

Inference in the Cox
Proportional Hazards Model
with Missing Covariate Data

Marian Pugh¹, James Robins, Stuart Lipsitz, David Harrington

Department of Biostatistics
Harvard School of Public Health

July, 1993

¹This investigation was supported by grants CA57253 and CA39929 awarded by the National Cancer Institute, DHHS.

Abstract

The Cox proportional hazards regression model is often used by clinical trials investigators who wish to quantify the effects of prognostic factors on survival. When confronted with missing covariate data, analysts reduce either the data set or the model by deleting incompletely observed cases or covariates. The first approach can lead to biased estimates of treatment effects, the second to model misspecification.

A set of estimating equations is proposed which are re-weighted versions of the usual score equations in the Cox model restricted to complete cases. The subject-specific weights are proportional to the reciprocal probability of having complete data, which may depend on the event time or other nonmissing covariates, and can be estimated from the data using logistic or probit regression.

The asymptotic distribution of the estimates of relative risk is derived, allowing the construction of confidence intervals and test statistics for inference. A simulation study illustrates the large sample unbiasedness of the estimates in a simplified clinical trials setting.

1.1 Introduction

In clinical trials investigators often wish to quantify the effects of prognostic factors on survival, in order to adjust for imbalances in predictive covariates which may persist after randomization, to increase the statistical power of tests for treatment effects on survival, and to model the natural history of the disease. Covariate measurements may be missing for some individuals, either by accident or study design. Measurements which require invasive or painful collection procedures, complicated laboratory analysis, or time consuming coding or compilation are most likely to be incomplete. Common strategies that analysts employ when confronted with missing data include deletion of cases with missing covariates, or removal of covariates with missing observations from the model. The first approach can lead to biased estimates of covariate effects, and the second to model misspecification. In addition, the deletion of incomplete cases can result in loss of efficiency due to reduction in sample size.

In this paper we propose a method of inference for right-censored failure time data when there are missing covariates on some cases. Our methods provide both test statistics and parameter estimates for the effect of covariates in the proportional hazards regression model. The estimating equations used for parameter estimation and testing are intuitive, re-weighted versions of the usual score equations in the Cox model restricted to subjects with complete covariate information. The subject-specific weights are proportional to the reciprocal of the probability of having complete data. The use of weights permits valid inferences to be drawn from the complete case analysis even if the probability that a subject has complete data depends on its follow-up time. Our assumption about the missing data mechanism is closely related to Rubin's (1976) definition of missing at ran-

dom. The weights for each case in the modified score equations are generally not known, but can be estimated from the data using logistic, probit, or other binary regression model. We show that the estimates of relative risk obtained from the weighted score equation are consistent, asymptotically normal and unbiased, with both known and estimated weights.

Several methods for handling missing covariate data have been proposed for the regression problem where the response is a right-censored failure time, and covariates are dichotomous or categorical. Schemper and Smith (1990) discuss the situation of testing for treatment effects, while adjusting for dichotomous covariates, some of which may be missing. They suggest a probability imputation technique, where the missing covariate is replaced by the means of non-missing values, calculated and imputed separately for each treatment group. They note that treatment comparisons are invalid if the missingness depends on unobserved values of the incomplete covariate, and the incomplete covariate is associated with treatment.

Schluchter and Jackson (1989) consider a parametric model consisting of a log-linear model for the hazard, which is assumed to be piecewise constant over time, and a multinomial model for the probabilities in the contingency table defined by the categorical covariates. They present both a generalized EM algorithm and a Newton-Raphson algorithm for estimating the model parameters and obtaining their large sample standard errors. In addition to the missing at random assumption, their method requires an additional assumption that the censoring distribution does not depend on any incomplete covariates.

Prentice (1986) examines relative risk regression in the context of a case-cohort study design. Covariate histories are obtained only for individuals who fail, and for

a randomly selected cohort of censored individuals. This design reduces the burden of data collection for censored cases, but results in little loss of efficiency, because the latter depends primarily on failed cases. Prentice then proposes a “pseudo-score” function which is an estimate of the usual partial likelihood score based on the complete cohort information. Lin and Ying (1992) extend the pseudo-score approach of Prentice to the general missing data problem under the Cox regression model. Their approach allows for time-dependent covariates, but requires more restrictive assumptions than our approach. They assume that for each time t the conditional probability of a component of the case’s covariate vector being missing at t , given that the case is at risk, is independent of all information in the conditioning event except risk set membership.

In this paper, we discuss the bias that can result from a complete case analysis, review counting process notation, motivate our weighted estimating equations, and characterize the asymptotic distribution of our estimate of relative risk. This paper concludes with a small simulation study which provides an example of the effectiveness of using weights in reducing estimation bias, and a discussion of the relative merits of the methodology. A lengthy proof of the asymptotic results presented in this paper is given in the chapter titled “Proofs of the Asymptotic Properties of the Weighted Estimator”. This material supplements the more illustrative and heuristic approach of this paper, but is not essential to its development or understanding.

1.2 The Bias in Complete Case Analyses

In the Cox proportional hazards regression model (Cox 1972), the effect of a p -dimensional time-independent covariate \mathbf{Z} on a failure time variable T is modeled

through the conditional hazard function for T ,

$$\lim_{h \downarrow 0} \frac{P(t \leq T < t + h | T \geq t, \mathbf{Z})}{h} = \lambda(t | \mathbf{Z}) = \lambda_0(t) e^{\beta' \mathbf{Z}}.$$

Generally, the observation of T is censored by a variable U , so that the observable outcomes are $X = \min(T, U)$ and $\delta_i = I_{\{T \leq X\}}$ instead of T itself. When there are no missing covariates, the data used in a regression analysis based on n cases are $\{X_i, \delta_i, \mathbf{Z}_i : i = 1, \dots, n\}$. Inference is based on the partial likelihood for the data, which is, assuming no ties,

$$L_n(\beta) = \prod_{i=1}^n \left[\frac{e^{\beta' \mathbf{Z}_i}}{\sum_{j=1}^n I_{\{X_j \geq X_i\}} e^{\beta' \mathbf{Z}_j}} \right]^{\delta_i} \quad (1.1)$$

and on the resulting score equation

$$\mathbf{U}_n(\beta) = \frac{\partial}{\partial \beta} \log L_n(\beta) = 0.$$

Under the standard assumptions of uninformative and independent censoring, maximum partial likelihood estimators are asymptotically unbiased and normally distributed (Andersen and Gill, 1982).

When there are missing covariates, using only cases with complete covariate information (“complete cases”) can lead to biased inference. Define $R_i = 1$ when case i has no missing data and 0 otherwise, and let $\hat{\beta}_{comp}$ denote the maximizer of (1.1) when the product is only over the set of subjects with $R_i = 1$. Let \mathbf{Z}_{all} be the components of \mathbf{Z} which are observed with probability one. A sufficient condition for $\hat{\beta}_{comp}$ to be consistent for the true β_0 is

$$\lim_{h \downarrow 0} \frac{P(t \leq T < t + h | X \geq t, R = 1, \mathbf{Z})}{h} = \lim_{h \downarrow 0} \frac{P(t \leq T < t + h | T \geq t, \mathbf{Z})}{h}. \quad (1.2)$$

Equation (1.2) is the condition of independent censoring when the censoring process is generalized to also “censor” individuals with data missing at time 0. The right hand side of (1.2) is called the net hazard function.

A simple re-expression of the left-hand side of (1.2) provides insight into sufficient conditions for consistency of $\hat{\beta}_{comp}$. The left-hand side of (1.2) can be rewritten as

$$\lim_{h \downarrow 0} \frac{P(R = 1 | t \leq T < t + h, X \geq t, \mathbf{Z})}{P(R = 1 | X \geq t, \mathbf{Z})} \lim_{h \downarrow 0} \frac{1}{h} P(t \leq T < t + h | X \geq t, \mathbf{Z}). \quad (1.3)$$

The second factor is the crude hazard, and under the usual independent censoring assumption, is equal to the net hazard. Equation (1.2) implies that the first factor in (1.3) is one. However, $\hat{\beta}_{comp}$ will be consistent for β_0 under the weaker condition that the first factor in (1.3) is a function of t (for all t) alone since any such function will cancel from the partial likelihood.

The independence assumptions which permit unbiased estimation in the usual complete case analysis are quite restrictive but will hold in special cases. For example, if conditional on $\{T_i, I_{\{X_i > t\}}, \mathbf{Z}_i\}$, R_i depends only on $I_{\{X_i > t\}}$ then (1.3) holds. This implies that case cohort designs which collect only partial covariate information will still yield unbiased estimates, as long as the cases omitted are a random sample of cases from each risk set at each failure time. As a second example, consider the situation when all cases are observed to fail, or more generally, T is independent of the censoring time U conditional on R and \mathbf{Z} . We may then write (1.2) as $\lambda(t | R = 1, \mathbf{Z}) = \lambda(t | \mathbf{Z})$, which implies that $\hat{\beta}_{comp}$ is consistent provided that the observed hazard among the complete cases is equal to the net hazard.

In the next section, we review the counting process notation for the Cox score equation, in preparation for introducing a weighted score equation which yields unbiased estimates under less restrictive assumptions.

1.3 Counting Process Notation for the Cox Score

Consider the setting where there are no missing data. We may rewrite the score as a martingale by switching to counting process notation. The information in (X_i, δ_i) can be represented by processes $\{N_i(t), Y_i(t) : t \geq 0\}$. The process N_i takes value one if the i^{th} individual has been observed to fail at or before time t , and zero otherwise, and $Y_i(t)$ takes value one if the i^{th} case is at risk at time t and value zero otherwise. That is,

$$N_i(t) = I_{\{T \leq t, \delta=1\}}.$$

$$Y_i(t) = I_{\{X \geq t\}}.$$

The right-continuous filtration specifying the information accruing over time is $\{\mathcal{F}_t : t \geq 0\}$, where $\mathcal{F}_t = \sigma\{\mathbf{Z}_i, N_i(s), Y_i(s^+) : 0 \leq s \leq t, i = 1, \dots, n\}$. The expression for the score $\mathbf{U}_n(\boldsymbol{\beta})$ may be written as a stochastic integral in counting process notation,

$$\mathbf{U}_n(\boldsymbol{\beta}) = \sum_{i=1}^n \int_0^\infty \{\mathbf{Z}_i - \bar{\mathbf{Z}}(\boldsymbol{\beta}, u)\} dN_i(u),$$

where

$$\bar{\mathbf{Z}}(\boldsymbol{\beta}, u) = \frac{\sum_{j=1}^n \mathbf{Z}_j Y_j(u) e^{\boldsymbol{\beta}' \mathbf{Z}_j}}{\sum_{j=1}^n Y_j(u) e^{\boldsymbol{\beta}' \mathbf{Z}_j}}.$$

If $dM_i(s) \equiv dN_i(s) - e^{\boldsymbol{\beta}' \mathbf{Z}_i} Y_i(s) \lambda_0(s) ds$, it is easy to show that $\mathbf{U}_n(\boldsymbol{\beta}) = \mathbf{U}_n(\boldsymbol{\beta}, \infty)$, where

$$\mathbf{U}_n(\boldsymbol{\beta}, t) = \sum_{i=1}^n \int_0^t \{\mathbf{Z}_i - \bar{\mathbf{Z}}(\boldsymbol{\beta}, u)\} dM_i(u).$$

At the true parameter value $\boldsymbol{\beta}_0$, $\mathbf{U}_n(\boldsymbol{\beta}_0, t)$ is a mean zero martingale with respect to $\{\mathcal{F}_t : t \geq 0\}$, and the partial likelihood score has expectation zero. Asymptotic normality follows from the martingale central limit theorem, and the moments

can be computed using results from the martingale calculus. Andersen and Gill (1982), Gill (1984) and Fleming and Harrington (1991) contain discussions of the martingale approach to the Cox model.

1.4 Weighting the Complete Case Score to Remove Bias

Before describing our weighted estimating equations, we first characterize the missing data mechanism. Throughout, we assume that $(X_i, \delta_i, \mathbf{Z}_i, R_i)$ are independent and identically distributed random vectors for $i = 1, \dots, n$, where $R_i = 1$ if \mathbf{Z}_i is fully observed, and is 0 otherwise. Let $\mathbf{Z}_{i,all}$ denote the components of \mathbf{Z}_i which are observed with probability one. The consistency and asymptotic normality of the estimators proposed below require the following assumptions:

$$P(R_i = 1 | X_i, \delta_i, \mathbf{Z}_i) = P(R_i = 1 | X_i, \delta_i, \mathbf{Z}_{i,all}) \equiv \pi_i, \quad (1.4)$$

and

$$\pi_i > \delta_2^2 > 0,$$

for some δ_2^2 , with probability one. If only one component of the vector \mathbf{Z}_i is not contained in $\mathbf{Z}_{i,all}$, equation (1.4) is precisely missing at random in Little and Rubin's sense (1987).

We propose the following weighted complete case pseudo-score function for inference in the Cox proportional hazards model with missing covariates:

$$\mathbf{U}_{wn}(\boldsymbol{\beta}) = \sum_{i=1}^n \int_0^\infty \frac{R_i}{\pi_i} \{ \mathbf{Z}_i - \bar{\mathbf{Z}}_w(\boldsymbol{\beta}, u) \} dN_i(u), \quad (1.5)$$

where,

$$\bar{\mathbf{Z}}_w(\boldsymbol{\beta}, u) = \frac{\sum_{j=1}^n \frac{R_j}{\pi_j} \mathbf{Z}_j Y_j(u) e^{\boldsymbol{\beta}' \mathbf{Z}_j}}{\sum_{j=1}^n \frac{R_j}{\pi_j} Y_j(u) e^{\boldsymbol{\beta}' \mathbf{Z}_j}}.$$

The subscript wn denotes a weighted score from a sample of n cases. We propose estimating β_0 by $\hat{\beta}$ solving $\mathbf{U}_{wn}(\hat{\beta}) = \mathbf{0}$. Our estimating function (1.5) is motivated by the weighted estimating equations proposed by Robins et al. (1992), for non-linear regression models with missing covariates in the absence of right censoring. Robins and Rotnitzky's (1992) estimating equation for the proportional hazards model is equivalent to (1.5), with dependent censoring taking the role of missingness. Note that only cases with complete data contribute to (1.5). Each complete case is weighted by the inverse of the conditional probability of selection into the subsample of complete cases. Cases which are under-represented in the sample will have their contribution to the score inflated to compensate for other missing cases with the same covariates.

We now show that $n^{-1}\mathbf{U}_{wn}(\beta)$ is asymptotically unbiased, that is, $n^{-1}\mathbf{U}_{wn}(\beta)$ converges almost surely to a function $\mathbf{U}(\beta)$ such that $\mathbf{U}(\beta_0) = \mathbf{0}$. The key step is establishing that $n^{-1}\mathbf{U}_{wn}(\beta)$ is asymptotically equivalent to the unweighted Cox score $n^{-1}\mathbf{U}_n(\beta)$, that is that $|n^{-1}\mathbf{U}_{wn}(\beta) - n^{-1}\mathbf{U}_n(\beta)| \xrightarrow{a.s.} 0$. Asymptotic unbiasedness follows from the properties of the unweighted Cox score. In fact, each of these two statistics is a member of a class of four closely related statistics that are asymptotically equivalent:

$$n^{-1}\mathbf{U}_{nw}(\beta) = n^{-1} \sum_{j=1}^n \int_0^\tau R_i \pi_i^{-1} \{ \mathbf{Z}_i - \bar{\mathbf{Z}}_w(\beta, u) \} dN_i(u), \quad (1.6)$$

$$n^{-1}\mathbf{U}_{nw}^\mu(\beta) = n^{-1} \sum_{j=1}^n \int_0^\tau R_i \pi_i^{-1} \{ \mathbf{Z}_i - \boldsymbol{\mu}(\beta, u) \} dN_i(u), \quad (1.7)$$

$$n^{-1}\mathbf{U}_n(\beta) = n^{-1} \sum_{j=1}^n \int_0^\tau \{ \mathbf{Z}_i - \bar{\mathbf{Z}}(\beta, u) \} dN_i(u). \quad (1.8)$$

$$n^{-1}\mathbf{U}_n^\mu(\beta) = n^{-1} \sum_{j=1}^n \int_0^\tau \{ \mathbf{Z}_i - \boldsymbol{\mu}(\beta, u) \} dN_i(u). \quad (1.9)$$

Note that the superscript μ notation in $U_{nw}^\mu(\boldsymbol{\beta})$ denotes the replacement of the empirical average $\bar{\mathbf{Z}}_w(\boldsymbol{\beta}, u)$ by its large sample limit $\boldsymbol{\mu}(\boldsymbol{\beta}, u)$.

A conditional expectation argument is useful in establishing that statistics (1.7) and (1.9) have the same large sample limit. In the sequel, we often use the argument

$$\begin{aligned} E \left[R\pi^{-1}f(\delta, X, \mathbf{Z}) \right] &= E \left[f(\delta, X, \mathbf{Z}) E \left[R\pi^{-1} | \delta, X, \mathbf{Z} \right] \right] \\ &= E f(\delta, X, \mathbf{Z}). \end{aligned}$$

The latter equality follows from the definition of $\pi = P(R = 1 | \delta, X, \mathbf{Z})$.

To show the asymptotic equivalence of statistics (1.6) and (1.7) it suffices to establish that $\bar{\mathbf{Z}}_w(\boldsymbol{\beta}, u)$ converges to $\boldsymbol{\mu}(\boldsymbol{\beta}, u)$, uniformly in u , ie. that

$$\sup_{0 \leq u \leq \tau} \left| \bar{\mathbf{Z}}_w(\boldsymbol{\beta}, u) - \boldsymbol{\mu}(\boldsymbol{\beta}, u) \right| \xrightarrow{a.s.} \mathbf{0}.$$

If we consider the centering term in the Cox score with no missing data, $\bar{\mathbf{Z}}(\boldsymbol{\beta}, u)$, to be a special case of the weighted version $\bar{\mathbf{Z}}_w(\boldsymbol{\beta}, u)$ with weights set identically to one, then this result is sufficient also to show the equivalence of (1.8) and (1.9). The following regularity conditions are helpful: we assume the existence of an interval $[0, \tau]$ on which the probability that a case has complete data and is at risk is bounded away from zero, and that the covariates are bounded.

The large sample limit of $\bar{\mathbf{Z}}_w(\boldsymbol{\beta}, u)$ is given by

$$\boldsymbol{\mu}(\boldsymbol{\beta}, u) \equiv \lim_{n \rightarrow \infty} \frac{n^{-1} \sum_{j=1}^n R_j \pi_j^{-1} Y_j(u) e^{\boldsymbol{\beta}' \mathbf{z}_j} \mathbf{z}_j}{n^{-1} \sum_{j=1}^n R_j \pi_j^{-1} Y_j(u) e^{\boldsymbol{\beta}' \mathbf{z}_j}} = \frac{E \left[Y_j(u) \mathbf{z}_j e^{\boldsymbol{\beta}' \mathbf{z}_j} \right]}{E \left[Y_j(u) e^{\boldsymbol{\beta}' \mathbf{z}_j} \right]},$$

with probability one. Pointwise convergence to this limit follows by applying the strong law of large numbers to the numerator and denominator of $\bar{\mathbf{Z}}_w(\boldsymbol{\beta}, u)$, which are each averages of bounded identically distributed random variables.

Recall the definition of $Y_j(u) = I_{\{X_j \geq u\}}$. Making this substitution into the numerator and denominator of $Z_w(\beta, u)$, we see that they are analogous to weighted empirical survivor functions. The Glivenko Cantelli Theorem (see Chung (1974) p. 133), which establishes the convergence of the empirical distribution function may be modified to prove that

$$\sup_{0 \leq u \leq \tau} \left| n^{-1} \sum_{j=1}^n R_j \pi_j^{-1} Y_j(u) e^{\beta' Z_j} Z_j - E \left[Y_j(u) Z_j e^{\beta' Z_j} \right] \right| \xrightarrow{a.s.} 0,$$

and

$$\sup_{0 \leq u \leq \tau} \left| n^{-1} \sum_{j=1}^n R_j \pi_j^{-1} Y_j(u) e^{\beta' Z_j} - E \left[Y_j(u) e^{\beta' Z_j} \right] \right| \xrightarrow{a.s.} 0.$$

Under the regularity conditions assumed, this implies that the convergence of $\bar{Z}_w(\beta, u)$ to $\mu(\beta, u)$ is uniform in u .

This completes the chain of asymptotic equivalence (1.6) \Leftrightarrow (1.7) \Leftrightarrow (1.9) \Leftrightarrow (1.8) linking the weighted and unweighted scores. Since the weighted score is asymptotically unbiased, Foutz's result (1977) can be used to establish the existence of a unique strongly consistent root. Instead, we use the equivalence of the weighted and unweighted scores and modify Gill's (1984) proof of the consistency of the maximum partial likelihood estimate. The details are deferred to the appendix.

1.5 Asymptotic Results for the Weighted Complete Case Score

The key assumptions and results are summarized in this section. Outlines of the proofs are deferred to the Appendix. The main result is that in large samples, the parameter estimates are unbiased and normally distributed, with an asymptotic variance which depends on whether the weights used in the estimating equation are known a priori or are estimated from the data.

With weights π_i known, the estimate $\hat{\beta}$ solves the weighted complete-case score equations

$$\mathbf{U}_{wn}(\beta) \equiv \sum_{i=1}^n \int_0^\tau \frac{R_i}{\pi_i} \left\{ \mathbf{Z}_i - \frac{\sum_{j=1}^n \frac{R_j}{\pi_j} \mathbf{Z}_j Y_j(u) e^{\beta' \mathbf{Z}_j}}{\sum_{j=1}^n \frac{R_j}{\pi_j} Y_j(u) e^{\beta' \mathbf{Z}_j}} \right\} dN_i(u) = \mathbf{0}. \quad (1.10)$$

Although the asymptotics assume no tied observation times, a small number of ties can be incorporated by noting that equation (1.10) is a weighted version of the Peto approximation to the score for tied data.

In practice the weights π_i are unknown and must be estimated. We shall do so by specifying a parametric model $\pi_i(\alpha)$ for the π_i . Specifically we assume that $\pi_i = \pi_i(\alpha_0)$, where α_0 is an unknown parameter vector, and that for each α , $\pi(\alpha) = \pi(X_i, \delta_i, \mathbf{Z}_{i,all}; \alpha)$ is a smooth function of X_i , δ_i , and $\mathbf{Z}_{i,all}$ taking values between 0 and 1. Typically the logit of $\pi_i(\alpha)$ is taken to be a linear function of X_i , δ_i and the components of $\mathbf{Z}_{i,all}$.

The dependence of the score on the parameterization for the weights is emphasized by adding α to the notation. Define

$$\mathbf{U}_{wn}(\alpha, \beta) \equiv \sum_{i=1}^n \int_0^\tau \pi_i(\alpha)^{-1} R_i \left\{ \mathbf{Z}_i - \frac{\sum_{j=1}^n \pi_j(\alpha)^{-1} R_j \mathbf{Z}_j Y_j(u) e^{\beta' \mathbf{Z}_j}}{\sum_{j=1}^n \pi_j(\alpha)^{-1} R_j Y_j(u) e^{\beta' \mathbf{Z}_j}} \right\} dN_i(u). \quad (1.11)$$

For fixed α let $\hat{\beta}(\alpha)$ solve

$$\mathbf{U}_{wn}(\alpha, \beta) = \mathbf{0}. \quad (1.12)$$

Hence $\hat{\beta}$ solving (1.10) is equivalent to $\hat{\beta}(\alpha_0)$ solving (1.11).

When the weights are unknown, we estimate π_i^{-1} by $\pi_i(\hat{\alpha})^{-1}$, where $\hat{\alpha}$ denotes the maximum likelihood estimate of α_0 . That is, $\hat{\alpha}$ solves

$$\mathbf{T}_n(\alpha) = \sum_{i=1}^n \frac{\partial}{\partial \alpha} [R_i \log \pi_i(\alpha) + (1 - R_i) \log(1 - \pi_i(\alpha))] = \mathbf{0}.$$

We then solve $\mathbf{U}_{wn}(\hat{\alpha}, \beta) = \mathbf{0}$ to obtain $\hat{\beta}(\hat{\alpha})$.

The asymptotic distribution of $\hat{\beta}(\hat{\alpha})$ is obtained through a Taylor expansion of the score $n^{-1/2}\mathbf{U}_{wn}(\alpha, \beta)$ around the true parameter values α_0, β_0 . Normality follows from the Central Limit Theorem since the scores $n^{-1/2}\mathbf{U}_{wn}(\alpha_0, \beta_0)$ and $n^{-1/2}\mathbf{T}_n(\alpha_0)$ are asymptotically equivalent to $n^{-1/2}$ times the sum of i.i.d. terms, with summands

$$\tilde{\mathbf{U}}_{wn,i}^{\mu}(\alpha_0, \beta_0) \equiv \int_0^{\tau} \pi_i(\alpha_0)^{-1} R_i \{Z_i - \mu(\beta_0, u)\} dM_i(u)$$

and

$$\mathbf{T}_{n,i}(\alpha_0) \equiv \left. \frac{\partial}{\partial \alpha} [R_i \log \pi_i(\alpha) + (1 - R_i) \log(1 - \pi_i(\alpha))] \right|_{\alpha=\alpha_0}$$

respectively. The subscript n, i , tilde and superscript μ denote that $\tilde{\mathbf{U}}_{wn,i}^{\mu}$ is the i th in a sum of n terms, each of which is a stochastic integral with respect to $dM_i(u)$, of covariates centered at μ . The briefer notation \mathbf{U}_i will be used in the remainder of this section. The derivation of the asymptotic variance uses Pierce's (1982) results for the limiting distribution of statistics in which nuisance parameters have been replaced by efficient estimates, as in Robins, Mark and Newey (1992) and Robins, Zhao and Lipsitz (1990).

Theorem 1 (*Asymptotic Distribution of the Estimate*) Under the modeling and regularity assumptions listed in the appendix,

$$n^{1/2}(\hat{\beta}(\hat{\alpha}) - \beta_0) \xrightarrow{L} N_p(\mathbf{0}, \Sigma^{-1} \mathbf{V} \Sigma^{-1}), \quad (1.13)$$

where $\Sigma \equiv -E \left[\frac{\partial}{\partial \beta} \mathbf{U}_i(\alpha_0, \beta_0) \right]$ and $\mathbf{V} \equiv E[\mathbf{U}_i \mathbf{U}_i'] - E[\mathbf{U}_i \mathbf{T}_i'] E[\mathbf{T}_i \mathbf{T}_i']^{-1} E[\mathbf{T}_i \mathbf{U}_i']$. Here \xrightarrow{L} denotes convergence in law, and N_p denotes the p -dimensional multivariate normal distribution. The limiting distribution of $\hat{\beta}(\alpha_0)$ is also given by equation (1.13) except with \mathbf{V} replaced by $E[\mathbf{U}_i \mathbf{U}_i']$. In both instances, expectations are

taken with respect to the true distribution of the covariate, survival, censoring, and missingness variables.

An analogous result is given in Robins and Rotnitzky (1992). Note that \mathbf{V} is the variance of the residual from a population regression of \mathbf{U}_i on \mathbf{T}_i , the summands in the scores for β and α , respectively. Estimating the weights, as opposed to using known weights, never decreases and usually increases the efficiency of our estimate of β because the scores for α contain information from incomplete cases which would otherwise be ignored. If the data are missing completely at random in the sense that π_i is a constant π , then $\hat{\beta}(\alpha)$ is exactly the complete case estimator $\hat{\beta}_{comp}$ of Section 2. Since $\hat{\beta}(\hat{\alpha})$ is always at least as efficient and usually more efficient than $\hat{\beta}(\alpha_0)$, $\hat{\beta}(\hat{\alpha})$ will improve on the efficiency of $\hat{\beta}_{comp}$ when π is constant.

The asymptotic variance of the estimates may be estimated from the data. The following estimator for Σ should be consistent under mild regularity assumptions:

$$\hat{\Sigma} \equiv n^{-1} \sum_{i=1}^n \int_0^\tau \pi_i(\hat{\alpha})^{-1} R_i \left[\frac{\mathbf{S}_w^{(2)}(\hat{\beta}(\hat{\alpha}), u)}{S_w^{(0)}(\hat{\beta}(\hat{\alpha}), u)} - \left\{ \frac{\mathbf{S}_w^{(1)}(\hat{\beta}(\hat{\alpha}), u)}{S_w^{(0)}(\hat{\beta}(\hat{\alpha}), u)} \right\}^{\otimes 2} \right] dN_i(u),$$

where

$$\mathbf{S}_w^{(k)}(\hat{\beta}(\hat{\alpha}), u) \equiv n^{-1} \sum_{j=1}^n \pi_j(\hat{\alpha})^{-1} R_j \lambda_j(u) e^{\hat{\beta}(\hat{\alpha})' \mathbf{Z}_j} \mathbf{Z}_j^{\otimes k}, \quad k = 0, 1, 2.$$

For any vector \mathbf{Z} , $\mathbf{Z}^{\otimes 2} = \mathbf{Z}\mathbf{Z}'$, $\mathbf{Z}^{\otimes 1} = \mathbf{Z}$, and $\mathbf{Z}^{\otimes 0} = 1$.

An estimate of \mathbf{V} may be obtained by first regressing $\hat{\mathbf{U}}_i$ on $\hat{\mathbf{T}}_i$, then computing the empirical covariance matrix of the residuals, where

$$\hat{\mathbf{U}}_i \equiv \int_0^\tau \pi_i(\hat{\alpha})^{-1} R_i \left\{ \mathbf{Z}_i - \frac{\mathbf{S}_w^{(1)}(\hat{\beta}(\hat{\alpha}), u)}{S_w^{(0)}(\hat{\beta}(\hat{\alpha}), u)} \right\} d\hat{M}_i(u),$$

and

$$\hat{\mathbf{T}}_i \equiv \frac{\partial}{\partial \alpha} [R_i \log \pi_i(\alpha) + (1 - R_i) \log(1 - \pi_i(\alpha))] \Big|_{\alpha = \hat{\alpha}}.$$

The estimate for \mathbf{U}_i requires an estimate of

$$dM_i(u) = dN_i(u) - \lambda_0(u)Y_i(u)e^{\beta_0' \mathbf{z}_i} du.$$

and hence for the baseline hazard $\lambda_0(u)$. We modify the Breslow (1972,1974) estimator for the cumulative hazard, obtaining:

$$\hat{\Lambda}_0(t) = \int_0^t \frac{\sum_{j=1}^n dN_j(u)}{\sum_{j=1}^n \pi_j(\hat{\alpha})^{-1} R_j Y_j(u) e^{\beta(\hat{\alpha})' \mathbf{z}_j}}.$$

and thus,

$$\hat{\lambda}_0(u) = \frac{\sum_{j=1}^n dN_j(u)}{\sum_{j=1}^n \pi_j(\hat{\alpha})^{-1} R_j Y_j(u) e^{\beta(\hat{\alpha})' \mathbf{z}_j}}.$$

Although it is tempting to replace $dM_i(u)$ by $dN_i(u)$ in our estimate of \mathbf{U}_i , this leads to correlated increments and invalidates the usual sums of squares estimates of covariance.

1.5.1 A Small Simulation Study

We simulated data from a simple randomized clinical trial testing a treatment effect with uniform patient accrual in which a nuisance covariate was missing at random for some cases. This simulation allowed us to check that the method reduced bias. More exhaustive simulations are being conducted to investigate other properties of the estimator such as rates of convergence, small sample behavior, and sensitivity to modeling assumptions, and will be reported elsewhere.

Censoring times were uniformly distributed on the interval from zero to three. The baseline hazard was the constant 1, and the conditional hazard depended on two covariates. The treatment indicator variable had values zero (treatment) and one (control) with equal probability, while the second covariate took values -1, 0, and 1 with probability .3, .3 and .4 respectively. The regression coefficients in the

Cox model were 1 (treatment indicator) and .5 (nuisance covariate). Because of the treatment coding convention, the coefficient of 1 indicates improved survival on the treatment arm of the study.

Our missing data mechanism models a scenario where investigators obtain data retrospectively on a covariate whose clinical importance has been only recently recognized. The investigators may be unable to obtain data for patients who died or who entered the study early in the trial, because of the cost of retrieving records from archives. Long event times can occur only with long follow-up times, so X_i is used as a proxy for early entry in the study in the model for probability of completeness. For the true missing data model, we use a linear model on a logistic scale, with probability of completeness related to a constant, event time and censoring status. Specifically,

$$\text{logit } \pi_i(\alpha_0) = 3 - 3X_i - .5\delta_i \quad (1.14)$$

In each simulation run, we estimated the parameters in the logistic model (1.14) using ordinary logistic regression.

One thousand samples, each with one thousand observations were simulated, and weighted and unweighted estimates of the parameters β_1 and β_2 and their standard errors, were computed for each sample. The proportion of cases without missing covariates was 72%, the percentage of cases with observed failures was 78%.

Table 1 about here

Table 1.1 shows the results from the simulation study. The amount of bias in the parameter estimates can be assessed by comparing the empirical means of the

estimates with the true parameter values of 1 and .5. The complete case analysis leads to substantial bias (25 %) and the weighted analysis removes that bias. The bias in the estimated standard errors can be quantified by comparing the empirical mean of the standard error estimates with the empirical standard deviation of the parameter estimates. The complete case analysis correctly estimates the parameter standard errors. Therefore, the poor coverage of the nominal 95% confidence intervals in the unweighted analysis, (21 % and 41 % for β_1 and β_2 respectively), is largely due to bias in estimating the parameters themselves.

1.6 Discussion

The current methods proposed in the literature for handling missing covariates in the proportional hazards model are useful in special situations, but are not generally applicable to many datasets that confront an investigator. Full likelihood methods are useful for missing data problems, but they require correct specification of the joint distribution of the (missing and observed) data, and the EM algorithm can be very slow to converge. The advantages of the likelihood method proposed by Schluchter and Jackson (1989) are the flexibility of log-linear modeling in testing the proportional hazards assumption and covariate interactions, and the gain in efficiency made possible by using all of the cases. However, mis-specification of the piecewise exponential model they propose could lead to substantial bias in estimation; further, their method requires categorical or grouped covariates, and thus will not always be useful. The method proposed by Schemper and Smith (1990) for testing treatment effects would not be adequate when modeling the natural history of the disease, where inference on the covariate effects themselves is of primary importance. The method of Lin and Ying (1992) requires the data

to be missing completely at random. Our proposed method requires the weaker condition of (1.4) that allows missingness to depend on survival and censoring time. Furthermore, the method we discuss here can be used with continuous covariates and is easily adapted to data with a small number of tied failures times.

Our method poses some intriguing and unanswered questions. Firstly, the small sample properties of our estimator need further evaluation. Because of the broad range of possible survival models and missing data configurations, it is not possible to draw definitive conclusions from the one simulation we report. Although our estimator reduces the bias, it will not necessarily be the most efficient estimator. Robins and Rotnitzky (1992) have derived a general expression for the efficient score in an arbitrary semi-parametric model (including the Cox proportional hazards model) with the data missing at random. They note that in the Cox proportional hazards model computation of a semi-parametric efficient estimator requires the solution of an integral equation which has no closed form. One could examine the further increases in efficiency that can be gained by numerically solving their integral equation.

The properties of our estimator when the model for π_i is misspecified have not yet been studied. However, suppose that given a linear logistic selection model $\pi(\boldsymbol{\alpha})$ for π_i , we added to the model additional terms such as powers of X_i and the components of $Z_{i,all}$ and their interactions. This will increase the number of parameters in our selection model. As noted by Robins, Mark and Newey (1992), as we increase the number of parameters we derive two benefits. First we reduce the large sample bias in $\hat{\beta}(\hat{\boldsymbol{\alpha}})$ due to the misspecification of the model for π_i . Second, even when the original linear logistic model was correctly specified the variance of $\hat{\beta}(\hat{\boldsymbol{\alpha}})$ will be non-increasing and will usually decrease. This follows

from the fact that as the number of free parameters in our model for π_i increases, the dimension of the score \mathbf{T}_i will increase and hence the number of regressors in the population regression of \mathbf{U}_i on \mathbf{T}_i will increase. But by standard likelihood theory the variance \mathbf{V} of the residuals from a population regression never increases and usually decreases with an increasing number of regressors. Thus it would be advantageous to use a richly parameterized model $\pi_i(\boldsymbol{\alpha})$ both to guard against misspecification bias and to increase the efficiency of $\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\alpha}})$. However, the apparent gain is tempered by two facts. First we show that no matter how many covariates we add to our selection model the asymptotic variance of $\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\alpha}})$ will never be less than

$$\boldsymbol{\Sigma}^{-1} \text{Var} \{ \mathbf{U}_i - \mathbf{G}_i^* \} \boldsymbol{\Sigma}^{-1}$$

where

$$\mathbf{G}_i^* = (R_i - \pi_i) E [\mathbf{U}_i | X_i, \delta_i, \mathbf{Z}_{i,all}] / \pi_i.$$

Second, our argument that $\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\alpha}})$ is asymptotically normal and unbiased for $\boldsymbol{\beta}_0$ assumes $\hat{\boldsymbol{\alpha}}$ is $n^{1/2}$ -consistent, which limits the number of free parameters in our model $\pi(\boldsymbol{\alpha})$ for $\boldsymbol{\alpha}_0$. However, results in Newey (1992) imply that $n^{1/4}$ - consistency is sufficient which in principle would allow us to use a non-parametric, eg. multivariate kernel regression estimate of π_i .

Acknowledgements

The authors thank Zhiliang Ying for suggesting the integration by parts approach which substantially simplified and shortened the proof of asymptotic normality of the reweighted score.

This investigation was supported by grants CA57253 and CA39929 awarded by the National Cancer Institute, DHHS.

Table 1.1: Empirical Estimates of Log Relative Risk for Weighted and Unweighted Complete Case Analyses.

	Weighted		Un-Weighted	
	β_1	β_2	β_1	β_2
true parameter value	1.000	0.500	1.000	0.500
mean of 1000 estimates	.9899	.4959	.7571	.3849
mean of 1000 std. errors	.1508	.0907	.0883	.0534
std. dev. of 1000 estimates	.1943	.1292	.0853	.0554
Empirical Coverage of Nominal 95% CI	.8960	.8650	.2080	.4110

Estimating the coefficient of Z_1 in the Cox model $\lambda = \exp(Z_1 + .5Z_2)$. Covariate Z_1 takes values 0 and 1 with equal probability, Z_2 takes values -1.0, and 1 with probabilities .3 and .4, and Z_2 is observed with probability $\logit \pi_i = 3 - 3X_i - .5\delta_i$. The follow-up times X are the minimum of an exponential failure time with hazard λ and a censoring time uniform on (0,3).

References

- Andersen, P.K. and Gill, R.D. (1982) Cox's regression model for counting processes: a large sample study. *The Annals of Statistics*, 10. 1100-1120.
- Chung, Kai Lai, (1974) *A Course in Probability Theory, Second Edition*. San Diego, Academic Press, Inc.
- Cox, D.R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society B* 24. 187-220.
- Fleming, T.R. and Harrington, D.P. (1991) *Counting Processes and Survival Analysis*. New York, John Wiley and Sons, Inc.
- Gill, R.D. (1984). Understanding Cox's regression model: A martingale approach. *Journal of the American Statistical Association* 79. 441-7.
- Glynn, R.J. and Laird, N.M. (1985) Regression Estimates and Missing Data: Complete Case Analyses. *Doctoral Thesis, Department of Biostatistics, Harvard School of Public Health, Boston. Massachusetts 02115*.
- Lin, D.Y., Ying, Z. (1992). Cox regression with incomplete covariate measurements. *Technical Report No. 112, February 1992, University of Washington, Department of Biostatistics, SC-32, School of Public Health and Community Medicine, Seattle, Washington 98195*.
- Little, R.J.A., Rubin, D.B. (1987). *Statistical Analysis with Missing Data*, New York, John Wiley and Sons, Inc.
- Pepe, M. S., Self, S.G. and Prentice, R. L. (1989). Further results on covariate measurements in cohort studies with time to response data. *Statistics in*

Medicine 8, 1167-1178.

Prentice, R.L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* 73, 1-11.

Robins, J., Mark, S. and Newey, W. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, 48, 479-495.

Robins, J., Rotnitzky, A. and Zhao, L.P. (1992). Analysis of semiparametric regression models for repeated outcomes under the presence of missing data. Submitted to *Journal of the American Statistical Association*

Robins, J. and Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. *AIDS Epidemiology: Methodologic Issues*. Jewell, N.P., Dietz, K. and Farewell, B. editors, Boston, Birkhauser, pp. 297-331.

Robins, J.M., Zhao, L.P. and Lipsitz, S.R. (1992). Estimation of regression coefficients when a regressor is not always observed. *Technical Report, Department of Biostatistics, Harvard School of Public Health, Boston, MA.*

Rubin, D.B. (1976). Inference and missing data. *Biometrika* 63, 81-92.

Schemper, M. and Smith, T.L. (1990). Efficient evaluation of treatment effects in the presence of missing covariate values. *Statistics in Medicine* 9, 777-784.

Schluchter, M.D. and Jackson, K.L. (1989). Log-linear analysis of censored survival data with partially observed covariates. *Journal of the American Statistical Association* 84, 42-52.