Proportional hazards regression with interval censored data using an inverse probability weight

Glenn Heller

Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering
Cancer Center, 307 East 63 St, New York, NY 10065, U.S.A.

email address: hellerg@mskcc.org

Telephone number: 646-735-8112

Fax number: 646-735-0010

Abstract The prevalence of interval censored data is increasing in medical studies due to the growing use of biomarkers to define a disease progression endpoint. Interval censoring results from periodic monitoring of the progression status. For example, disease progression is established in the interval between the clinic visit where progression is recorded and the prior clinic visit where there was no evidence of disease progression. A methodology is proposed for estimation and inference on the regression coefficients in the Cox proportional hazards model with interval censored data. The methodology is based on estimating equations and uses an inverse probability weight to select event time pairs where the ordering is unambiguous. Simulations are performed to examine the finite sample properties of the estimate and a colon cancer data set is used to demonstrate its performance relative to the conventional partial likelihood estimate that ignores the interval censoring.

Keywords Estimating equation; Interval censoring; Proportional hazards; Selection probability

1 Introduction

In medical studies, radiographic scans and blood based biomarkers are increasingly being accepted as clinical endpoints in measuring the time to progression. In contrast to survival time, where patient death is the single determining point, disease progression is established in an interval between the clinic visit where progression is recorded and the prior clinic visit where there was no evidence of disease progression. To estimate the relationship between progression free survival time and covariates, it is common to employ the proportional hazards model, using a partial likelihood analysis based on the recorded progression time. However, if the interval width between clinic visits is significant relative to the variability of the true (but unobserved) progression times, then the ranks of the recorded progression times may be substantially different from the ranks of the true progression times, resulting in an inaccurate partial likelihood estimate of the proportional hazards regression coefficients.

Incorporation of interval censoring into the proportional hazards model does not enable canceling of the baseline hazard function, and as a result, estimation of the regression coefficients and the derivation of its asymptotic properties have proven challenging. Full likelihood approaches that require estimation of the baseline hazard to estimate the regression coefficient were developed by partitioning the time axis based on the endpoints of the event time intervals (Finkelstein 1986); employment of an EM algorithm based on piecewise constant event times (Goetghebeur and Ryan 2000); and use of a local likelihood to jointly estimate the regression coefficient and the baseline hazard function (Betensky et al. 2002). Although these methods do not require a global parametric distribution for the survival time, the need to estimate the baseline hazard offsets a key benefit of the semiparametric proportional hazards framework.

Monte Carlo methods have been explored to avoid baseline hazard estimation in the Cox model with interval censored data. Pan (2000) uses a multiple imputation procedure to fill-in failure times for the interval censored events and then applies the standard partial likelihood analysis. Satten (1996) considers a marginal likelihood approach, using Gibbs sampling to generate possible event time ranks. Simulations are used to demonstrate the small sample properties of the regression estimates, but the distribution theory for the estimated regression coefficients has not been developed. Satten, Datta, and Williamson (1998) derive an asymptotic distribution for the regression estimates, but require a parametric specification of the baseline hazard to impute failure times.

Estimating equations have been used to avoid estimation of the baseline hazard function. Zhang et al. (2005) proposed a solution under the constraint that the covariate in the proportional hazards model are discrete. In addition, due to the complexity of the regression estimate, they rely on a heuristic derivation of its asymptotic distribution and suggest a bootstrap estimate for the asymptotic variance (Zhang 2009). An excellent survey of statistical approaches applied to interval censored data is found in Sun (2006).

Software for the interval censored proportional hazards model is sparse (Gomez et al. 2009). In R, there exists one package (Henschel, Heiss, and Mansmann 2009), derived from an algorithm based on the joint maximization of the baseline survival function and the regression coefficient (Pan 1999). However, under the standard assumption that the baseline survival function S_0 is piecewise constant, the number of parameters grows with the sample size and the asymptotic distribution of the coefficient estimate $\hat{\boldsymbol{\beta}}$ is unclear. Huang and Wellner (1997) consider a profile likelihood approach $L(\hat{S}_{0\boldsymbol{\beta}}, \boldsymbol{\beta})$ to estimate the var$(\hat{\boldsymbol{\beta}})$, but note that the high dimensional inverse Hessian required for the variance estimate may lead to computational issues. In addition, they note that it remains uncertain whether $\hat{S}_{0\boldsymbol{\beta}}$ is a smooth function of $\boldsymbol{\beta}$, which would argue against a bootstrap estimate of var$(\hat{\boldsymbol{\beta}})$.

Thus to date, although endpoints such as progression free survival that produce interval censored data are increasing, the application of these approaches is not routinely incorporated into a Cox model analysis. As noted, the reasons include: either concurrent estimation of the infinite dimensional baseline hazard, computational difficulty, or the lack of an asymptotic distribution for the regression coefficient estimate.

In this paper, an alternative methodology is proposed to estimate the proportional hazards regression coefficient and to derive the asymptotic distribution of this estimate when interval censoring is present. The method does not require estimation of the baseline hazard and uses standard estimating equation techniques to produce the estimate and its asymptotic distribution. Sections 2 and 3 develop the methodology for estimation and inference. In Section 4, simulations are performed to examine the finite sample adequacy of the parameter estimate and coverage based on asymptotic confidence intervals. An analysis of colon cancer data is undertaken in Section 5 and Section 6 contains concluding remarks.

2 An estimating equation under proportional hazards

The proportional hazards relationship between the event time T and covariate vector

$\boldsymbol{X}$ is specified by

$$h(t|\boldsymbol{x}) = h_0(t)\exp(\boldsymbol{x}^T\boldsymbol{\beta}_1), \qquad (1)$$

where $h(t|\boldsymbol{x})$ represents the subject-specific conditional hazard function, $h_0(t)$ is an unspecified baseline hazard function, common for all subjects, and $\exp(\boldsymbol{x}^T\boldsymbol{\beta}_1)$ is the relative risk for an individual with k-dimensional covariate vector $\boldsymbol{x}$. As a result, the proportional hazards specification contains a finite dimensional parameter $\boldsymbol{\beta}_1$ and an infinite dimensional parameter $h_0(t)$.

With right censored data, each individual is associated with an event time, censoring time, and covariate vector $\{T_i, C_i, \boldsymbol{X}_i\}$. It is assumed that the event time and censoring time are independent conditional on the covariate vector. The minimum time and censoring indicator are observed for each subject and are denoted by

$$Y_i = \min(T_i, C_i) \qquad d_i^* = I(T_i \leq C_i) \qquad i = 1, \ldots, n.$$

Estimation and inference for $\boldsymbol{\beta}_1^0$, the true value of $\boldsymbol{\beta}_1$, is accomplished through the partial likelihood function (Cox, 1975). The score equation, derived from the partial likelihood, is based on the ranks of the observed times and does not contain the infinite dimensional parameter $h_0(t)$.

With interval censored data, however, the precise event times are not observed, and thus the ranks of these event times are unknown. Instead, the event time T_i for the i^{th} subject is known to lie in an interval $[L_i, R_i]$. For example, when scheduled clinic visits are used to assess disease progression, R_i represents the time from the start of treatment to the visit when disease progression is determined and L_i represents the time to the previous clinic visit. If the event has not occurred at the time of analysis, the event time is subject to right censoring. The notation used to indicate whether right censoring occurs is $d_i = I[R_i < C_i]$. The event time interval for a right censored subject is $[L_i, +\infty]$, where L_i represents the follow up time to the last negative recording. If the failure time is known precisely, when for example a death has occurred, the interval may be denoted as $[R_i, R_i]$. Finally, it is assumed that conditional on the covariate vector, the scheduling times are independent of the event time. This conditional independence assumption is common with interval censored data.

An alternative specification of the proportional hazards model is

$$m(t_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}_1^0 + \epsilon_i$$

where ϵ_i are independent, identically distributed, standard extreme value random variables, and m is an unknown, but monotone function of the survival times. This specification shows that the proportional hazards model is a member of the linear transformation family (Dabrowska and Doksum 1988; Cheng, Wei, and Ying 1995). Under the proportional hazards specification, for any pair of observations $\{(T_i, \boldsymbol{X}_i), (T_j, \boldsymbol{X}_j)\}$,

$$\Pr(T_i > T_j | \boldsymbol{X}_i = \boldsymbol{x}_i, \boldsymbol{X}_j = \boldsymbol{x}_j) = \frac{\exp[\boldsymbol{x}_{ji}^T \boldsymbol{\beta}_1^0]}{1 + \exp[\boldsymbol{x}_{ji}^T \boldsymbol{\beta}_1^0]}, \qquad (2)$$

where $\boldsymbol{x}_{ji} = \boldsymbol{x}_j - \boldsymbol{x}_i$. This result suggests an alternative estimating equation approach for inference on $\boldsymbol{\beta}_1^0$. Using the notation $S_{ij} = I[T_i > T_j]$, with no censoring, an unbiased estimating equation is

$$\sum_i \sum_{j \neq i} \{S_{ij} - E[S_{ij} | \boldsymbol{x}_i, \boldsymbol{x}_j]\} = 0.$$

An additional component is needed to account for the finite follow up period typical with survival time data. For an event time pair ordering to be observed, the minimum of the pair must be less than the maximum follow up time of the study, τ. Denoting the bounded ordering of the event time pair as $S_{ij}^{\tau} = I[T_i > T_j, T_j < \tau]$, then under proportional hazards (Fine, Ying, and Wei 1998),

$$E[S_{ij}^{\tau}|\boldsymbol{x}_i, \boldsymbol{x}_j] = \frac{\exp[\boldsymbol{x}_{ji}^T \boldsymbol{\beta}_1^0]}{1 + \exp[\boldsymbol{x}_{ji}^T \boldsymbol{\beta}_1^0]} \left\{1 - \exp\left(-\beta_2^0 \exp[\boldsymbol{x}_i^T \boldsymbol{\beta}_1^0] - \beta_2^0 \exp[\boldsymbol{x}_j^T \boldsymbol{\beta}_1^0]\right)\right\}$$

where $\beta_2^0 = \int_{t=0}^{\tau} h_0(t)$. Thus, accounting for the finite follow up period requires one additional parameter β_2 and the $(k+1)$ parameters are represented as $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \beta_2)^T$.

3 Inverse probability weighted estimating equation

With interval censoring, the ordering of the event time pairs (S_{ij}^{τ}) may not be observed. A sufficient condition for observing $\{T_i > T_j, T_j < \tau\}$ is $\{L_i > R_j, R_j < C_j\}$. This condition or its complementary event is indicated by

$$\Delta_{ij} = I[R_j < C_j]I[L_i > R_j] + I[R_i < C_i]I[L_j > R_i].$$

Li and Pu (2003) considered this condition of nonoverlapping event time interval pairs to estimate $\boldsymbol{\beta}$. Their proposal, however, was limited to the accelerated failure time model with a single covariate and required a strong independence assumption between the covariate and the assessment schedule.

A weighted unbiased estimating equation within the proportional hazards framework, based on the selected event time pairs is

$$\sum_i \sum_{j \neq i} \frac{\Delta_{ij}}{\pi^*(s_{ij}^{\tau}, \boldsymbol{x}_i, \boldsymbol{x}_j)} W_{ij}(\boldsymbol{\beta}) \left\{S_{ij}^{\tau} - E[S_{ij}^{\tau}|\boldsymbol{x}_i, \boldsymbol{x}_j]\right\} = 0, \tag{3}$$

where $\pi^*(s_{ij}^{\tau}, \boldsymbol{x}_i, \boldsymbol{x}_j) = \Pr[\Delta_{ij} = 1|S_{ij}^{\tau} = s_{ij}^{\tau}, \boldsymbol{X}_i = \boldsymbol{x}_i, \boldsymbol{X}_j = \boldsymbol{x}_j]$ is the selection probability and $W_{ij}(\boldsymbol{\beta}) = \partial E[S_{ij}^{\tau}|\boldsymbol{x}_i, \boldsymbol{x}_j]/\partial \boldsymbol{\beta}$.

If π^* is known, then inference for $\boldsymbol{\beta}$ can be accomplished through the estimating equation (3). If π^* is unknown, but the unobserved S_{ij}^{τ} are missing at random, i.e. $\pi^*(s_{ij}^{\tau}, \boldsymbol{x}_i, \boldsymbol{x}_j) = \pi^*(\boldsymbol{x}_i, \boldsymbol{x}_j)$, then either setting $\pi^* = 1$ or using a working model for π^* would produce an unbiased estimating equation. The unobserved ordered pairs, however, are not missing at random. Specifically, note that $I(L_i > R_j) = I(L_i > R_j)I(T_i > T_j)$ and

$$\Pr[\Delta_{ij} = 1 | S_{ij}^{\tau} = s_{ij}^{\tau}, \boldsymbol{X}_i = \boldsymbol{x}_i, \boldsymbol{X}_j = \boldsymbol{x}_j] =$$
$$s_{ij}^{\tau}\Pr[\Delta_{ij} = 1 | S_{ij}^{\tau} = 1, \boldsymbol{X}_i = \boldsymbol{x}_i, \boldsymbol{X}_j = \boldsymbol{x}_j] + s_{ji}^{\tau}\Pr[\Delta_{ij} = 1 | S_{ji}^{\tau} = 1, \boldsymbol{X}_i = \boldsymbol{x}_i, \boldsymbol{X}_j = \boldsymbol{x}_j].$$

An instructive approach for the missing not at random case is to incorporate an observable auxillary variable that essentially captures the information in S_{ij}^{τ} for predicting Δ_{ij} (Ibrahim, Lipsitz, and Horton 2001). Since S_{ij}^{τ} informs the order between (T_i, T_j), an observable auxiliary variable that orders the event time interval pair $[L_i, R_i]$ and $[L_j, R_j]$ is

$$A_{ij} = I[R_j < C_j]I[L_i \geq L_j]I[R_i > R_j].$$

The proposal is to use the observable A_{ij} in place of S_{ij}^{τ} to model the selection probability; the inverse probability weight component in the estimating equation (3). This replacement is based on an estimating equation modified to include the auxillary variables and still retain its mean zero property

$$\sum_i \sum_{j \neq i} \frac{\Delta_{ij} W_{ij}(\boldsymbol{\beta}) \left\{ S_{ij}^{\tau} - E[S_{ij}^{\tau} | \boldsymbol{x}_i, \boldsymbol{x}_j] \right\}}{\Pr[\Delta_{ij} = 1 | S_{ij}^{\tau} = s_{ij}^{\tau}, A_{ij} = a_{ij}, A_{ji} = a_{ji}, \boldsymbol{X}_i = \boldsymbol{x}_i, \boldsymbol{X}_j = \boldsymbol{x}_j]} = 0, \qquad (4)$$

where $0/0$ is defined as zero for any term in the summand and throughout the paper. To create an observable estimating equation, the selection probability in (4) is replaced by $\Pr[\Delta_{ij} = 1|A_{ij} = a_{ij}, A_{ji} = a_{ji}, \boldsymbol{X}_i = \boldsymbol{x}_i, \boldsymbol{X}_j = \boldsymbol{x}_j]$. If $a_{ij} + a_{ji} = 0$, this replacement is trivial since $\Delta_{ij} = 0$ with probability 1. If $a_{ij} + a_{ji} = 1$, a critical assumption for the viability of this substitution is that conditional on $(A_{ij}, A_{ji}, \boldsymbol{X}_i, \boldsymbol{X}_j)$, S^{τ}_{ij} is ignorable for the prediction of Δ_{ij}. This ignorability assumption essentially transforms the problem into a missing at random framework. Although it is likely that this condition is only an approximation, the simulations in this paper suggest that the approximation is reasonable.

To implement this alternative specification of the selection probability, data pairs are eliminated when right censoring masks the potential ordering of the intervals $[L_i, R_i]$, $[L_j, R_j]$. The indication that right censoring obscures the ordering is denoted by

$$1 - \omega_{ij} = (1 - d_i)I[L_i \leq L_j] + (1 - d_j)I[L_j < L_i]$$

and the selection probability conditional on the complementary space where $(A_{ij}, A_{ji}, \boldsymbol{X}_i, \boldsymbol{X}_j)$ is informative for Δ_{ij} is

$$\pi(a_{ij}, a_{ji}, \boldsymbol{x}_i, \boldsymbol{x}_j) \equiv \Pr[\Delta_{ij} = 1|A_{ij} = a_{ij}, A_{ji} = a_{ji}, \boldsymbol{X}_i = \boldsymbol{x}_i, \boldsymbol{X}_j = \boldsymbol{x}_j, \omega_{ij} = 1].$$

Incorporation of the selection probability enables the use of the observed event time pairs to create an asymptotic mean zero estimating equation in the presence of interval censored data

$$\sum_i \sum_{j \neq i} \frac{\Delta_{ij}}{\pi(a_{ij}, a_{ji}, \boldsymbol{x}_i, \boldsymbol{x}_j)} W_{ij}(\boldsymbol{\beta}) \left\{ S^{\tau}_{ij} - E[S^{\tau}_{ij}|\boldsymbol{x}_i, \boldsymbol{x}_j] \right\} = 0. \tag{5}$$

Use of this estimating equation requires an estimate for the selection probability component $\pi(a_{ij}, a_{ji}, \boldsymbol{x}_i, \boldsymbol{x}_j)$. A common approach is to apply a logistic regression model, which is specified as

$$\pi(a_{ij}, a_{ji}, \boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\gamma}) = a_{ij} \frac{\exp[\gamma_0 + \boldsymbol{x}_{ji}^T \boldsymbol{\gamma}_1]}{1 + \exp[\gamma_0 + \boldsymbol{x}_{ji}^T \boldsymbol{\gamma}_1]} + a_{ji} \frac{\exp[\gamma_0 + \boldsymbol{x}_{ij}^T \boldsymbol{\gamma}_1]}{1 + \exp[\gamma_0 + \boldsymbol{x}_{ij}^T \boldsymbol{\gamma}_1]}. \tag{6}$$

Estimation of the Cox regression coefficient is accomplished through two estimating equations. First, $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_0, \hat{\boldsymbol{\gamma}}_1^T)^T$ is the solution to the estimating equation

$$Q_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}) = n^{-3/2} \sum_i \sum_{j \neq i} D_{ij}(\boldsymbol{\gamma}) \left\{ \frac{\Delta_{ij}}{\pi(a_{ij}, a_{ji}, \boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\gamma})} - 1 \right\} = 0, \tag{7}$$

where $D_{ij}(\boldsymbol{\gamma}) = \partial \pi(a_{ij}, a_{ji}, \boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\gamma}) / \partial \boldsymbol{\gamma}$. The estimated selection probabilities are used to estimate $\boldsymbol{\beta}$ from the profile estimating equation

$$Q_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \hat{\boldsymbol{\gamma}}) = n^{-3/2} \sum_i \sum_{j \neq i} \frac{\Delta_{ij}}{\pi(a_{ij}, a_{ji}, \boldsymbol{x}_i, \boldsymbol{x}_j; \hat{\boldsymbol{\gamma}})} W_{ij}(\boldsymbol{\beta}) \left\{ S_{ij}^{\tau} - E[S_{ij}^{\tau} | \boldsymbol{x}_i, \boldsymbol{x}_j] \right\} = 0. \tag{8}$$

The estimating equations (7) and (8), as a function of the true parameter values, are U-statistics of degree 2. As a result, the following theorem summarizes the asymptotic distribution of the parameter estimates determined by the estimating equations.

Theorem: Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)^T$ and $Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ be the $2k+2$ vector of estimating functions defined in (7) and (8). Under the following conditions:

(C1) The proportional hazards assumption (1) between the unobserved event time T and the covariate vector $\boldsymbol{X}$ is valid.

(C2) The selection probability in (4) is approximated by $\pi(a_{ij}, a_{ji}, \boldsymbol{x}_i, \boldsymbol{x}_j)$, creating a missing at random structure.

(C3) The logistic regression model (6) is the proper specification for the selection probability $\pi(a_{ij}, a_{ji}, \boldsymbol{x}_i, \boldsymbol{x}_j)$.

Then $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges in distribution to $N(0, U^{-1}VU^{-1})$, where $U = \lim_{n\to\infty} \mathrm{E}\{n^{-1/2} \partial Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}\}$ and $V = \lim_{n\to\infty} n^{-1} \mathrm{var}\{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\}$. In addition, the asymptotic variance of $\hat{\boldsymbol{\beta}}_1$ is the upper left $k \times k$ submatrix of $U^{-1}VU^{-1}$. The proof is standard and follows from a one-term Taylor expansion of $Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$. The estimate of the matrix U is computed by replacing the expectation with the sample average and the parameters with the parameter estimates. The variance-covariance matrix V is obtained through U-statistic theory. Letting q_{ijl} denote the (i, j) element in the summand of the l^{th} component of $Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$, the (l, m) element of V is

$$n^{-3} \sum_i \sum_j \sum_{k \neq j} (q_{ijl} + q_{jil})(q_{ikm} + q_{kim}).$$

4 Simulations

A set of simulations were generated to assess the performance of the interval censored estimating equation estimates $\hat{\beta}_{ee}$. A comparable set of simulations were produced using the conventional partial likelihood estimate $\hat{\beta}_{pl}$, choosing the time of the first recorded positive event as the failure time, and an estimate of $\boldsymbol{\beta}$ using Pan's joint maximization algorithm (Pan 1999). Pan's method, however, does not produce an asymptotic variance for the coefficient estimate. The underlying event time data were generated to satisfy the proportional hazards specification

$$T_i = \exp[4 + x_i \beta] \times \epsilon_i$$

where $\beta = 1$, x_i was generated as a standard normal random variable, and ϵ_i was distributed as a Weibull random variable with shape parameter λ and scale parameter equal to 1. The shape parameter λ was varied as $\{0.693, 1.1, 1.386, 1.609\}$ and the Cox coefficient is $\beta_1 = -\lambda\beta$. This choice of shape parameters results in odds parameters equal to $\{2, 3, 4, 5\}$, based on the odds concordance interpretation of the

Cox coefficient

$$\frac{\Pr[T_j > T_i|\boldsymbol{x}_i, \boldsymbol{x}_j]}{\Pr[T_i > T_j|\boldsymbol{x}_i, \boldsymbol{x}_j]} = \exp[-\boldsymbol{x}_{ji}\boldsymbol{\beta}_1].$$

This sequence of shape parameters translates into event times with decreasing variability.

The monitoring schedule was developed under the clinical scenario that scans to detect disease progression were scheduled either every 12 months or every 24 months. Thus, for subject i, the simulated 12 month scan schedule was

$$\sigma_{k,i} = 12k + U_{ik}(-2, 2) \qquad k = 1, \ldots, K$$

where the $U_{ik}(-2, 2)$ represent independent uniform random variables, with support between -2 and 2, and K indicates the maximum number of scans. The uniform random variable is included to signify the scenario that a subject would arrive for the scan within a 2 month neighborhood of the scheduled visit. To examine the effect of the length of follow up on the regression estimates, the number of scans (K) was varied as $\{5, 10, 25\}$ for scans scheduled every 12 months and $\{3, 6, 15\}$ for scans scheduled every 24 months. The left and right endpoints for each subject's event time interval were computed as

$$\begin{aligned}(L_i, R_i) \quad = \quad & (0, \sigma_{1i})I[T_i < \sigma_{1,i}] + \sum_{k=2}^{K}\{(\sigma_{k-1,i}, \sigma_{k,i})I[\sigma_{k-1,i} < T_i < \sigma_{k,i}]\} + \\ & (\sigma_{K,i}, M)I[T_i > \sigma_{K,i}]\end{aligned}$$

In the simulations, the right endpoint of a subject whose underlying event time was greater than the maximum follow up time was assigned a very large value (M). In addition, a uniform censoring random variable was generated as $C_i \sim U(0, \tau)$ to simulate the clinical trial scenario of subjects entering the study at different time points and lost to follow up. If C was less than the scan time where the progression was observed, then the follow-up time was censored at the scan time just prior to C. The maximum follow up time of the study τ is equal to the scan interval width times the maximum number of scans scheduled. The sample size for each simulation was 100 and the results of each simulation were based on 5,000 replications.

The estimating equation removes data pairs that do not meet the selection criteria. To assess the potential loss of information with this procedure, the simulation estimated root mean squared error of $\hat{\beta}_{ee}$ was compared to the simulation estimated root mean squared error of the partial likelihood estimate and Pan's regression coefficient estimate. Pan's approach, based on a full likelihood, is used to gauge the efficiency of the proposed estimate.

The results for $\hat{\beta}_{ee}$ are accurate over the range of simulations examined (Table 1). The bias was small. The average asymptotic standard error of the $\hat{\beta}_{ee}$ provided a good approximation to its simulation standard error. The 95% empirical coverage was uniformly good based on the confidence interval $\hat{\beta} \pm 1.96 \times \mathrm{se}(\hat{\beta})$. In contrast, the conventional partial likelihood approach produced a bias in $\hat{\beta}_{pl}$ and poorer coverage rates, which were magnified when the scan interval width was larger and the variability of the event times was smaller (Table 2).

The quality of the estimate $\hat{\beta}_{ee}$ decreased as the scan interval width increased and the variability of the underlying event time decreased. These parameter settings, along with higher censoring rates, tended to reduce the number of unambiguously ordered event time pairs. A comparison of the root mean squared error for the estimating equation estimator and the partial likelihood estimator shows that when the scan interval width is 12, the root mean squared error of the partial likelihood estimator is always smaller. The results are mixed when the scan interval is 24, where $\hat{\beta}_{ee}$ has a smaller root mean square error when the censoring is either 25% or 45% and the variability of the underlying event times is low. These results suggest an additional weight incorporated into the estimating equations, targeted to reduce the variability of $\hat{\beta}_{ee}$, would prove useful. Surprisingly, the root mean squared error for Pan's full likelihood does not dominate. The Pan estimate incurs significant bias when the event time variability is small.

5 Colon cancer data analysis

Ninety patients with locally advanced colorectal cancer were treated surgically at Memorial Sloan-Kettering Cancer Center. Patients were monitored for their time to recurrence or death. Tumors of the colon produce the protein Carcinoembryonic Antigen (CEA). The prognostic value of CEA for predicting the risk of recurrence in this population is unclear, since CEA may also be elevated in response to other diseases.

An analysis was undertaken to explore the prognostic significance of the baseline CEA measure, taken just after surgery, in assessing the risk of recurrence. The purpose of the analysis was to determine whether the clinician should use CEA in the assessment of peri-surgical treatment options. To determine recurrence, patients underwent a CT scan at six months and one year after surgery, and were subsequently scheduled for yearly scans. Although follow up scans were intended for the life of the patient, patients with no cancer related health problems typically ceased returning to the clinic at some point in their follow up. As a result of the scan schedule, the event tumor recurrence is interval censored. An analysis based on the partial likelihood estimate, using the time of the first recorded positive recurrence on the CT scan as the event, was compared to the estimating equation estimate to explore the differences in the two approaches. Approximately 50% of the patients recurred and no patient died prior to their recurrence. The estimated probability of remaining alive and recurrence free at 10 years was 0.36. This calculation was based on a Kaplan-Meier estimate that accounts for interval censoring (Wellner and Zhan 1997). The relationship between baseline CEA and the time to recurrence was first summarized through the partial likelihood estimate. CEA values ranged from $(0.0, 52.6)$, with the median value equal to 3.7. The distribution of CEA was right skewed and a square root transformation was applied. The estimated coefficient was $\hat{\beta}_{pl} = 0.171$ and the estimated $\text{se}(\hat{\beta}_{pl}) = 0.078$, suggesting a positive relationship between baseline CEA and the risk of recurrence.

Accounting for the interval censoring and applying the estimating equation approach produced $\hat{\beta}_{ee} = 0.235$ and the estimated $\text{se}(\hat{\beta}_{ee}) = 0.114$. The results of the partial likelihood and the estimating equation methods are summarized in Table 3. Interestingly, both the estimate and its standard error are greater than the partial likelihood estimates. One explanation is that the estimating equation approach uses only data pairs where there is a definitive event time ordering. In contrast, the partial likelihood approach ignores the uncertainty in the event time ordering and uses all the events in the estimation process. This can produce an event time ordering that is partially dictated by the visitation schedule to the clinic, resulting in an attenuation of the partial likelihood estimate $\hat{\beta}_{pl}$. In addition, by implicitly assuming more information in the event time ordering than exists in the presence of interval censored data, the standard error of $\hat{\beta}_{pl}$ is expected to be smaller, giving a false sense of confidence as to the location of the true regression coefficients.

6 Discussion

Interval censored data methods applied to the proportional hazards model have not gained widespread acceptance. Methodology that has incorporated interval censoring requires either estimation of the baseline hazard or a computationally intensive approach. Due to this inadequacy, interval censoring is often ignored, and the right endpoint of the event time interval is used to compute the partial likelihood estimate. The result is a biased estimate, with the bias increasing as a function of the interval censoring width combined with the precision of the underlying event times. In this paper, the proportional hazards specification is used to create an estimating equation that incorporates an inverse selection probability weight. If the selection probability model and the proportional hazards model are properly specified, then a set of unbiased estimating equations can be formed under a surrogacy condition. The asymptotic distribution of the estimated regression coefficients follows directly from the estimating equations. This proposal, which produces a Cox model estimate that is computationally straightforward and has an estimable asymptotic distribution, is timely due to the increasing use of clinical endpoints that are subject to interval censoring.

References

Betensky RA, Lindsey JC, Ryan LM, and Wand MP (2002) A local likelihood proportional hazards model for interval censored data. Statistics in Medicine 21:263-275

Cheng SC, Wei LJ, Ying Z (1995) Analysis of transformation models with censored data. Biometrika 82:835-845

Cox DR (1975) Partial likelihood. Biometrika 62:269-276

Dabrowska DM, Doksum KA (1988) Partial likelihood in transformation models with censored data. Scandanavian Journal of Statistics 15:1-23

Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman and Hall, New York.

Fine JP, Ying Z, Wei LJ (1998) On the linear transformation model for censored data. Biometrika 85:980-986

Finkelstein DM (1986) A proportional hazards model for interval-censored failure time data. Biometrics 42:845-854

Goetghebeur E, Ryan, L (2000) Semiparametric regression analysis of interval-censored data. Biometrics 56:1139-1144

Gomez G, Calle ML, Oller R, Langohr K (2009) Tutorial on methods for interval-censored data and their implementation in R. Statistical Modelling 9:259-297

Henschel V, Heiss C, Mansmann U (2009) intcox: Compendium to apply the iterative convex minorant algorithm to interval censored event data. http://127.0.0.1:21646/library/intcox/vignettes/intcox.pdf

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

Ibrahim JG, Lipsitz SR, Horton N (2001) Using auxiliary data for parameter estimation with non-ignorably missing outcomes. Applied Statistics 50:361-373

Li L, Pu Z (2003) Rank estimation of log-linear regression with interval-censored data. Lifetime Data Analysis 9:57-70.

Pan W (1999) Extending the iterative convex minorant to the Cox model for interval censored data. Journal of Computational and Graphical Statistics 8:109-120

Pan W (2000) A multiple imputation approach to Cox regression with interval-censored data. Biometrics 56:199-203

Satten GA (1996) Rank-based inference in the proportional hazards model with interval-censored data. Biometrika 83:355-370

Satten GA, Datta S, Williamson JM (1998) Inference based on imputed failure times for the proportional hazards model with interval-censored data. Journal of the American Statistical Association 93:318-327

Sun J (2006) The statistical analysis of interval-censored failure time data. Springer, New York

Wellner JA, Zhan Y (1997) A hybrid algorithm for computation of the nonparametric maximum likelihood estimator from censored data. Journal of the American Statistical Association 92:945-959

Zhang Z, Sun L, Zhao X, Sun J (2005) Regression analysis of interval censored failure time data with linear transformation models. The Canadian Journal of

Statistics 33:61-70

Zhang, Z (2009) Linear transformation models for interval-censored data: prediction of survival probability and model checking. Statistical Modelling 9:321-343

Table 1: Simulation results based on the inverse probability weighted estimating equation. The columns in the table represent: Scan, scan schedule; Shape, variability of the underlying event times; PRC, percent right censored; Bias, bias of the parameter estimator; SSE, standard deviation of the simulation estimates; ASE, average estimated standard error; CP, empirical coverage probability; RMSE, root mean square error of the estimator.

Scan	Shape	PRC	Bias	SSE	ASE	CP	RMSE
12	0.693	0.25	-0.004	0.161	0.160	0.949	0.161
		0.45	-0.007	0.193	0.183	0.943	0.193
		0.65	-0.009	0.230	0.221	0.947	0.230
	1.110	0.25	-0.001	0.191	0.200	0.952	0.191
		0.45	-0.013	0.232	0.230	0.949	0.232
		0.65	-0.054	0.318	0.296	0.939	0.322
	1.386	0.25	0.000	0.218	0.301	0.950	0.218
		0.45	-0.021	0.271	0.279	0.943	0.272
		0.65	-0.079	0.398	0.364	0.943	0.406
	1.609	0.25	0.005	0.237	0.303	0.949	0.237
		0.45	-0.023	0.301	0.339	0.946	0.302
		0.65	-0.126	0.475	0.425	0.938	0.491
24	0.693	0.25	0.030	0.165	0.163	0.939	0.168
		0.45	0.028	0.203	0.195	0.929	0.205
		0.65	0.015	0.273	0.258	0.932	0.273
	1.110	0.25	0.051	0.197	0.202	0.921	0.203
		0.45	0.021	0.263	0.254	0.930	0.264
		0.65	-0.049	0.398	0.364	0.934	0.401
	1.386	0.25	0.084	0.226	0.349	0.908	0.241
		0.45	0.021	0.318	0.318	0.920	0.319
		0.65	-0.086	0.513	0.589	0.927	0.520
	1.609	0.25	0.111	0.258	0.425	0.885	0.281
		0.45	0.015	0.372	0.447	0.920	0.372
		0.65	-0.156	0.686	0.727	0.929	0.704

Table 2: Simulation results from the partial likelihood estimate and the Pan estimate. The columns in the table represent: Scan, scan schedule; Shape, variability of the underlying event times; PRC, percent right censored; Bias, bias of the parameter estimator; SSE, standard deviation of the simulation estimates; ASE, average estimated standard error; CP, empirical coverage probability; RMSE, root mean square error of the estimator.

			PL estimate					Pan estimate		
Scan	Shape	PRC	Bias	SSE	ASE	CP	RMSE	Bias	SSE	RMSE
12	0.693	0.25	0.006	0.135	0.137	0.953	0.135	-0.013	0.136	0.137
		0.45	0.012	0.158	0.155	0.948	0.158	0.005	0.150	0.150
		0.65	0.046	0.182	0.183	0.947	0.188	0.034	0.178	0.181
	1.110	0.25	0.034	0.155	0.154	0.935	0.159	0.063	0.139	0.153
		0.45	0.044	0.179	0.177	0.931	0.184	0.084	0.151	0.173
		0.65	0.056	0.224	0.216	0.917	0.231	0.118	0.192	0.225
	1.386	0.25	0.060	0.178	0.171	0.908	0.188	0.157	0.145	0.214
		0.45	0.076	0.200	0.196	0.906	0.214	0.176	0.156	0.235
		0.65	0.102	0.257	0.242	0.885	0.277	0.229	0.194	0.300
	1.609	0.25	0.093	0.196	0.183	0.868	0.217	0.259	0.147	0.298
		0.45	0.113	0.224	0.211	0.861	0.251	0.273	0.157	0.315
		0.65	0.147	0.283	0.263	0.848	0.319	0.334	0.191	0.385
24	0.693	0.25	0.048	0.131	0.134	0.932	0.140	-0.011	0.141	0.141
		0.45	0.062	0.152	0.154	0.924	0.164	0.004	0.164	0.164
		0.65	0.106	0.190	0.191	0.890	0.218	0.006	0.209	0.209
	1.110	0.25	0.125	0.153	0.147	0.815	0.198	0.039	0.152	0.157
		0.45	0.155	0.173	0.168	0.797	0.232	0.045	0.179	0.185
		0.65	0.224	0.224	0.212	0.676	0.317	0.046	0.231	0.236
	1.386	0.25	0.215	0.177	0.157	0.645	0.278	0.114	0.158	0.195
		0.45	0.259	0.196	0.179	0.623	0.325	0.110	0.188	0.218
		0.65	0.346	0.252	0.228	0.564	0.428	0.131	0.234	0.268
	1.609	0.25	0.305	0.197	0.164	0.496	0.363	0.191	0.161	0.250
		0.45	0.349	0.221	0.189	0.499	0.413	0.177	0.193	0.262
		0.65	0.457	0.274	0.240	0.456	0.533	0.217	0.234	0.319

Table 3: Estimated coefficients and standard errors for the square root of CEA from the partial likelihood and estimating equations.

	sqrt(CEA)	
Method	Coef	SE
Partial likelihood	0.171	0.078
Estimating Equation	0.235	0.114