$ using (2) and the value $\hat{\boldsymbol{\theta}}$ that maximizes (4) is equivalent to solving the estimating equations

$$\mathbb{S}_\beta(\boldsymbol{\beta}, \boldsymbol{\theta}) = \mathbf{0} \qquad (5)$$

and

$$\mathbb{U}^0_\theta(\boldsymbol{\beta},\boldsymbol{\theta}) \equiv \sum_{i=1}^n \mathcal{U}^0_\theta(\mathbf{d}_i;\boldsymbol{\beta},\boldsymbol{\theta}) = \mathbf{0} \tag{6}$$

simultaneously for $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$, where

$$\mathcal{U}^0_\theta(\mathbf{d}_i;\boldsymbol{\beta},\boldsymbol{\theta}) = \delta_i \frac{\partial \ln[F(u_i|\mathbf{x}_i;\boldsymbol{\beta},\boldsymbol{\theta}) - F(l_i|\mathbf{x}_i;\boldsymbol{\beta},\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}} + (1-\delta_i)\frac{\partial \ln[1 - F(l_i|\mathbf{x}_i;\boldsymbol{\beta},\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}}. \tag{7}$$

Note that our proposed estimator may be compared to that of Finkelstein (1986), who considered estimating $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ by solving the score equations (6) and (8), with

$$\mathbb{U}^0_\beta(\boldsymbol{\beta},\boldsymbol{\theta}) \equiv \sum_{i=1}^n \mathcal{U}^0_\beta(\mathbf{d}_i;\boldsymbol{\beta},\boldsymbol{\theta}) = \mathbf{0}, \tag{8}$$

where

$$\mathcal{U}^0_\beta(\mathbf{d}_i;\boldsymbol{\beta},\boldsymbol{\theta}) = \delta_i \frac{\partial \ln[F(u_i|\mathbf{x}_i;\boldsymbol{\beta},\boldsymbol{\theta}) - F(l_i|\mathbf{x}_i;\boldsymbol{\beta},\boldsymbol{\theta})]}{\partial \boldsymbol{\beta}} + (1-\delta_i)\frac{\partial \ln[1 - F(l_i|\mathbf{x}_i;\boldsymbol{\beta},\boldsymbol{\theta})]}{\partial \boldsymbol{\beta}}. \tag{9}$$

Our proposal can be directly compared to the fully parametric approach using the EM algorithm. By writing $\mathbb{U}^0_\beta(\boldsymbol{\beta},\boldsymbol{\theta})$ as

$$\mathbb{U}^0_\beta(\boldsymbol{\beta},\boldsymbol{\theta}) = \int \mathbb{U}^c_\beta(\mathbf{t}|\boldsymbol{\delta},\mathbf{x};\boldsymbol{\beta},\boldsymbol{\theta})\, dF(\mathbf{t}|\mathbf{d};\boldsymbol{\beta},\boldsymbol{\theta}),$$

where

$$\mathbb{U}^c_\beta(\mathbf{t}|\boldsymbol{\delta},\mathbf{x};\boldsymbol{\beta},\boldsymbol{\theta}) \equiv \sum_{i=1}^n \mathcal{U}^c_\beta(t_i,\delta_i,\mathbf{x}_i;\boldsymbol{\beta},\boldsymbol{\theta}) \tag{10}$$

and

$$\mathcal{U}^c_\beta(t_i,\delta_i,\mathbf{x}_i;\boldsymbol{\beta},\boldsymbol{\theta}) = \delta_i \frac{\partial \ln f(t_i|\mathbf{x}_i;\boldsymbol{\beta},\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} + (1-\delta_i)\frac{\partial \ln[1 - F(t_i|\mathbf{x}_i;\boldsymbol{\beta},\boldsymbol{\theta})]}{\partial \boldsymbol{\beta}}, \tag{11}$$

we see that $\mathbb{U}^c_\beta(\mathbf{t}|\boldsymbol{\delta},\mathbf{x};\boldsymbol{\beta},\boldsymbol{\theta})$ is the score function for $\boldsymbol{\beta}$ that we would obtain if we had observed the failure times $\mathbf{t}$; hence both (1) and (8) involve averaging a "full-data" score equation over imputed times. As the partial likelihood score function $\mathbf{S}_\beta(\mathbf{t}|\boldsymbol{\delta},\mathbf{x};\boldsymbol{\beta})$ is independent of the baseline distribution and depends only on the ranks of the imputed times rather than on their actual values, we may expect (1) to be less sensitive to correct specification of the baseline distribution than (8). For future use, we define $\mathcal{U}^0_\theta(\mathbf{d}_i;\boldsymbol{\beta},\boldsymbol{\theta})$ and $\mathcal{U}^c_\theta(t_i,\delta_i,\mathbf{x}_i;\boldsymbol{\beta},\boldsymbol{\theta})$ to be the score functions obtained by taking partial derivatives with respect to $\boldsymbol{\theta}$ rather than $\boldsymbol{\beta}$ in (9) and (11).

When some observations are right censored, averaging the score function for the full-data proportional hazards model over imputed failure times may be done in one of two nonequivalent ways. Because right censoring is easily handled within the framework of the proportional hazards model, when $u_i = \infty$, we may choose not to impute a failure time for this observation but instead use the form of the full-data proportional hazards model that considers the ith observation to be right censored. Alternatively, we may consider the censoring interval for a right-censored observation in the same way as any other censoring interval, and impute a failure time. Thus we define our "censoring" indicator variable in the following way. If u_i is finite, then $\delta_i = 1$; whereas if $u_i = \infty$, then $\delta_i = 1$ if a failure time is to be imputed and $\delta_i = 0$ if one is using the form of the full-data proportional hazards model in which the ith observation is censored (in which case $t_i = l_i$). When imputing failure times for individuals for which $u_i = \infty$, the model for the baseline cdf must be specified for all times t; if these observations are treated as right censored, then the baseline cdf need be specified only up to the largest of the finitely valued u_i's.

3. CONSISTENCY AND ASYMPTOTIC NORMALITY OF $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$

We now establish the existence of a $\sqrt{n}$-consistent solution of the estimating equations (5) and (6), as well as the asymptotic normality of such a solution. Let $\boldsymbol{\beta}_0$ and $\boldsymbol{\theta}_0$ denote the true values of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. First, note that by (1), assuming that $F_0(t|\boldsymbol{\theta}_0)$ and the proportional hazards assumption are correct, the unconditional expected value of $\mathbb{S}_\beta(\boldsymbol{\beta}_0,\boldsymbol{\theta}_0)$ is equal to the unconditional expected value of $\mathbf{S}_\beta(\mathbf{t}|\boldsymbol{\delta},\mathbf{x};\boldsymbol{\beta}_0)$, which in turn is 0 (Cox 1975). Hence the estimating equation (5) is unbiased. By the usual properties of maximum likelihood estimation, the estimating equation (6) is also unbiased. If $\mathbb{S}_\beta(\boldsymbol{\beta},\boldsymbol{\theta})$ were a sum of iid terms for each observation, then we could use standard results for estimating equations to conclude asymptotic consistency and normality of $(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\theta}})$. But this not being the case, a deeper analysis is required to establish these asymptotic results.

The details on the regularity assumptions needed for the following theorems are deferred to the Appendix; detailed proofs are available from the authors. For notational convenience, let γ and γ_0 denote $(\boldsymbol{\beta},\boldsymbol{\theta})$ and $(\boldsymbol{\beta}_0,\boldsymbol{\theta}_0)$, and let $\|\ \|$ denote the Euclidian norm.

Theorem 3.1. Given any $\varepsilon > 0$, there exist a $M < \infty$ and an integer n_0 such that

$$\Pr[(5) \text{ and } (6) \text{ have a solution } \hat{\gamma} \text{ with } n^{1/2}\|\hat{\gamma}-\gamma_0\| \le M] \ge 1-\varepsilon \quad \forall \quad n \ge n_0.$$

To establish the asymptotic normality of $\sqrt{n}(\hat{\gamma}-\gamma_0)$, we first show that $n^{-1/2}(\mathbb{S}_\beta(\boldsymbol{\beta}_0,\boldsymbol{\theta}_0), \mathbb{U}^0_\theta(\boldsymbol{\beta}_0,\boldsymbol{\theta}_0))$ is asymptotically normally distributed. Lin and Wei (1989) showed that $n^{-1/2}\mathbf{S}_\beta(\mathbf{t}|\boldsymbol{\delta},\mathbf{x};\boldsymbol{\beta}_0)$ can be written as a sum of iid terms plus terms that are $o_p(1)$; specifically

$$n^{-1/2}\mathbf{S}_\beta(\mathbf{t}|\boldsymbol{\delta},\mathbf{x};\boldsymbol{\beta}_0) = n^{-1/2}\sum_{i=1}^n \boldsymbol{\phi}(t_i,\mathbf{x}_i,\delta_i;\boldsymbol{\beta}_0) + o_p(1), \tag{12}$$

where

$$\boldsymbol{\phi}(t_i, \mathbf{x}_i, \delta_i; \boldsymbol{\beta}) = \delta_i\left(\mathbf{x}_i - \frac{\mathbf{s}^{(1)}(\boldsymbol{\beta}, t_i)}{s^{(0)}(\boldsymbol{\beta}, t_i)}\right) - e^{\boldsymbol{\beta}' \cdot \mathbf{x}_i} \int_0^{t_i} \left(\mathbf{x}_i - \frac{\mathbf{s}^{(1)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)}\right) \frac{dG(t)}{s^{(0)}(\boldsymbol{\beta}, t)} \quad (13)$$

and where

$$G(t) = E\left[\frac{1}{n}\sum_{i=1}^{n} I(t_i \le t, \delta_i = 1)\right],$$

$$\mathbf{S}^{(r)}(\boldsymbol{\beta}, t) = \frac{1}{n}\sum_{i=1}^{n} I[t_i \ge t] e^{\boldsymbol{\beta}' \cdot \mathbf{x}_i} \mathbf{x}_i^{\otimes r}, \quad (14)$$

and

$$\mathbf{s}^{(r)}(\boldsymbol{\beta}, t) = E[\mathbf{S}^{(r)}(\boldsymbol{\beta}, t)]. \quad (15)$$

Lemma A.1 in the Appendix establishes that termwise integration of (12) is allowable, enabling us to write

$$n^{-1/2}\mathbb{S}_\beta(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0) = n^{-1/2}\sum_{i=1}^{n} \boldsymbol{\psi}(\mathbf{d}_i; \boldsymbol{\beta}_0, \boldsymbol{\theta}_0) + o_p(1), \quad (16)$$

where

$$\boldsymbol{\psi}(\mathbf{d}_i; \boldsymbol{\beta}_0, \boldsymbol{\theta}_0) = \begin{cases} \int_{l_i}^{u_i} \boldsymbol{\phi}(t_i, \mathbf{x}_i, \delta_i; \boldsymbol{\beta}_0)\, dF(t_i | l_i, u_i, \mathbf{x}_i; \boldsymbol{\beta}_0, \boldsymbol{\theta}_0) & \text{if } \delta_i = 1 \\ \boldsymbol{\phi}(l_i, \mathbf{x}_i, \delta_i; \boldsymbol{\beta}_0) & \text{if } \delta_i = 0 \end{cases} \quad (17)$$

Using (16), it is easily established that

$$n^{-1/2}(\mathbb{S}_\beta(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0), \mathbb{U}_\theta^0(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \boldsymbol{\Psi}), \quad (18)$$

where $\boldsymbol{\Psi}$, the variance–covariance matrix of $\boldsymbol{\psi}(\mathbf{d}_i; \boldsymbol{\beta}_0, \boldsymbol{\theta}_0)$ and $\mathcal{U}_\theta^0(\mathbf{d}_i; \boldsymbol{\beta}_0, \boldsymbol{\theta}_0)$, is defined in (A.3)–(A.6) in the Appendix. Asymptotic normality of $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$ follows in part from (18) and is summarized in the following theorem.

Theorem 3.2. Let $\hat{\gamma} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$ solve (5) and (6) and $n^{1/2}(\hat{\gamma} - \gamma_0) = \mathcal{O}_p(1)$. Then

$$n^{1/2}(\hat{\gamma} - \gamma_0) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \mathbf{V}),$$

where the variance–covariance matrix $\mathbf{V}$ has the sandwich form

$$\mathbf{V} = \mathbf{A}^{-1} \cdot \boldsymbol{\Psi} \cdot \mathbf{A}^{-1} \quad (19)$$

and where $-n\mathbf{A}$, the expected value of the Jacobian matrix of the score equations, is defined in (A.7)–(A.12) in the Appendix.

A final result that follows from Lemma A.2 in the Appendix and Theorem 3.1 is that $\mathbf{A}$ can be consistently estimated by $\mathbf{A}_n(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$, where $-n\mathbf{A}_n$ is the Jacobian matrix of the score equations, and is given by

$$\mathbf{A}_n(\boldsymbol{\beta}, \boldsymbol{\theta}) \equiv \begin{bmatrix} \mathbf{A}_{11}^n(\boldsymbol{\beta}, \boldsymbol{\theta}) & \mathbf{A}_{12}^n(\beta, \boldsymbol{\theta}) \\ \mathbf{A}_{21}^n(\beta, \boldsymbol{\theta}) & \mathbf{A}_{22}^n(\beta, \boldsymbol{\theta}) \end{bmatrix} = -\frac{1}{n}\begin{bmatrix} \frac{\partial \mathbb{S}_\beta(\boldsymbol{\beta}, \boldsymbol{\theta})}{\partial \boldsymbol{\beta}} & \frac{\partial \mathbb{S}_\beta(\boldsymbol{\beta}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\ \frac{\partial \mathbb{U}_\theta^0(\boldsymbol{\beta}, \boldsymbol{\theta})}{\partial \boldsymbol{\beta}} & \frac{\partial \mathbb{U}_\theta^0(\boldsymbol{\beta}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \end{bmatrix}. \quad (20)$$

$\mathbf{A}_{21}^n$ and $\mathbf{A}_{22}^n$ are easily obtained by differentiation of (6) with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, whereas $\mathbf{A}_{11}^n$ and $\mathbf{A}_{12}^n$ are given by

$$\mathbf{A}_{11}^n(\boldsymbol{\beta}, \boldsymbol{\theta}) = \frac{1}{n}\int [\mathcal{I}_\beta(\mathbf{t}|\boldsymbol{\delta}, \mathbf{x}; \boldsymbol{\beta}) - \mathbf{S}_\beta(\mathbf{t}|\boldsymbol{\delta}, \mathbf{x}; \boldsymbol{\beta}) \times [\mathbb{U}_\beta^c(\mathbf{t}|\boldsymbol{\delta}, \mathbf{x}; \boldsymbol{\beta}, \boldsymbol{\theta}) - \mathbb{U}_\beta^0(\boldsymbol{\beta}, \boldsymbol{\theta})]^T]\, dF(\mathbf{t}|\mathbf{d}; \boldsymbol{\beta}, \boldsymbol{\theta}), \quad (21)$$

where $\mathcal{I}_\beta(\mathbf{t}|\boldsymbol{\delta}, \mathbf{x}; \boldsymbol{\beta})$ is the observed information matrix for the full-data proportional hazards model and

$$\mathbf{A}_{12}^n(\boldsymbol{\beta}, \boldsymbol{\theta}) = -\frac{1}{n}\int \mathbf{S}_\beta(\mathbf{t}|\boldsymbol{\delta}, \mathbf{x}; \boldsymbol{\beta}) \times [\mathbb{U}_\theta^c(\mathbf{t}|\boldsymbol{\delta}, \mathbf{x}; \boldsymbol{\beta}, \boldsymbol{\theta}) - \mathbb{U}_\theta^0(\boldsymbol{\beta}, \boldsymbol{\theta})]^T\, dF(\mathbf{t}|\mathbf{d}; \boldsymbol{\beta}, \boldsymbol{\theta}). \quad (22)$$

4. SOLVING THE SCORE EQUATION

The score (2) is difficult to solve using standard numerical methods, as the integrals are usually intractable. We show here how these score equations can be solved using Monte Carlo techniques. Let $\hat{\boldsymbol{\theta}}(\boldsymbol{\beta})$ denote the value of $\boldsymbol{\theta}$ that solves (6) for a fixed value of $\boldsymbol{\beta}$. For each $j \ge 1$, let $\mathbf{t}_{jk}$, $k = 1, \ldots, K$ be iid samples from the product distribution $F(\mathbf{t}|\mathbf{d}; \boldsymbol{\beta}_j, \hat{\boldsymbol{\theta}}(\boldsymbol{\beta}_j))$. Note that

$$\bar{\mathbf{S}}_j = \frac{1}{K}\sum_{k=1}^{K} \mathbf{S}_\beta(\mathbf{t}_{jk}|\boldsymbol{\delta}, \mathbf{x}; \boldsymbol{\beta}_j) \quad (23)$$

is an unbiased estimator of $\mathbb{S}_\beta(\boldsymbol{\beta}_j, \hat{\boldsymbol{\theta}}(\boldsymbol{\beta}_j))$. Hence our task is to solve (2) where we can obtain values only of an unbiased estimator of $\mathbb{S}_\beta$. The stochastic approximation scheme of Ruppert, Reisch, Deriso, and Carroll (1984) can be used to construct the sequence $\boldsymbol{\beta}_j$ recursively; specifically, we choose

$$\boldsymbol{\beta}_{j+1} = \boldsymbol{\beta}_j - \frac{1}{j}\hat{\mathbf{D}}_j^{-1} \cdot \bar{\mathbf{S}}_j, \qquad j = 1, 2, \ldots, \quad (24)$$

where $\hat{\mathbf{D}}_j$ is a sequence of values that converges to

$$\mathbf{D}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}) \equiv \left.\frac{\partial \mathbb{S}_\beta}{\partial \boldsymbol{\beta}}\right|_{\substack{\beta = \hat{\beta}, \\ \theta = \hat{\theta}}} = -n(\mathbf{A}_{11}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}) - \mathbf{A}_{12}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})\mathbf{A}_{22}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})^{-1}\mathbf{A}_{21}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})). \quad (25)$$

We describe construction of the sequence $\hat{\mathbf{D}}_j$ later.

Note that $\mathbf{A}_{11}$ and $\mathbf{A}_{12}$ are each the expected value of a function of $\mathbf{t}$ with respect to the distribution $F(\mathbf{t}|\mathbf{d}; \beta, \boldsymbol{\theta})$.

Hence $\mathbf{A}$ can be estimated after j steps by $\hat{\mathbf{A}}_j$, where

$$\hat{\mathbf{A}}_j = \begin{bmatrix} \hat{\mathbf{A}}_{11j} & \hat{\mathbf{A}}_{12j} \\ \mathbf{A}_{21}(\boldsymbol{\beta}_j, \hat{\boldsymbol{\theta}}(\boldsymbol{\beta}_j)) & \mathbf{A}_{22}(\boldsymbol{\beta}_j, \hat{\boldsymbol{\theta}}(\boldsymbol{\beta}_j)) \end{bmatrix}$$

and where

$$\hat{\mathbf{A}}_{11j} = \frac{1}{j}\sum_{j'=1}^{j}\frac{1}{K}\sum_{k=1}^{K} - \mathcal{I}_\beta(\mathbf{t}_{j'k}|\boldsymbol{\delta}, \mathbf{x}; \boldsymbol{\beta}_{j'}) + \mathbf{S}_\beta(\mathbf{t}_{j'k}|\boldsymbol{\delta}, \mathbf{x}; \boldsymbol{\beta})\mathbb{U}^c_\beta(\mathbf{t}_{j'k}|\boldsymbol{\delta}, \mathbf{x}; \boldsymbol{\beta}_{j'}, \hat{\boldsymbol{\theta}}(\boldsymbol{\beta}_{j'}))^T$$

and

$$\hat{\mathbf{A}}_{12j} = \frac{1}{j}\sum_{j'=1}^{j}\frac{1}{K}\sum_{k=1}^{K}\mathbf{S}_\beta(\mathbf{t}_{j'k}|\boldsymbol{\delta}, \mathbf{x}; \boldsymbol{\beta}_{j'}) \times \mathbb{U}^c_\theta(\mathbf{t}_{j'k}|\boldsymbol{\delta}, \mathbf{x}; \boldsymbol{\beta}_{j'}, \hat{\boldsymbol{\theta}}(\boldsymbol{\beta}_{j'}))^T.$$

Note that in defining $\hat{\mathbf{A}}_{11j}$ and $\hat{\mathbf{A}}_{12j}$ we have used the score (2) to eliminate terms involving $\mathbb{U}^0_\beta(\boldsymbol{\beta}, \boldsymbol{\theta})$ and $\mathbb{U}^0_\theta(\boldsymbol{\beta}, \boldsymbol{\theta})$ from (21) and (22) that are equal to 0.

From (25), after j steps $\mathbf{D}$ can be estimated by $\hat{\mathbf{D}}_j$, given by

$$\hat{\mathbf{D}}_j = -n(\hat{\mathbf{A}}_{11j} - \hat{\mathbf{A}}_{12j}\mathbf{A}_{22}(\boldsymbol{\beta}_j, \hat{\boldsymbol{\theta}}(\boldsymbol{\beta}_j))^{-1}\mathbf{A}_{21}(\boldsymbol{\beta}_j, \hat{\boldsymbol{\theta}}(\boldsymbol{\beta}_j))). \tag{26}$$

To develop an estimator of $\boldsymbol{\Psi}$, we begin with (16)–(17). Lin and Wei (1989) showed that the function ϕ in (12)–(13) can be estimated by $\hat{\boldsymbol{\phi}}(t_i, \mathbf{x}_i, \delta_i, \boldsymbol{\beta})$, where

$$\hat{\boldsymbol{\phi}}(t_i, \mathbf{x}_i, \delta_i, \boldsymbol{\beta}) = \delta_i\left(\mathbf{x}_i - \frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t_i)}{S^{(0)}(\boldsymbol{\beta}, t_i)}\right) - \sum_{i'=1}^{n}\frac{\delta_{i'}I(t_i \geq t_{i'})e^{\boldsymbol{\beta}\cdot\mathbf{x}_i}}{nS^{(0)}(\boldsymbol{\beta}, t_{i'})}\left\{\mathbf{x}_{i'} - \frac{\mathbf{S}^{(1)}(\boldsymbol{\beta}, t_{i'})}{S^{(0)}(\boldsymbol{\beta}, t_{i'})}\right\}$$

and where $S^{(0)}(\boldsymbol{\beta}, t_i)$ and $\mathbf{S}^{(1)}(\boldsymbol{\beta}, t_i)$ are defined in (14). Note that $\hat{\boldsymbol{\phi}}(t_i, \mathbf{x}_i, \delta_i, \boldsymbol{\beta})$ is a function of $\boldsymbol{\beta}$ and rank information only and is the influence function for the full-data proportional hazards model (Reid and Crépeau 1985). As each $\boldsymbol{\psi}(\mathbf{d}_i, \boldsymbol{\beta}, \boldsymbol{\theta})$ is the expected value of $\boldsymbol{\phi}(t_i, \mathbf{x}_i, \delta_i, \boldsymbol{\beta})$ with respect to the distribution $F(t_i|\mathbf{d}_i; \boldsymbol{\beta}, \boldsymbol{\theta})$, after j steps we may estimate $\boldsymbol{\psi}(\mathbf{d}_i; \boldsymbol{\beta}, \boldsymbol{\theta})$ by $\hat{\boldsymbol{\psi}}_{ij}$, where

$$\hat{\boldsymbol{\psi}}_{ij} = \begin{cases} \frac{1}{j}\sum_{j'=1}^{j}\frac{1}{K}\sum_{k=1}^{K}\hat{\boldsymbol{\phi}}(t_{ij'k}, \mathbf{x}_i, \delta_i; \boldsymbol{\beta}_{j'}) & \text{if } \delta_i = 1 \\ \hat{\boldsymbol{\phi}}(l_i, \mathbf{x}_i, \delta_i; \boldsymbol{\beta}_j) & \text{if } \delta_i = 0 \end{cases}$$

and where t_{ijk} is the ith component of $\mathbf{t}_{jk}$. Note that $\hat{\boldsymbol{\psi}}_{ij}$ is an estimate of the contribution of the ith individual to the score $\mathbb{S}_\beta(\boldsymbol{\beta}_j, \hat{\boldsymbol{\theta}}(\boldsymbol{\beta}_j))$. Hence from (A.4)–(A.6), an approximant to $\boldsymbol{\Psi}$, denoted by $\boldsymbol{\Psi}_j$, where

$$\hat{\boldsymbol{\Psi}}_j = \begin{bmatrix} \hat{\boldsymbol{\Psi}}_{11j} & \hat{\boldsymbol{\Psi}}_{12j} \\ \hat{\boldsymbol{\Psi}}^T_{12j} & \hat{\boldsymbol{\Psi}}_{22j} \end{bmatrix},$$

can then be constructed by taking

$$\hat{\boldsymbol{\Psi}}_{11j} = \frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{\psi}}_{ij}\hat{\boldsymbol{\psi}}^T_{ij},$$

$$\hat{\boldsymbol{\Psi}}_{12j} = \frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{\psi}}_{ij}\mathcal{U}^0_\theta(\mathbf{d}_i; \boldsymbol{\beta}_j, \hat{\boldsymbol{\theta}}(\boldsymbol{\beta}_j))^T,$$

and

$$\hat{\boldsymbol{\Psi}}_{22j} = \frac{1}{n}\sum_{i=1}^{n}\mathcal{U}^0_\theta(\mathbf{d}_i; \boldsymbol{\beta}_j, \hat{\boldsymbol{\theta}}(\boldsymbol{\beta}_j))\mathcal{U}^0_\theta(\mathbf{d}_i; \boldsymbol{\beta}_j, \hat{\boldsymbol{\theta}}(\boldsymbol{\beta}_j))^T.$$

It is important to note that each of the quantities required to estimate $\mathbf{A}$, $\mathbf{D}$, and $\boldsymbol{\Psi}$ can be calculated recursively, so that only one set of $\mathbf{t}_{jk}$'s is needed at a time and past values of $\mathbf{t}_{jk}$ never need be stored. To determine how close $\boldsymbol{\beta}_j$ is to $\hat{\boldsymbol{\beta}}$, Ruppert et al. (1984) showed that conditional on the observed data,

$$j^{1/2}(\boldsymbol{\beta}_j - \hat{\boldsymbol{\beta}}) \sim \mathrm{N}(0, \mathbf{D}^{-1}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})\boldsymbol{\Sigma}\mathbf{D}^{-1}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})), \tag{27}$$

where

$$\boldsymbol{\Sigma} = \frac{1}{K}\operatorname{cov}_{F(\cdot|\mathbf{d};\hat{\beta},\hat{\theta})}[\mathbf{S}_\beta(\mathbf{t}|\boldsymbol{\delta}, \mathbf{x}; \hat{\boldsymbol{\beta}})] \tag{28}$$

is a measure of the variability introduced into our evaluation of $\mathbb{S}_\beta(\boldsymbol{\beta}, \boldsymbol{\theta})$ by approximating it with the block average (23). A recursive estimator of $\boldsymbol{\Sigma}$ denoted by $\hat{\boldsymbol{\Sigma}}_j$, can be constructed by taking

$$\hat{\boldsymbol{\Sigma}}_{j+1} = \hat{\boldsymbol{\Sigma}}_j + \frac{1}{j}(\bar{\mathbf{S}}_j\bar{\mathbf{S}}_j^T - \hat{\boldsymbol{\Sigma}}_j), \tag{29}$$

which results in estimating $\boldsymbol{\Sigma}$ by the empirical covariance matrix of the block-averaged values $\bar{\mathbf{S}}_j$.

Using the stochastic approximation scheme, we now describe how $\boldsymbol{\beta}_{j+1}$ is obtained, having just obtained $\boldsymbol{\beta}_j$. First, $\hat{\boldsymbol{\theta}}(\boldsymbol{\beta}_j)$ is obtained by maximizing (4) with respect to $\boldsymbol{\theta}$ with $\boldsymbol{\beta} = \boldsymbol{\beta}_j$; the search can be started at $\hat{\boldsymbol{\theta}}(\boldsymbol{\beta}_{j-1})$ for computational efficiency. Then for $k = 1, \ldots, K$, the vectors of imputed failure times $\mathbf{t}_{jk}$ are obtained from $F(\mathbf{t}|\mathbf{d}; \boldsymbol{\beta}_j, \hat{\boldsymbol{\theta}}(\boldsymbol{\beta}_j))$. As each $\mathbf{t}_{jk}$ is generated, $\bar{\mathbf{S}}_j$ is calculated recursively; at the same time, $\hat{\mathbf{A}}_{11(j-1)}$, $\hat{\mathbf{A}}_{12(j-1)}$ and the $\hat{\boldsymbol{\psi}}_{i(j-1)}$'s are updated recursively until $\hat{\mathbf{A}}_{11j}$, $\hat{\mathbf{A}}_{12j}$, and the $\hat{\boldsymbol{\psi}}_{ij}$'s are obtained. Then $\hat{\mathbf{D}}_j$ is calculated using (26) and $\boldsymbol{\beta}_{j+1}$ is calculated using (24). Finally, $\hat{\boldsymbol{\Sigma}}_{j+1}$ is calculated using (29) and $\hat{\boldsymbol{\Sigma}}_j$. The entire process is repeated until some stopping rule is met. It is not necessary to calculate $\hat{\boldsymbol{\Psi}}_j$ at each stage, unless a sequential stopping rule requiring an estimate of $\mathbf{V}$ is being used; in our simulation examples we take J stochastic approximation steps where J is fixed, at which point $\hat{\boldsymbol{\Psi}}_J$ is calculated. The starting value of $\boldsymbol{\beta}$, $\boldsymbol{\beta}_1$, is arbitrary. After J steps, an estimate of $\mathbf{V}$, denoted by $\hat{\mathbf{V}}$, can be obtained by $\hat{\mathbf{V}} = \hat{\mathbf{A}}_J^{-1} \cdot \hat{\boldsymbol{\Psi}}_J \cdot \hat{\mathbf{A}}_J^{-1}$.

The parameter K, first used in (23) and which we call the block size, determines the number of rankings generated before a new value of $\boldsymbol{\beta}$ is selected. The advantage of using $K > 1$ is that a new value of $\hat{\boldsymbol{\theta}}(\boldsymbol{\beta})$ must be determined each time that $\boldsymbol{\beta}$ is changed. Equations (27)–(28) show that the choice of K does not effect the variance of $\boldsymbol{\beta}_J$ about $\hat{\boldsymbol{\beta}}$ as long as $J \cdot K$, the total number of rankings generated, is constant. The variance of $\boldsymbol{\beta}_J$ about $\hat{\boldsymbol{\beta}}$ can be approximated by $\hat{\mathbf{V}}_J = \hat{\mathbf{D}}_J^{-1}\hat{\boldsymbol{\Sigma}}_J\hat{\mathbf{D}}_J^{-1}$.

Satten (1996) showed that the total sampling variance of $\boldsymbol{\beta}_J$ about the true value $\boldsymbol{\beta}_0$ is the sum of the variance of $\boldsymbol{\beta}_J$ about $\hat{\boldsymbol{\beta}}$ due to using stochastic approximation and the sampling variance of $\hat{\boldsymbol{\beta}}$ about $\boldsymbol{\beta}_0$. Thus the estimated total sampling variance of $\boldsymbol{\beta}_J$ can be obtained as the sum of $\hat{\mathbf{V}}$ and $\hat{\mathbf{V}}_J$. However, in most cases the stochastic approximation should be iterated until $\hat{\mathbf{V}}_J$ is negligible compared to $\hat{\mathbf{V}}$.

In theory, the value of $\boldsymbol{\beta}_1$ is arbitrary; in practice, a good starting value is important. To ensure a good starting value, we use a "burn-in" period with $K = 1$, during which $\hat{\mathbf{D}}_j$ is approximated by $\hat{\mathbf{D}}_j = -(1/j)\sum_{j'=1}^{j} \mathcal{I}_\beta(\mathbf{t}_{j'1}|\boldsymbol{\delta}, \mathbf{x}; \boldsymbol{\beta}_{j'})$. Starting with $\boldsymbol{\beta}_1 = \mathbf{0}$, the stochastic approximation is run for J_b steps, and the values $\boldsymbol{\beta}_{J_b}$ and $\hat{\boldsymbol{\theta}}(\boldsymbol{\beta}_{J_b})$ are then used as starting values for the full scheme described earlier. If convergence is still a problem, Ruppert et al. (1984) have suggested replacing the factor $1/j$ in (24) by $1/(j + k_0)$ so that the initial steps are not too large.

5. SIMULATIONS

To assess the performance of our method, we conducted analyses using simulated data with known distributions. The goals of our simulation studies were to compare the performance of the missing–failure time and fully parametric methods when the baseline distribution has been incorrectly specified, to compare the efficiency of the parametric and missing–failure time methods when the baseline distribution has been correctly specified, to compare both the missing–failure time and parametric methods to the missing-rank method of Satten (1996), and to assess the performance of the variance–covariance estimators $\hat{\mathbf{V}}$ and $\hat{\mathbf{V}}_J$.

For each of the two baseline distributions, we generated 100 datasets each containing 500 failure times, with a binary covariate corresponding to a hazard ratio of $\exp(\beta) = 2$ ($\beta = .693$). In each dataset, 250 observations were generated for each covariate value. The baseline distributions selected were a Weibull distribution with shape parameter 2 and location parameter .01 and a log-logistic distribution with shape parameter 4 and location parameter .01.

Each simulated dataset was then subject to three types of censoring: "light," "heavy," and "current-status." In light censoring, the censoring interval for the ith ordered failure time overlaps only with the $(i-1)$th and $(i+1)$th ordered failure times. Heavy censoring was achieved by starting an independent renewal process for each observation in each dataset at time 0; the increments followed a lognormal distribution increments with mean 18.7 and standard error 209.4 for the Weibull case and mean 15 and standard error 109.8 for the log-logistic case. A maximum number of renewals was set; if one of the renewal intervals contained the true failure time, then this interval was used as the censoring interval; otherwise, the observation was considered to be right censored at the last renewal time. Approximately 20% of the observations were right censored. Current status data were created by generating a single independent log-logistic random variate τ for each failure time in each dataset and choosing as the censoring interval which of the intervals $[0, \tau)$ or $[\tau, \infty)$ contained the true failure time. The extent of censoring can be measured by the average proportion of intervals that each interval overlaps with; this proportion is approximately 2/500 for the light-censored data, .78 for the heavy-censored data using the loglogistic baseline, .72 for the heavy-censored data using the Weibull baseline, .92 for the current-status data using the log-logistic baseline, and .91 for current-status data using the Weibull baseline.

For each baseline distribution and for each censoring type, we then estimated the log-hazard ratio β and the standard error of β using the parametric, missing–failure time, and missing-rank methods. For reference, we also estimated β and its standard error using the exact failure times using the standard proportional hazards model. For the missing–failure time and missing-rank methods, we used $J = 400$ stochastic approximation steps with a block size of $K = 50$ and estimated the standard error of the Monte Carlo approximant to the true parameter estimate. All analyses using the missing–failure time and parametric methods assumed that the baseline distribution was a Weibull distribution. Analyses of heavily censored and current-status datasets using the missing–failure time method were performed in two ways: first, observations for which $u_i = \infty$ were considered to be right censored, and second, these observations were considered to be interval censored, and a failure time was imputed.

Table 1 summarizes the results of our simulations using the log-logistic as the baseline distribution. For the light-censored case, both the missing–failure time and missing-rank methods give point estimates and standard errors that are very close to the values obtained using exact failure times. For heavy censored and current-status datasets, the missing–failure time method also performs well. An interesting trend is that as the amount of censoring increases, the performance of the missing-rank method decreases and the performance of the parametric method improves. For current-status datasets, the parametric method actually has a smaller bias than the nonparametric method; numerical experiments (results not shown) indicate this bias in the missing-rank method appears to be a finite–sample size phenomenon. Note that although the same misspecified baseline distribution is used in both the missing–failure time and parametric methods, using (2) rather than (8) does protect against misspecification of the baseline distribution, yielding unbiased estimates of the hazard ratio for the three types of censoring considered. Similar results were obtained when failure times were imputed for the right-censored observations.

Table 2 summarizes the results of our simulations using the Weibull distribution as the baseline. For the light-censored case, all three methods give point estimates and standard errors which are very close to the values obtained using exact failure times. For both the light- and heavy-censored datasets, the efficiency of the missing–failure time method and the missing-rank method relative to the parametric method, (defined as the average over the 100 datasets of the estimated variance of $\hat{\boldsymbol{\beta}}$ divided by the estimated vari-

Table 1. Simulation Results for Log-Logistic Baseline[a]

Method	β^*	$SE(\beta^*)$	$SE(\hat{\beta} - \beta^*)$
	No censoring		
Exact[b]	.703	.0937	
	(.464, .946)	(.0879, .103)	
	Light censoring		
Parametric[c]	.855	.0971	
	(.515, 1.099)	(.0928, .103)	
Missing failure time RC[d]	.703	.0944	5.5×10^{-6}
	(.465, .945)	(.0882, .103)	$(3.3 \times 10^{-6}, 18.2 \times 10^{-6})$
Missing rank[e]	.703	.0946	1.1×10^{-5}
	(.465, .945)	(.0916, .0990)	$(.6 \times 10^{-5}, 3.9 \times 10^{-5})$
	Heavy censoring		
Parametric	.825	.133	
	(.463, 1.121)	(.123, .149)	
Missing failure time RC	.713	.132	7.0×10^{-4}
	(.464, 1.008)	(.120, .148)	$(5.3 \times 10^{-4}, 9.1 \times 10^{-4})$
Missing failure time IC[f]	.712	.133	8.9×10^{-4}
	(.446, 1.006)	(.118, .153)	$(7.4 \times 10^{-4}, 11.7 \times 10^{-4})$
Missing rank	.726	.133	4.7×10^{-4}
	(.453, 1.025)	(.123, .150)	$(4.3 \times 10^{-4}, 5.3 \times 10^{-4}$
	Current status		
Parametric	.739	.153	
	(.394, 1.264)	(.141, .169)	
Missing failure time RC	.714	.147	6.4×10^{-4}
	(.389, 1.233)	(.125, .165)	$(4.7 \times 10^{-4}, 8.7 \times 10^{-4})$
Missing failure time IC	.712	.158	1.4×10^{-3}
	(.393, 1.230)	(.141, .178)	$(1.1 \times 10^{-3}, 2.3 \times 10^{-3})$
Missing rank	.742	.162	5.4×10^{-4}
	(.399, 1.288)	(.148, .182)	$(5.1 \times 10^{-4}, 5.9 \times 10^{-4})$

[a] Values shown are means of 100 simulations, and the ranges of the 100 simulations are shown in parentheses.
[b] Exact refers to fitting the proportional hazards model to the actual failure times with no censoring.
[c] Parametric refers to a fully parametric model for analyzing interval-censored data.
[d] Missing failure time RC refers to the proposed estimating equation approach for interval-censored data using the right-censoring form of the proportional hazards model.
[e] Missing rank refers to Satten's rank-based proportional hazards model for interval censored data.
[f] Missing failure time IC refers to the proposed estimating equation approach for interval-censored data using the methodology in which a failure time is imputed for every individual.

ance of $\hat{\beta}$ obtained using the parametric method) is nearly 1. Again, note that the difference between the missing-rank method estimates and the true value increases as the extent of censoring increases; this probably explains the diminished efficiency of the missing-rank method. (Coverage of the confidence intervals for the missing-rank method is better than the point estimates would lead one to expect.) Similar results were obtained when failure times were imputed for the right-censored observations in the missing–failure time method.

In all cases the standard error of β_J about $\hat{\beta}$ is smaller for the missing–failure time method than for the missing-rank method, even though both used the same number of stochastic approximation steps and the same block size. This is because the successive rankings in the missing-rank method are independent, whereas those of the missing–failure time method carry dependence induced by the Gibbs sampler. This results in faster convergence for the missing–failure time method.

6. DISCUSSION

An important but subtle distinction must be made between the rank-based methods used here and those of Satten (1996). For example, consider two observations with failure times t_1 and t_2 and censoring intervals [5, 10) and [15, 20). The only requirement of a method based purely on rank information is that the first observation must precede the second observation. Note that any failure times t_1 and t_2 that satisfy $t_1 < t_2$ are consistent with the rank information, but imputed failure times must also fall within the censoring intervals. This added restriction leads to rank-based inference that is not fully independent of the baseline hazard rate; however, this dependence appears to be very small in practice. The advantage of (2) over the missing-rank approach of Satten (1996) is that it is easier to sample rankings from (3) than from the nonparametric distribution of rankings when solving the estimating equations.

In simulation studies we have shown that our proposed estimator provides an advantage over a fully parametric estimator in that dependence of the estimates on correct specification of the baseline distribution is greatly reduced. Even in situations where the baseline distribution was misspecified, our new estimator provided apparently unbiased estimation, and even outperformed the method of Satten (1996) in cases of extreme censoring. In our simulations we made no effort to determine if the misspecification of the baseline distribution could be determined, as the goal of the simula-

Table 2. Simulation Results for Weibull Baseline[a]

Method	β^*	$SE(\beta^*)$	$SE(\hat{\beta} - \beta^*)$	Efficiency
		No censoring		
Exact[b]	.697	.0944		
	(.416, .947)	(.0890, .103)		
		Light censoring		
Parametric[c]	.698	.0911		
	(.422, .930)	(.0787, .0956)		
Missing failure time RC[d]	.697	.0944	13.2×10^{-6}	
	(.416, .947)	(.0890, .103)	(4.6×10^{-6},27.8×10^{-6})	.93
Missing rank[e]	.697	.0945	1.2×10^{-5}	
	(.416, .947)	(.0914, .0985)	($.6 \times 10^{-5}$, 3.7×10^{-5})	.93
		Heavy censoring		
Parametric	.706	.125		
	(.450, 1.009)	(.119, .136)		
Missing failure time RC	.709	.126	6.2×10^{-4}	
	(.443, 1.019)	(.117, .139)	(5.0×10^{-4}, 7.7×10^{-4})	1.00
Missing failure time IC[f]	.709	.128	8.3×10^{-4}	
	(.444, 1.014)	(.119, .140)	(7.0×10^{-4}, 10.5×10^{-4})	.96
Missing rank	.722	.131	4.6×10^{-4}	
	(.461, 1.045)	(.123, .140)	(4.3×10^{-4}, 4.9×10^{-4})	.92
		Current status		
Parametric	.708	.136		
	(.308, 1.000)	(.125, .154)		
Missing failure time RC	.708	.132	6.2×10^{-4}	
	(.312, 1.015)	(.107, .155)	(4.7×10^{-4}, 8.4×10^{-4})	1.08
Missing failure time IC	.708	.141	1.0×10^{-3}	
	(.305, 1.016)	(.126, .168)	($.9 \times 10^{-3}$, 1.4×10^{-3})	.93
Missing rank	.746	.147	5.1×10^{-4}	
	(.339, 1.088)	(.133, .168)	(4.7×10^{-4}, 5.6×10^{-4})	.86

[a] Values shown are means of 100 simulations, and the ranges of the 100 simulations are shown in parentheses.
[b] Exact refers to fitting the proportional hazards model to the actual failure times with no censoring.
[c] Parametric refers to a fully parametric model for analyzing interval censored data.
[d] Missing failure time RC refers to the proposed estimating equation approach for interval censored data using the right-censoring form of the proportional hazards model.
[e] Missing rank refers to Satten's rank based proportional hazards model for interval censored data.
[f] Missing failure time IC refers to the proposed estimating equation approach for interval censored data using the methodology in which a failure time is imputed for every individual.

tions was to explore the sensitivity of our estimator and the parametric estimator to this type of error. We stress that in actual use, a careful analysis would include exploration of the proper baseline, or, at the very least, use of a flexible family of baseline models (such as spline models).

Our proposed missing–failure time estimator has several advantages over the missing-rank approach of Satten (1996). First, generating the imputed failure times is much easier than the rank-generating scheme used in the missing-rank method, which requires a Gibbs sampler. This simplicity results in a faster computer program as well as a more efficient stochastic approximation scheme. Second, when datasets are very heavily censored (such as in current-status data), the missing-rank method seems to require larger sample sizes to produce an unbiased estimate. In other simulations (results not shown), the missing–failure time method performed well for datasets with 100 observations each. Third, failure times for right-censored observations do not need to be imputed; this results in considerable computational efficiency for data that are heavily right-censored, such as current-status data. Fourth, having an imputed failure time available allows straightforward generalizations of the interval-censoring problem in various ways. For example, the missing–failure time method may be immediately generalized to time-dependent covariates (as long as the value of the covariate is known at all times from 0 to the right endpoint of the censoring interval) by simply using the time-dependent covariate forms of the full-data partial likelihood score function and the function $\hat{\phi}$; there does not appear to be a similar generalization of the nonparametric method. Finally, the rank-based method does not have a firm theoretical foundation, and formal determination of the asymptotic properties of the rank-based estimator remains open.

Our simulation studies revealed a finite-sample bias for the rank-based marginal likelihood method that increases with the extent of censoring. For current-status data, our simulations suggest that the missing-rank method should be avoided unless a very large sample is available. In cases with less extreme censoring, such as the heavily censored data considered in our simulations, the missing-rank method performed well (see also Satten 1996) and remains the only option when a methodology that does not require estimation of the baseline distribution is required.

APPENDIX: TECHNICAL DETAILS

Throughout we assume that $(t_i, l_i, u_i, \delta_i, \mathbf{x}_i)$ are iid and $\mathbf{x}$ is bounded; if $u_i = \infty$ implies use of the right-censored form of the proportional hazards model, then we further assume the standard proportional hazards regularity condition that the support of the failure time t contains that of the censoring time and that this containment is proper if the failure time is bounded. Let $\|\mathbf{a}\|$ denote the Euclidean norm of $\mathbf{a}$ if $\mathbf{a}$ is a vector and the operator norm of $\mathbf{a}$ if $\mathbf{a}$ is a matrix. We assume, without specifying it on a case-by-case basis, that certain quantities are differentiable in the parameters $\boldsymbol{\gamma} = (\boldsymbol{\beta}, \boldsymbol{\theta})$, at least in a neighborhood of the true value $\gamma_0 = (\beta_0, \boldsymbol{\theta}_0)$, and that the order of certain differentiations and integrations can be interchanged. For $\varepsilon > 0$ and $\mathbf{b} \in \mathcal{R}^p$ for some p, let $N_\varepsilon(\mathbf{b}) = \{\mathbf{b}' \| \|\mathbf{b}' - \mathbf{b}\| < \varepsilon\}$ denote the ε neighborhood of $\mathbf{b}$. Throughout, $\vee$ denotes the supremum of functions of γ over $N_{M/\sqrt{n}}(\gamma_0)$, where M is a finite constant. If the function does not involve $\boldsymbol{\theta}$, then $\vee$ denotes supremum over $N_{M/\sqrt{n}}(\beta_0)$.

We say that an iid sequence indexed by γ, $\{y_i(\gamma), i \geq 1\}$ satisfies the uniform weak law of large numbers (U-WLLN) condition at γ_0 if $E(y_1(\gamma))$ is continuous at γ_0; if for some $\delta > 0$, $E(y_1^*) < \infty$, where $y_1^* = \sup_{\|\gamma - \gamma_0\| < \delta} \|y_1(\gamma)\|$; and if

$$\limsup_{\rho \downarrow 0} E[V_1(\gamma)] = 0 \tag{A.1}$$

for each $\gamma \in N_\delta(\gamma_0)$, where $V_1(\gamma) = \sup\{\|y_1(\gamma') - y_1(\gamma)\| : \gamma' \in N_\rho(\gamma)\}$. If this condition holds, then $n^{-1} \sum y_i(\gamma) \xrightarrow{p} E[y_1(\gamma)]$ uniformly in $\gamma \in N_{\delta/2}(\gamma_0)$ (see, e.g., Datta 1988). Note that (A.1) can be replaced with a more familiar sufficient condition; such as

$$\|y_1(\gamma') - y_1(\gamma)\| \leq d(\|\gamma' - \gamma\|) W_1 \qquad \forall \ \ \gamma, \gamma' \in N_\delta(\gamma_0),$$

where $d(x) \to 0$ as $x \to 0$ and where $E[W_1] < \infty$.

Let $\mathcal{U}^0(\mathbf{d}_i; \boldsymbol{\beta}, \boldsymbol{\theta})^T = (\mathcal{U}^0_\beta(\mathbf{d}_i; \boldsymbol{\beta}, \boldsymbol{\theta})^T, \mathcal{U}^0_\theta(\mathbf{d}_i; \boldsymbol{\beta}, \boldsymbol{\theta})^T)$ and $\mathcal{U}^c(t, \delta_i, \mathbf{x}_i; \boldsymbol{\beta}, \boldsymbol{\theta})^T = (\mathcal{U}^c_\beta(t, \delta_i, \mathbf{x}_i; \boldsymbol{\beta}, \boldsymbol{\theta})^T, \mathcal{U}^c_\theta(t, \delta_i, \mathbf{x}_i; \boldsymbol{\beta}, \boldsymbol{\theta})^T)$. The following regularity conditions are assumed to hold:

C1. $E\|\mathcal{U}^0(\mathbf{d}_1; \beta_0, \boldsymbol{\theta}_0)\|^2 < \infty$; $\|\mathcal{U}^c(t_1, \delta_1, \mathbf{x}_1; \beta_0, \boldsymbol{\theta}_0)\|$ and $w(\mathbf{d}_1, t_1, \delta_1) = \sup_{\gamma \in N_\varepsilon(\gamma_0)} \|(\partial/\partial\gamma)\mathcal{U}^c(t_1; \mathbf{d}_1; \beta_0, \boldsymbol{\theta}_0)\|$ for some $\varepsilon > 0$ have finite moment-generating functions in some neighborhood of 0.
C2. The sequences $\{\partial\psi/\partial\gamma\}$, $\{\partial\mathcal{U}^0/\partial\gamma\}$, $\{\phi\mathcal{U}^{0T}\}$, and $\{\phi\mathcal{U}^{cT}\}$ each satisfies the U-WLLN condition at γ_0.
C3. The matrix $\mathbf{A}$ given in (A.7) is nonsingular.

The following two lemmas are used to prove the theorems stated in Section 3.

Lemma A.1. As $n \to \infty$,

$$n^{-1/2} \mathbb{S}_\beta(\gamma) = n^{-1/2} \sum_{i=1}^{n} \psi(\mathbf{d}_i; \gamma) + o_p(1) \tag{A.2}$$

uniformly in $\gamma \in N_{M/\sqrt{n}}(\gamma_0)$, for each $M < \infty$, where $\psi(\mathbf{d}_i; \gamma)$ is defined in (17).

Lemma A.2. For any $M < \infty$,

$$\vee \|\mathbf{A}^n_{ij}(\gamma) - \mathbf{A}_{ij}\| = o_p(1).$$

Proofs of these two lemmas as well as of the two theorems are available from the authors.

The matrices $\mathbf{A}$ and $\boldsymbol{\Psi}$ were introduced in Section 3. Their formal definitions are

$$\boldsymbol{\Psi} = \begin{bmatrix} \boldsymbol{\Psi}_{11} & \boldsymbol{\Psi}_{12} \\ \boldsymbol{\Psi}_{12}^T & \boldsymbol{\Psi}_{22} \end{bmatrix}, \tag{A.3}$$

where

$$\boldsymbol{\Psi}_{11} = E[\psi(\mathbf{d}_1; \beta_0, \boldsymbol{\theta}_0)\psi(\mathbf{d}_1; \beta_0, \boldsymbol{\theta}_0)^T], \tag{A.4}$$

$$\boldsymbol{\Psi}_{12} = E[\psi(\mathbf{d}_1; \beta_0, \boldsymbol{\theta}_0)\mathcal{U}^0_\theta(\mathbf{d}_1; \beta_0, \boldsymbol{\theta}_0)^T], \tag{A.5}$$

and

$$\boldsymbol{\Psi}_{22} = E[\mathcal{U}^0_\theta(\mathbf{d}_1; \beta_0, \boldsymbol{\theta}_0)\mathcal{U}^0_\theta(\mathbf{d}_1; \beta_0, \boldsymbol{\theta}_0)^T] \tag{A.6}$$

and

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}, \tag{A.7}$$

where

$$\mathbf{A}_{11} = \mathcal{Q}(\beta_0) - E\{\phi(t_1, \mathbf{x}_1, \delta_1; \beta_0)[\mathcal{U}^c_\beta(t_1, \delta_1, \mathbf{x}_1; \beta_0, \boldsymbol{\theta}_0) - \mathcal{U}^0_\beta(\mathbf{d}_1; \beta_0, \boldsymbol{\theta}_0)]^T\}, \tag{A.8}$$

$$\mathbf{A}_{12} = E\{\phi(t_1, \mathbf{x}_1, \delta_1, \beta_0)[\mathcal{U}^c_\theta(t_1, \delta_1, \mathbf{x}_1; \beta_0, \boldsymbol{\theta}_0) - \mathcal{U}^0_\theta(\mathbf{d}_1; \beta_0, \boldsymbol{\theta}_0)]^T\}, \tag{A.9}$$

$$\mathbf{A}_{21} = E\left\{ \left. \frac{\partial\mathcal{U}^0_\theta(\mathbf{d}_1; \boldsymbol{\beta}, \boldsymbol{\theta})}{\partial\boldsymbol{\beta}} \right|_{\substack{\beta=\beta_0,\\ \theta=\theta_0}} \right\}, \tag{A.10}$$

$$\mathbf{A}_{22} = E\left\{ \left. \frac{\partial\mathcal{U}^0_\theta(\mathbf{d}_1; \boldsymbol{\beta}, \boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \right|_{\substack{\beta=\beta_0,\\ \theta=\theta_0}} \right\}, \tag{A.11}$$

and

$$\mathbf{Q}(\boldsymbol{\beta}) = \int \left\{ \frac{\mathbf{s}^{(2)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)} - \frac{\mathbf{s}^{(1)}(\boldsymbol{\beta}, t)^{\otimes 2}}{s^{(0)}(\boldsymbol{\beta}, t)^2} \right\} s^{(0)}(\boldsymbol{\beta}, t)\lambda_0(t)\, dt. \tag{A.12}$$

[Received May 1995. Revised August 1997.]

REFERENCES

Cox, D. R. (1972), "Regression Models and Life Tables" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 26, 187–220.

Cox, D. R. (1975), "Partial Likelihood," *Biometrika*, 62, 269–276.

Datta, S. (1988), "Asymptotically Optimal Bayes Compound and Empirical Bayes Estimators in Exponential Families with Compact Parameter Spaces," Ph. D. dissertation, Michigan State University, Dept. of Statistics and Probability.

DeLong, D. M., Guirguis, G. H., and So, Y. C. (1994), "Efficient Computation of Subset Selection Probabilities With Application to Cox Regression," *Biometrika*, 81, 607–611.

Diamond, I. D., McDonald, J. W., and Shah, I. H. (1986), "Proportional Hazards Models for Current Status Data: Application to the Study of Differentials in Age at Weaning in Pakistan," *Demography*, 23, 607–620.

Finkelstein, D. M. (1986), "A Proportional Hazards Model for Interval-Censored Failure Time Data," *Biometrics*, 42, 845–854.

Grummer-Strawn, L. M. (1993), "Regression Analysis of Current-Status Data: An Application to Breast-Feeding," *Journal of the American Statistical Association*, 88, 758–765.

Hoff, R. (1994), "Preparation for HIV Vaccine Trails: Moving From Baseline Studies to Efficacy Trials," *AIDS Research and Human Retroviruses*, 10, supp. 2, S191–S193.

Huang, J. (1996), "Efficient Estimation for the Proportional Type 1 Interval Censored Data," *The Annals of Statistics*, 24, 540–568.

Kalbfleisch, J. D., and Prentice, R. L. (1972), Discussion of "Regression Models and Life Tables" by D. R. Cox, *Journal of the Royal Statistical Society*, Ser. B, 26, 215–216.

——— (1973), "Marginal Likelihoods Based on Cox's Regression and Life Model," *Biometrika*, 60, 267–278.

——— (1980), *The Statistical Analysis of Failure Time Data*, New York: Wiley.

Kooperberg, C., and Clarkson, D. B. (1997), "Hazard Regression With Interval Censored Data," *Biometrics*, 53, 1485–1494.

Lin, D. Y., and Wei, L. J. (1989), "The Robust Inference for the Cox Proportional Hazards Model," *Journal of the American Statistical Association*, 89, 1074–1078.

Peto, R. (1972), Discussion of "Regression Models and Life Tables" by D. R. Cox, *Journal of the Royal Statistical Society*, Series B, 26, 205–207.

Reid, N., and Crépeau, H. (1985), "Influence Functions for the Proportional Hazards Regression," *Biometrika*, 72, 1–9.

Ruppert, D., Reisch, R. L., Deriso, R. B., and Carroll, R. J. (1984), "Optimization Using Stochastic Approximation and Monte Carlo Simulation (With Application to Harvesting of Atlantic Menhaden)," *Biometrics*, 40, 535–545.

Satten, G. A. (1996), "Rank-Based Inference in the Proportional Hazards Model for Interval Censored Data," *Biometrika*, 83, 355–370.

Sinha, D., Tanner, M. A., and Hall, W. J. (1994), "Maximization of the Marginal Likelihood of Grouped Survival Data," *Biometrika*, 81, 53–60.