Cox Regression with Incomplete Covariate Measurements

Author(s): D. Y. Lin and Z. Ying

Source: *Journal of the American Statistical Association*, Dec., 1993, Vol. 88, No. 424 (Dec., 1993), pp. 1341-1349

Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association

Stable URL: https://www.jstor.org/stable/2291275

Cox Regression With Incomplete Covariate Measurements

D. Y. Lin and Z. Ying*

This article provides a general solution to the problem of missing covariate data under the Cox regression model. The estimating function for the vector of regression parameters is an approximation to the partial likelihood score function with full covariate measurements and reduces to the pseudolikelihood score function of Self and Prentice in the special setting of case-cohort designs. The resulting parameter estimator is consistent and asymptotically normal with a covariance matrix for which a simple and consistent estimator is provided. Extensive simulation studies show that the large-sample approximations are adequate for practical use. The proposed approach tends to be more efficient than the complete-case analysis, especially for large cohorts with infrequent failures. For case-cohort designs, the new methodology offers a variance-covariance estimator that is much easier to calculate than the existing ones and allows multiple subcohort augmentations to improve efficiency. Real data taken from clinical and epidemiologic studies are analyzed.

KEY WORDS: Case-cohort design; Censoring; Failure time; Missing data; Proportional hazards; Survival analysis.

1. INTRODUCTION

The commonly used Cox (1972) regression model postulates that the hazard function for the failure time T associated with a $p \times 1$ vector of possibly time-varying covariates $\mathbf{Z}$ takes the form of

$$\lambda(t; \mathbf{Z}) = \lambda_0(t)\exp\{\beta_0'\mathbf{Z}(t)\}, \tag{1}$$

where $\lambda_0(t)$ is an unspecified baseline hazard function and β_0 is a $p \times 1$ vector of unknown regression parameters. The failure time is assumed to be subject to independent right censorship.

Let $(T_i, C_i, \mathbf{Z}_i)$ $(i = 1, \ldots, n)$ be n independent replicates of $(T, C, \mathbf{Z})$, where C is the censoring time. Define $X_i = \min(T_i, C_i)$, $\Delta_i = I(T_i \le C_i)$ and $Y_i(t) = I(X_i \ge t)$, where $I(\cdot)$ is the indicator function. If the information on $(X_i, \Delta_i, \mathbf{Z}_i)$ is available for all $i = 1, \ldots, n$, then the partial likelihood score function for β_0 is

$$\mathbf{U}_f(\beta) = \sum_{i=1}^{n} \Delta_i \{\mathbf{Z}_i(X_i) - \bar{\mathbf{Z}}(\beta, X_i)\}, \tag{2}$$

where

$$\bar{\mathbf{Z}}(\beta, t) = \frac{\sum_{l=1}^{n} Y_l(t)\exp\{\beta'\mathbf{Z}_l(t)\}\mathbf{Z}_l(t)}{\sum_{l=1}^{n} Y_l(t)\exp\{\beta'\mathbf{Z}_l(t)\}}. \tag{3}$$

Note that $\bar{\mathbf{Z}}(\beta, t)$ is the conditional expectation of $\mathbf{Z}_l(t)$ on $\{l: X_l \ge t\}$ with respect to a probability distribution proportional to $\exp\{\beta'\mathbf{Z}_l(t)\}$. The maximum partial likelihood estimator (MPLE) $\hat{\beta}$, defined as the solution to the score equation $\{\mathbf{U}_f(\beta) = \mathbf{0}\}$, has been shown to be approximately normal in large samples with mean β_0 and with covariance matrix $-\{\partial\mathbf{U}_f(\hat{\beta})/\partial\beta\}^{-1}$ (Andersen and Gill 1982; Tsiatis 1981).

In many applications, measurements on certain components of the covariate vector are missing on some study subjects. It is particularly rare to observe entire covariate histories on all subjects when continuous time-varying covariates are involved. The following general examples illustrate the main features of the problem.

Example 1.1 In clinical trials, it is often desirable to adjust for the effects of other covariates when testing for treatment differences so as to correct for baseline imbalances on prognostic variables and/or to increase statistical power. But measurements on some important prognostic factors may be lacking on a subset of patients.

Example 1.2 For modeling the natural history of a given disease, certain prognostic variables (e.g., age, cholesterol level) require only clinical evaluations and blood samples, whereas others (e.g., histological stage) may require biopsy and thus are invasive and expensive. Clearly, measurements are more readily available on the former than on the latter.

Example 1.3 When studying the effect of a biological marker process on the development of a clinical endpoint, serial measurements on the marker require regular visits to the clinics/centers by the participants. The availability of measurements depends on the actual timings of the visits. The measurements usually become rather sparse near the end of follow-up.

Example 1.4 Epidemiologic cohort studies and disease prevention trials often involve the follow-up of several thousand subjects for a number of years. Randomization assignments and certain characteristics (e.g., demographics) are known on all study subjects. On the other hand, the assembly of covariate histories which requires biochemical analysis of blood samples or other specimens or the hand coding of individual diet records can be prohibitively expensive if done on all cohort members.

The "naive" ways of handling incomplete covariate measurements in the Cox regression analysis are (1) to impute

* D. Y. Lin is Assistant Professor, Department of Biostatistics, SC-32, University of Washington, Seattle, WA 98195. Z. Ying is Associate Professor, Department of Statistics, University of Illinois, Champaign, IL 61820. This work was supported by the National Institute of General Medical Sciences and the National Institute of Allergy and Infectious Diseases (for Lin) and by the National Science Foundation and the National Security Agency (for Ying). Part of this research was conducted while Ying was visiting the Mathematical Sciences Research Institute at Berkeley. The authors thank the reviewers and Barbara McKnight for their helpful comments.

missing values, (2) to disregard subjects with any missing covariate values, and (3) to include in the model only those covariates that have complete measurements on every subject. The first approach can induce considerable bias into the parameter and variance estimators; the second can lead to substantial reduction in efficiency if subjects with uncensored failure times are deleted; and the third may not only reduce the power, but also distort the size of the partial-likelihood–based tests.

There exist formal solutions to the problem of missing covariate data in the literature of case-cohort designs. The case-cohort design was proposed by Prentice (1986) as a means of reducing the cost of large epidemiologic cohort studies described in Example 1.4. Under this design, raw covariate data are collected on all study subjects and then processed only on the cases (i.e., the subjects who experience failures) and on a random subset of the entire cohort selected in advance of cohort folow-up. Self and Prentice (1988) described an estimating equation for the parameter vector β_0 that results from replacing the full-cohort conditional expectation $\bar{\mathbf{Z}}(\beta, t)$ defined in (3) by its subcohort counterpart. These authors also provided the asymptotic distribution theory and efficiency results for their parameter estimator. For a large cohort with infrequent events, the efficiency with which β_0 may be estimated depends strongly on the number of cases whereas the marginal contributions from the noncases are small; therefore, the loss in efficiency of the parameter estimator under the case-cohort design relative to the MPLE under the full-cohort design will be minimal provided that there is an adequate subcohort size at each uncensored failure time. Unfortunately, the analytical variance estimators for this design provided by Prentice (1986) and by Self and Prentice (1988) are very complicated, and the bootstrap variance estimators suggested by Wacholder, Gail, Pee, and Brookmeyer (1989) seem too time-consuming to calculate for large studies. A more serious concern with the case-cohort design has been the potential lack of subcohort members for the latest events. It then seems desirable to augment the subcohort with additional random samples when it is wearing thin (Prentice 1986). But the augmentation will further complicate the variance formulas if the existing variance estimation strategies are adopted. Despite these limitations, the case-cohort design has been used in a number of studies (see, for example, Chyou, Nomura, Hankin, and Stemmermann 1990; Overvad et al. 1991; Sorensen and Sonne-Holm 1988).

In this article, we develop a general procedure for making inference about β_0 in the presence of missing covariate data. Mimicking the idea of Self and Prentice (1988), we estimate the conditional expectation $\bar{\mathbf{Z}}(\beta, t)$ from the subjects who have complete measurements on all covariate components at time t or from some representative sample of the entire cohort. The estimating function for β_0 is then taken as the sum over the uncensored failure times of the observed value of $\mathbf{Z}_i(X_i)$ minus its "estimated" conditional expectation. If the jth component of $\mathbf{Z}_i(X_i)$ with $\Delta_i = 1$ is missing, then the jth component of the difference is excluded from the summation. The estimator based on this estimating function, which we shall refer to as the approximate partial likelihood estimator (APLE), will be shown to be consistent and asymptotically normal under reasonable conditions. In addition, a simple and consistent estimator for the limiting covariance matrix will be constructed.

The loss in efficiency of the APLE relative to the MPLE with full covariate measurements will be small unless the covariate vectors associated with uncensored failure times are frequently missing. The proposed estimator is generally more efficient than the MPLE based only on complete cases (i.e., subjects who have complete covariate measurements); the gain in efficiency will be most profound if the covariates of primary interest (e.g., treatment indicators) have measurements on (virtually) all subjects but the secondary covariates are heavily missing.

The case-cohort design may be regarded as a special situation of the general missingness framework presented here. As a result, we provide a variance estimator for this design that is much simpler than the existing ones. Furthermore, the new methodology allows several useful extensions of the case-cohort design. For instance, raw covariate data need not be collected on all study participants. More important, multiple subcohorts may be augmented successively without the need to modify the variance formula.

The rest of this article is organized as follows. Section 2 presents the asymptotic theory for the APLE and corresponding cumulative baseline hazard estimators. The underlying technical developments are deferred to the Appendix. Section 3 reports the results from our extensive Monte Carlo studies, which demonstrate the adequacy of the asymptotic approximations for practical sample sizes as well as the advantages of the APLE over the "naive" estimators. Section 4 illustrates the proposed methods with real data taken from clinical and epidemiologic studies. Section 5 provides several extensions.

2. ASYMPTOTIC THEORY

Suppose that the data consist of iid random quintuplets $\{X_i, \Delta_i, \mathbf{Z}_i(\cdot), H_{0i}(\cdot), \mathbf{H}_i(\cdot)\}$ $(i = 1, \ldots, n)$, where $\mathbf{Z}_i(\cdot) = \{Z_{1i}(\cdot), \ldots, Z_{pi}(\cdot)\}'$ may not be completely observed, $H_{0i}(\cdot)$ is an indicator function, and $\mathbf{H}_i(\cdot)$ is a $p \times p$ diagonal matrix with indicator functions $\{H_{1i}(\cdot), \ldots, H_{pi}(\cdot)\}$ as the diagonal elements. For the problem of partially incomplete covariate vectors described in Examples 1.1–1.3, $H_{ji}(t) = 1$ if $Z_{ji}(t)$ is available and $H_{ji}(t) = 0$ otherwise $(j = 1, \ldots, p)$, and $H_{0i}(t) = I\{H_{ji}(t) = 1 \text{ for all } j = 1, \ldots, p\}$. For the original case-cohort design, $\mathbf{H}_i(\cdot) = \mathbf{I}_p$, the $p \times p$ identity matrix, and $H_{0i}(t) = 1$ if and only if the ith subject belongs to the subcohort at time t. As will be seen later, $H_{0i}(t)$ determines whether or not the ith subject is included in the estimation of $\bar{\mathbf{Z}}(\beta, t)$, and $H_{ji}(t)$ $(j = 1, \ldots, p)$ indicates whether or not the ith subject contributes directly to the jth component of the estimating function.

Assume that, conditional on $\{X_i \geq t\}$, the missing indicators $\{H_{ji}(t); j = 0, 1, \ldots, p\}$ are independent of all other random variables, which corresponds to the missing completely at random (MCAR) assumption of Rubin (1976). Possible ways of relaxing this assumption will be discussed in Section 5. Define $\mathbf{h}(t) = \mathcal{E}\{\mathbf{H}_1(t)\}$ and $h_j(t) = \mathcal{E}\{H_{j1}(t)\}$ $(j = 0, 1, \ldots, p)$, where $\mathcal{E}$ denotes expectation. Note that $\mathbf{h}(t)$ and $h_j(t)$ $(j = 0, 1, \ldots, p)$ are the common expectations

of $\mathbf{H}_i(t)$ and $H_{ji}(t)$ $(j = 0, 1, \ldots, p)$ for all $i = 1, \ldots, n$ and that $\mathbf{h}(t)$ is a $p \times p$ diagonal matrix with $\{h_1(t), \ldots, h_p(t)\}$ as the diagonal elements. We assume that $h_0(t) \geq \alpha_0$ for some $\alpha_0 > 0$ and all t. This condition, together with the MCAR assumption, ensures that at least in large samples the full-cohort conditional expectation $\bar{\mathbf{Z}}(\beta, t)$ can be reliably estimated from the covariate measurements available at time t.

It is convenient to introduce the following notation:

$$\mathbf{S}^{(r)}(\beta, t) = n^{-1} \sum_{i=1}^{n} H_{0i}(t) Y_i(t) \exp\{\beta' \mathbf{Z}_i(t)\} \mathbf{Z}_i(t)^{\otimes r},$$

$$\mathbf{s}^{(r)}(\beta, t) = \mathcal{E}\{\mathbf{S}^{(r)}(\beta, t)\}, \quad \mathbf{E}(\beta, t) = \mathbf{S}^{(1)}(\beta, t)/S^{(0)}(\beta, t),$$

$$\mathbf{e}(\beta, t) = \mathbf{s}^{(1)}(\beta, t)/s^{(0)}(\beta, t),$$

where for a column vector $\mathbf{a}$, $\mathbf{a}^{\otimes 0} = 1$, $\mathbf{a}^{\otimes 1} = \mathbf{a}$, and $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}'$. Then the approximate partial-likelihood score function can be written as

$$\tilde{\mathbf{U}}(\beta) = \sum_{i=1}^{n} \Delta_i \mathbf{H}_i(X_i)\{\mathbf{Z}_i(X_i) - \mathbf{E}(\beta, X_i)\}. \quad (4)$$

The APLE $\tilde{\beta}$ is the root to the estimating equation $\{\tilde{\mathbf{U}}(\beta) = \mathbf{0}\}$, which can be solved by the Newton–Raphson algorithm. Note that $\partial\tilde{\mathbf{U}}(\beta)/\partial\beta = -\sum_{i=1}^{n} \Delta_i \mathbf{H}_i(X_i)\{\mathbf{S}^{(2)}(\beta, X_i)/S^{(0)}(\beta, X_i) - \mathbf{E}(\beta, X_i)^{\otimes 2}\}$.

To develop a rigorous asymptotic theory for the APLE $\tilde{\beta}$, we impose the following regularity conditions:

A. All covariates have bounded total variations, namely, $\int_0^\infty |dZ_{ji}(t)| + |Z_{ji}(0)| \leq K$ for some $K > 0$ and all i, j.

B. There exist $k_0 > 0$ and $\eta_0 > 0$ such that for $j = 0, 1, \ldots, p$ and $r = 0, 1,$

$$\sup_{|d| \leq n^{-k_0}} \left[n^{-1} \left| \sum_{i=1}^{n} \{H_{ji}(t) - H_{ji}(t + d)\} \right| + |h_j(t) - h_j(t + d)| \right] = o_p(n^{-(1/2)-\eta_0})$$

and

$$\sup_{|d| \leq n^{-k_0}} \left[n^{-1} \left\| \sum_{i=1}^{n} \{\mathbf{Z}_i(t) - \mathbf{Z}_i(t + d)\} \right\| + \|\mathbf{s}^{(r)}(\beta_0, t) - \mathbf{s}^{(r)}(\beta_0, t + d)\| \right] = o_p(n^{-(1/2)-\eta_0}).$$

C. $\mathcal{E}(X_1^{\theta_0}) < \infty$ for some $\theta_0 > 0$.

Condition C is certainly satisfied in any reasonable setups. Conditions A and B require that $\mathbf{Z}$ and $\mathbf{H}$ do not have too many fluctuations; the usual time-dependent covariates and missing indicators satisfy both requirements. It is interesting to note that Conditions A–C are much simpler to interpret and to check than are the regularity conditions of Self and Prentice (1988) for the case-cohort design.

Our asymptotic theory for $\tilde{\beta}$ consists of two parts, presented in Theorems 2.1 and 2.2. The first part gives the asymptotic distribution of $\tilde{\mathbf{U}}(\beta_0)$ when it is regarded as a test statistic; the second part deals with the consistency and asymptotic normality of the proposed estimator $\tilde{\beta}$. The proofs of the theorems are relegated to the Appendix.

Theorem 2.1 The random vector $n^{-1/2}\tilde{\mathbf{U}}(\beta_0)$ is asymptotically normal with mean $\mathbf{0}$ and covariance matrix $\mathbf{B}(\beta_0) = \mathcal{E}\{\mathbf{W}_1(\beta_0)^{\otimes 2}\}$, where

$$\mathbf{W}_i(\beta) = \Delta_i \mathbf{H}_i(X_i)\{\mathbf{Z}_i(X_i) - \mathbf{e}(\beta, X_i)\} - \int_0^{X_i} \{\mathbf{h}(t)/h_0(t)\} H_{0i}(t) \exp\{\beta' \mathbf{Z}_i(t)\} \times \{\mathbf{Z}_i(t) - \mathbf{e}(\beta, t)\} \lambda_0(t)\, dt. \quad (5)$$

The covariance matrix $\mathbf{B}(\beta_0)$ can be consistently estimated by $\mathbf{B}_n(\beta_0) = n^{-1} \sum_{i=1}^{n} \hat{\mathbf{W}}_i(\beta_0)^{\otimes 2}$, where

$$\hat{\mathbf{W}}_i(\beta) = \Delta_i \mathbf{H}_i(X_i)\{\mathbf{Z}_i(X_i) - \mathbf{E}(\beta, X_i)\} - n^{-1} \sum_{l=1}^{n} \Delta_l Y_i(X_l) H_{0i}(X_l) \mathbf{H}_l(X_l) \exp\{\beta' \mathbf{Z}_i(X_l)\} \times \{\mathbf{Z}_i(X_l) - \mathbf{E}(\beta, X_l)\}/S^{(0)}(\beta, X_l). \quad (6)$$

In deriving the asymptotic distribution of $\tilde{\beta}$, we need to expand its estimating function $\tilde{\mathbf{U}}(\beta)$ at β_0. Thus the slope function plays an important role. Let $\mathbf{A}_n(\beta) = -n^{-1}\partial\tilde{\mathbf{U}}(\beta)/\partial\beta$, $\mathbf{A}(\beta) = \lim_{n\to\infty} \mathbf{A}_n(\beta)$, and $\tilde{\mathbf{u}}(\beta) = \lim_{n\to\infty} n^{-1}\tilde{\mathbf{U}}(\beta)$, where by ergodicity the limits exist almost surely. Note that $\tilde{\mathbf{u}}(\beta_0) = \mathbf{0}$.

Theorem 2.2 Suppose that $\mathbf{A}(\beta_0)$ is nonsingular. Then there exists a solution $\tilde{\beta}$ to $\{\tilde{\mathbf{U}}(\beta) = \mathbf{0}\}$ such that the random vector $n^{1/2}(\tilde{\beta} - \beta_0)$ is asymptotically normal with mean $\mathbf{0}$ and covariance matrix $\mathbf{A}^{-1}(\beta_0)\mathbf{B}(\beta_0)\mathbf{A}^{-1}(\beta_0)'$. Furthermore, if $\{\tilde{\mathbf{u}}(\beta) = \mathbf{0}\}$ has no roots other than β_0, then within any bounded region containing β_0 the equation $\{\tilde{\mathbf{U}}(\beta) = \mathbf{0}\}$ has a unique root for all large n. A consistent estimator of the limiting covariance matrix is $\mathbf{A}_n^{-1}(\tilde{\beta})\mathbf{B}_n(\tilde{\beta})\mathbf{A}_n^{-1}(\tilde{\beta})'$.

In many important settings, the assumptions that $\mathbf{A}(\beta_0)$ is nonsingular and that $\{\tilde{\mathbf{u}}(\beta) = \mathbf{0}\}$ has a unique root are equivalent to the usual mild condition that $\mathbf{A}_f(\beta) = \lim_{n\to\infty}\{-n^{-1}\partial\mathbf{U}_f(\beta)/\partial\beta\}$, a nonnegative definite matrix, is positive definite at β_0, which guarantees the consistency and asymptotic normality of the MPLE. For example, if $(H_{1i}, \ldots, H_{pi})$ $(i = 1, \ldots, n)$ are time-invariant, as is true for time-independent covariates and for case-cohort studies, then $\mathbf{h}(t) \equiv \mathbf{h}$ is also time-invariant. Because $\mathbf{E}(\beta, t)$ always converges to the same limit regardless of the patterns for $H_{0i}(\cdot)$, we have $\mathbf{A}(\beta) = \mathbf{h}\mathbf{A}_f(\beta)$ and $\tilde{\mathbf{u}}(\beta) = \mathbf{h}\mathbf{u}_f(\beta)$, where $\mathbf{u}_f(\beta) = \lim_{n\to\infty} n^{-1}\mathbf{U}_f(\beta)$. The equivalence of the assumptions is evident because $\mathbf{h}$ is nonsingular. As another example, if $H_{1i}(X_i) = \cdots = H_{pi}(X_i)$ for all the uncensored X_i, as is true for case-cohort designs, then $\mathbf{A}(\beta)$ is symmetric. In this scenario, it can be verified that $\rho\mathbf{A}(\beta) \leq \mathbf{A}_f(\beta) \leq \rho^{-1}\mathbf{A}(\beta)$ for some $0 < \rho \leq 1$. Clearly, the positive definiteness of $\mathbf{A}_f(\beta_0)$ ensures the positive definiteness of $\mathbf{A}(\beta_0)$, which in turn implies that $\{\tilde{\mathbf{u}}(\beta) = \mathbf{0}\}$ has a unique root.

For the case-cohort design, the variance estimator $\mathbf{A}_n^{-1}(\tilde{\beta})\mathbf{B}_n(\tilde{\beta})\mathbf{A}_n^{-1}(\tilde{\beta})'$ is much easier to calculate than the estimators of Prentice (1986) and Self and Prentice (1988), especially in the presence of time-dependent covariates. Another advantage of the proposed estimator is that its form

remains unchanged under multiple subcohort augmentations. Furthermore, incomplete covariate measurements on the cases are allowed.

The matrix $\mathbf{A}_n^{-1}(\tilde{\beta})\mathbf{B}_n(\tilde{\beta})\mathbf{A}_n^{-1}(\tilde{\beta})'$ reduces to the robust variance–covariance estimator of Lin and Wei (1989) when there are no missing covariate values. As in Lin and Wei (1989), we may also construct score-type statistics for testing subsets of β_0. Simulation results of Lin and Wei (1989) indicated that the score-type test preserves the size better than the Wald-type test in small samples. For simplicity, we will confine our attention to the Wald statistic in this article.

Self and Prentice (1988) provided the asymptotic efficiency results for case-cohort studies. We have investigated the asymptotic efficiency of the APLE $\tilde{\beta}$ for partially incomplete covariate vectors. It is instructive to consider the following example.

Example 2.1 Suppose that $\mathbf{Z}$ consists of two independent time-invariant covariates Z_1 and Z_2 and that the measurements are complete on Z_1 but incomplete on Z_2. For simplicity, assume that $\beta_0 = \mathbf{0}$ and $C = c$, with c being a fixed constant. Let $F_0(\cdot)$ denote the distribution function corresponding to the baseline hazard $\lambda_0(\cdot)$. It can be shown through some calculations of expectations and integrals that the limiting covariance matrix for $n^{1/2}\tilde{\beta}$ is

$$\mathbf{A}_f^{-1}\begin{bmatrix}1 & 0\\ 0 & h_2^{-1}\end{bmatrix}\begin{bmatrix}b_1 & 0\\ 0 & b_2\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & h_2^{-1}\end{bmatrix}\mathbf{A}_f^{-1},$$

where $\mathbf{A}_f = \operatorname{cov}(\mathbf{Z})F_0(c)$, $b_1 = \operatorname{var}(Z_1)(F_0(c) + 2h_2^{-1}(1 - h_2)[F_0(c) + \{1 - F_0(c)\}\log\{1 - F_0(c)\}])$, and $b_2 = \operatorname{var}(Z_2)h_2F_0(c)$. Because $\log(1 - x) < -x - .5x^2$ (for $0 < x < 1$), we have $b_1 < \operatorname{var}(Z_1)F_0(c)[1 + h_2^{-1}(1 - h_2)F_0(c) \times \{1 + F_0(c)\}]$. Comparing the right side of the preceding inequality with the variance expression for $n^{1/2}\hat{\beta}_1$ based on complete cases only, we see that the proposed estimator is more efficient provided that $1 + h_2^{-1}(1 - h_2)F_0(c) \times \{1 + F_0(c)\} \le h_2^{-1}$ or $F_0(c) \le (\sqrt{5} - 1)/2 \approx .618$. Furthermore, the variance of $n^{1/2}\tilde{\beta}_1$ decreases to that of $n^{1/2}\hat{\beta}_1$ with full covariate data as $h_2^{-1}(1 - h_2)F_0(c)\{1 + F_0(c)\}$ approaches 0. Note that $F_0(c)$ is the probability of experiencing the specified event by the end of the study. This probability is usually small in comparative clinical trials and epidemiologic studies (except for advanced cancers).

Table 1 tabulates the efficiency of $\tilde{\beta}_1$ for various combinations of h_2 and $F_0(c)$. The loss of efficiency for $\tilde{\beta}_1$ relative to $\hat{\beta}_1$ with full covariate measurements is minimal if $F_0(c)$ is small and h_2 is moderately large. The proposed method tends to be more efficient than the complete-case analysis especially for small h_2 and small $F_0(c)$, but it can be slightly less efficient than the latter if $F_0(c)$ is too large.

It is often desirable to estimate survival probabilities. Corresponding to the APLE $\tilde{\beta}$ is a natural estimator

$$\tilde{\Lambda}(\tilde{\beta}, t) = \sum_{i=1}^{n} \frac{I(X_i \le t)\Delta_i H_{0i}(X_i)}{nS^{(0)}(\tilde{\beta}, X_i)} \tag{7}$$

of the cumulative baseline hazard function $\Lambda_0(t) = \int_0^t \lambda_0(u) \times du$. The asymptotic properties of this estimator are stated in the following theorem; the proof is given in the Appendix.

Table 1. Asymptotic Relative Efficiency Results for Example 2.1

h_2	$F_0(c)$	Proposed vs. full data	Proposed vs. complete cases
.2	.05	.83	4.15
	.20	.54	2.69
	.50	.29	1.45
	.80	.17	.86
.5	.05	.95	1.90
	.20	.82	1.65
	.50	.62	1.24
	.80	.46	.91
.8	.05	.99	1.23
	.20	.95	1.19
	.50	.87	1.08
	.80	.77	.96

Theorem 2.3 The process $n^{1/2}\{\tilde{\Lambda}(\tilde{\beta}, \cdot) - \Lambda_0(\cdot)\}$ converges weakly to a 0 Gaussian process with mean 0 and covariance function

$$\psi(t, s) = \int_0^{\min(t,s)} \frac{d\Lambda_0(u)}{s^{(0)}(\beta_0, u)} + \mathbf{J}'(t)\mathbf{A}^{-1}(\beta_0)\mathbf{B}(\beta_0)\mathbf{A}^{-1}(\beta_0)'\mathbf{J}(s) - \mathbf{J}'(s)\mathbf{A}^{-1}(\beta_0)\mathbf{G}(t) - \mathbf{J}'(t)\mathbf{A}^{-1}(\beta_0)\mathbf{G}(s), \tag{8}$$

where

$$\mathbf{J}(t) = \int_0^t \frac{\mathbf{s}^{(1)}(\beta_0, u)\, d\Lambda_0(u)}{s^{(0)}(\beta_0, u)},$$

$$\mathbf{G}(t) = \mathcal{E}\left[\int_0^{\infty}\int_0^{\min(t,v)} \frac{H_{01}(u)\exp\{\beta_0'\mathbf{Z}_1(u)\}\, d\Lambda_0(u)}{s^{(0)}(\beta_0, u)} \times \left\{\mathbf{H}_1(v) - \frac{H_{01}(v)}{h_0(v)}\mathbf{h}(v)\right\}\{\mathbf{Z}_1(v) - \mathbf{e}(\beta_0, v)\} \times Y_1(v)\exp\{\beta_0'\mathbf{Z}_1(v)\}\, d\Lambda_0(v)\right].$$

Furthermore, ψ can be consistently estimated by replacing the unknown quantities in (8) by their respective sample estimators.

The estimator (7) may not be the best choice if $H_{0i}(X_i) = 0$ for most of the nonzero Δ_i's. For the original case-cohort design, formula (2.5) of Self and Prentice (1988) should be used instead. If the subcohort membership varies over time due to augmentations, then the following estimator is recommended:

$$\tilde{\Lambda}^*(\tilde{\beta}, t) = \sum_{i=1}^{n} \frac{I(X_i \le t)\Delta_i \sum_{l=1}^{n} H_{0l}(X_i)Y_l(X_i)}{nS^{(0)}(\tilde{\beta}, X_i)\sum_{l=1}^{n} Y_l(X_i)}.$$

It follows from the arguments given in the proof of Theorem 2.3 that the process $n^{1/2}\{\tilde{\Lambda}^*(\tilde{\beta}, \cdot) - \Lambda_0(\cdot)\}$ also converges weakly to a 0-mean Gaussian process. The limiting covariance function contains more terms than formula (8), though each term can be easily estimated. The details are available from the authors.

3. SIMULATION STUDIES

Extensive Monte Carlo studies were carried out to investigate the finite-sample behavior of $\tilde{\beta}$. We considered the

Table 2. Monte Carlo Estimates for the Sampling Means and Variances of the β_1 Estimators and for the Sizes of the Wald Tests at the 5% Level for Testing $H_0: \beta_1 = 0$ Under the Model $\lambda(t|Z_1, Z_2) = \exp(.5Z_2)$

			20% Missing			50% Missing			80% Missing		
n	*Censoring*	*Method*	*Mean*	*Var.*	*Size*	*Mean*	*Var.*	*Size*	*Mean*	*Var.*	*Size*
200	90%	a)	.0066	.055	.056	.0067	.071	.040	—	—	—
		b)	.0062	.068	.048	−.0008	.122	.062	—	—	—
		c)	.0050	.052	.043	.0050	.052	.043	.0050	.052	.043
		d)	.0056	.053	.047	.0056	.053	.047	.0056	.053	.047
200	50%	a)	.0018	.013	.040	.0026	.021	.053	.0028	.060	.056
		b)	.0034	.013	.044	.0028	.022	.054	.0125	.065	.060
		c)	.0002	.010	.045	.0002	.010	.045	.0002	.010	.045
		d)	.0011	.010	.047	.0011	.010	.047	.0011	.010	.047
500	90%	a)	−.0002	.021	.061	.0015	.024	.055	.0046	.043	.051
		b)	−.0003	.025	.053	.0048	.041	.053	.0131	.120	.048
		c)	.0008	.020	.055	.0008	.020	.055	.0008	.020	.055
		d)	.0000	.020	.054	.0000	.020	.054	.0000	.020	.054
500	50%	a)	−.0021	.005	.051	−.0014	.008	.048	−.0004	.020	.055
		b)	−.0026	.005	.044	−.0042	.008	.054	−.0044	.022	.054
		c)	−.0018	.004	.036	−.0018	.004	.036	−.0018	.004	.036
		d)	−.0015	.004	.045	−.0015	.004	.045	−.0015	.004	.045

NOTE: Z_1 and Z_2 are independent standard normal variables. The measurements of Z_1 are complete; the missing percentage pertains to Z_2. Method (a) refers to the approximate partial likelihood analysis, method (b) deletes all cases that are incomplete on Z_2, method (c) excludes Z_2 from model fitting, and method (d) shows the results that would be obtained if the missing values were observed. Each block is based on 1,000 replications. The random number generator of Wichmann and Hill (1982) is used. "—" indicates that no reliable estimates are available due to heavy censoring and heavy missingness.

Cox model with unit exponential baseline hazard and with two standard normal covariates, Z_1 and Z_2. Censorship was imposed by the generation of independent uniform random variables on the interval $(0, \theta)$, where θ was a suitably chosen real number so that observations in each simulation sample had a desired probability of being censored. Our primary interest was in making inference about β_1 when measurements are complete on Z_1 but incomplete on Z_2. For comparison, the method based on complete cases only and that of modeling Z_1 only were also evaluated. The full-data analysis, which uses the actual values of the "missing" covariate measurements, provided a natural benchmark for evaluating the proposed method and the two "naive" methods.

Tables 2 and 3 summarize the key results for independent Z_1 and Z_2. The following conclusions are based on these two tables and related studies:

1. The bias of the APLE is fairly small and decreases with the sample size. This is also true with respect to the bias of the variance estimator for the APLE.
2. The Wald test based on the APLE has adequate size.
3. As compared to the complete-case analysis, the proposed method has smaller variance for the parameter estimator and higher power for the Wald test. This is particularly evident when the missingness and censoring are heavy. These findings are consistent with the asymptotic relative efficiency results given in Example 2.1.
4. When Z_2 is omitted, $\hat{\beta}_1$ is biased towards 0, which is consistent with the analytical result of Struthers and Kalbfleisch (1986). The proposed Wald test is more powerful than the Wald test based on modeling Z_1 only if the extent of missingness is light and β_2 is large.

We also conducted extensive studies for correlated Z_1 and Z_2. The results indicated that conclusions 1–3 remain essentially valid when Z_1 and Z_2 are correlated. For the Z_1-

Table 3. Monte Carlo Estimates for the Sampling Means and Variances of the β_1 Estimators and for the Powers of the Wald Tests at the 5% Level for Testing $H_0: \beta_1 = 0$ Under the Model $\lambda(t|Z_1, Z_2) = \exp(.25Z_1 + .5Z_2)$

			20% Missing			50% Missing			80% Missing		
n	*Censoring*	*Method*	*Mean*	*Var.*	*Power*	*Mean*	*Var.*	*Power*	*Mean*	*Var.*	*Power*
200	90%	a)	.264	.056	.230	.272	.072	.192	—	—	—
		b)	.263	.068	.192	.265	.121	.125	—	—	—
		c)	.250	.051	.211	.250	.051	.211	.250	.051	.211
		d)	.259	.052	.231	.259	.052	.231	.259	.052	.231
200	50%	a)	.258	.013	.620	.266	.022	.435	.300	.063	.234
		b)	.258	.014	.616	.262	.023	.430	.282	.067	.193
		c)	.232	.010	.636	.232	.010	.636	.232	.010	.636
		d)	.256	.011	.721	.256	.011	.721	.256	.011	.721
500	90%	a)	.248	.021	.432	.254	.025	.378	.270	.044	.267
		b)	.246	.025	.350	.252	.041	.232	.264	.117	.113
		c)	.242	.019	.429	.242	.019	.429	.242	.019	.429
		d)	.248	.020	.453	.248	.020	.453	.248	.020	.453
500	50%	a)	.251	.005	.942	.254	.008	.815	.268	.021	.468
		b)	.250	.005	.938	.250	.009	.783	.252	.023	.400
		c)	.228	.004	.956	.228	.004	.956	.228	.004	.956
		d)	.250	.004	.974	.250	.004	.974	.250	.004	.974

NOTE: See the comments for Table 2.

only analysis, $\hat{\beta}_1$ is biased upwards when Z_1 and Z_2 are positively correlated and downwards when the two covariates are negatively correlated. The size of the associated (two-sided) Wald test always exceeds the nominal level regardless of the sign of the correlation. The bias and anti-conservativeness are quite disturbing even for a moderate correlation.

4. REAL EXAMPLES

We have applied the new methodology to a number of real data sets of the types of Examples 1.1–1.4 and obtained satisfactory results. Due to space limitation, only two examples are provided here. The interested reader may request a FORTRAN program from the first author.

4.1 AZT Trial

The clinical benefit of the anti-AIDS drug, zidovudine (AZT), was first demonstrated in a double-blind, placebo-controlled clinical trial conducted in 1986 (Fischl et al. 1987). Of the 281 patients enrolled in the study, 144 were assigned to AZT and 137 to placebo. By the time the study was terminated, 76 patients had developed at least one opportunistic infection.

The development of opportunistic infections is triggered by the depletion of CD4 cells. It is, therefore, natural to adjust for the effect of the baseline CD4 count when assessing the benefit of AZT. But not all subjects had CD4 counts taken near the beginning of the study. In our illustration, we regard baseline CD4 counts as missing if no measurements were made within 1 week before the study. By this definition, 12 patients had missing baseline CD4 counts.

Table 4 shows the results of Cox regression analyses based on various procedures. The method of imputation replaces missing CD4 counts by the measurements nearest to the baseline. The proposed method provides a little stronger evidence for the treatment success than the other methods.

Table 4. Cox Regression Analyses of Time to the First Opportunistic Infection for the AZT Trial

Method	Parameter	Est.	S.E.	Est./S.E.	P value
Treatment-only	Treatment	−.438	.123	−3.55	.0004
Imputation	Treatment	−.451	.123	−3.66	.00025
	log (CD4)	−.440	.081	−5.43	<.0001
Complete-case	Treatment	−.454	.126	−3.61	.0003
	log (CD4)	−.465	.084	−5.56	<.0001
Proposed	Treatment	−.459	.124	−3.71	.0002
	log (CD4)	−.466	.084	−5.55	<.0001

4.2 Welsh Nickel Refiners Study

Men employed in a nickel refinery in South Wales were investigated to determine the risk of developing carcinoma of the bronchi and nasal sinuses associated with the refining of nickel. The cohort was identified using the weekly payrolls of the company and followed from the year 1934 until 1981. Appendix VIII of Breslow and Day (1987) contained complete records for 679 workers employed before 1925, to whom attention is henceforth confined. The follow-up through 1981 uncovered 56 deaths from cancer of the nasal sinus.

Table 5. Cox Regression Analyses of Time From the First Employment to the Nasal Sinus Cancer Death for the Welsh Nickel Refiners Study

	Design		
Parameter	Full cohort	Case-cohort	Modified case-cohort
log(AFE-10)			
Est.	2.22	1.74	1.69
S.E.	.44	.48	.43
P value	<.00001	.0003	.00008
(YFE-1915)/10			
Est.	−.09	−.18	−.13
S.E.	.32	.43	.38
P value	.76	.68	.74
$(YFE-1915)^2/100$			
Est.	−1.26	−1.28	−1.31
S.E.	.51	.73	.56
P value	.013	.078	.021
log(EXP + 1)			
Est.	.77	.67	.65
S.E.	.17	.27	.23
P value	.00001	.015	.0057

Breslow and Day (1987, pp. 222–223) analyzed the mortality data on the nasal sinus cancer using the Cox model. These authors considered the survival time to be the years since first employment; they found three significant risk factors: AFE (age at first employment), YFE (year at first employment), and EXP (exposure level). Their final results are reproduced in Table 5 under the heading "Full cohort."

The third column in Table 5 displays the results from fitting the same model to data obtained from a "hypothetical" case-cohort design that randomly selected 100 subcohort members from the entire cohort. These results are similar to, but less convincing than, the original findings. For example, log(EXP + 1) is significant at the 5% level but not at the 1% level. Inspections of the risk sets revealed that the subcohort had fewer than 50 members for 20 of the 56 deaths and had sizes of 17, 15, 13, and 11 for the last 4 events. When 50 randomly chosen members were augmented to the subcohort immediately after its size fell below 50, the risk sets retained more than 20 subjects throughout. The resulting estimates are shown in the last column of Table 5. In terms of hypothesis testing, the basic conclusions from this modified case-cohort design are practically the same as those of the full-cohort analysis.

In order to obtain a more reliable assessment of the case-cohort and modified case-cohort designs, we fit the same model to 1,000 replications of the two designs that were generated by randomly selecting the subcohorts from the full data set. The average standard error estimates for log(EXP + 1) were .31 and .25 under the case-cohort and modified case-cohort designs. The null hypothesis of no log(EXP + 1) effect was rejected 921 and 985 times at the 5% significance level, and 688 and 892 times at the 1% significance level. The efficiency gains of the modified case-cohort design over the case-cohort design were also evident in the other three covariates.

5. REMARKS

In many observational studies, the failure time T may be subject to left truncation in the sense that T can be observed

only when it is larger than some time interval, say L. For the Welsh nickel refiners study, the workers had been working in the company for various periods before follow-up was initiated. To incorporate left truncation in our analysis, all we need to do is to redefine the risk indicators as $Y_i(t) = I(X_i \geq t \geq L_i)\,(i = 1, \ldots, n)$. In fact, this modification was made implicitly in Section 4.2.

Regression forms such as $\{1 + \beta_0'\mathbf{Z}(t)\}$ are more plausible than the exponential form in (1) for some applications. The results of Section 2 can be easily extended to general relative risk models. Extensions to other families of regression models are currently under investigation.

A key step in the proof of Theorem 2.1 is to express $\tilde{\mathbf{U}}(\beta_0)$ as a sum of iid random vectors plus some asymptotically negligible terms. The iid representation also facilitates direct extensions of our results to the setting of multivariate failure time data where marginal distributions are postulated with Cox models, although a slightly different representation is required if all marginal models assume a common baseline hazard function. The interested reader is referred to Wei, Lin, and Weissfeld (1989) and Lee, Wei, and Amato (1992) for analyses with full covariate data.

The validity of the APLE $\tilde{\beta}$ and the related inference procedures depends critically on the MCAR assumption. This assumption is satisfied in many randomized clinical trials and in situations where measurements are missing by design (e.g., case-cohort studies). A less stringent requirement is the missing at random (MAR) assumption (Rubin 1976), which states that the probability of missing on certain components of $\mathbf{Z}$ (say $\mathbf{Z}^*$) depends on some completely observed variables (say $\mathbf{Q}$) but not on $\mathbf{Z}^*$. In this case, if we divide the range of $\mathbf{Q}$ into appropriate strata, then the MCAR assumption may be reasonable within each stratum. It is straightforward to modify estimating function (4) for the stratified analysis. In some applications, the probability of missing on $\mathbf{Z}^*$ depends on $\mathbf{Z}^*$ and possibly on $\mathbf{Q}$ as well. This case is very hard to deal with, though recent developments in the area of informative censoring (e.g., Link 1989) might provide useful clues.

APPENDIX: PROOFS

The estimating function $\tilde{\mathbf{U}}(\beta)$ is considerably more complicated than the partial likelihood score function $\mathbf{U}_f(\beta)$ with full covariate measurements and the pseudolikelihood score function under the original case-cohort design. Most notably, $\tilde{\mathbf{U}}(\beta_0)$ is not a martingale integral, and $\tilde{\mathbf{U}}(\beta)$ may not be a concave function. Consequently, the techniques used by Andersen and Gill (1982), Lin and Wei (1989), and Self and Prentice (1988) are not sufficient for investigating the asymptotic behavior of $\tilde{\beta}$. Our approach is based on approximations to some basic weighted empirical processes, which are established in Lemmas 2 and 3. The proofs of the lemmas are omitted here due to space limitations, but they can be found in Lin and Ying (1992). In the sequel, $\bar{Y}(t) = \sum_{i=1}^n Y_i(t)$ and $N_i(t) = \Delta_i I(X_i \leq t)$. For a $p \times 1$ vector $\mathbf{a}$, a_j denotes its jth component. When its limits are not shown explicitly, the summation $\sum$ is taken from 1 to n.

Lemma 1. (a) Under Condition A, $\max_{1\leq i\leq n} \Lambda_0(X_i) = O_p(\log n)$; and (b) under Conditions A and C, $\max_{1\leq i\leq n} X_i = o_p(n^{k_1})$ for any $k_1 > \theta_0^{-1}$.

Lemma 2. Under Conditions A–C,

(a) for every $\varepsilon > 0$ and $j = 0, \ldots, p$,

$$\sup_t \left| \sum_{i=1}^n H_{ji}(t) e^{\beta_0'\mathbf{Z}_i(t)} Y_i(t) - \sum_{i=1}^n h_j(t) e^{\beta_0'\mathbf{Z}_i(t)} Y_i(t) \right| = o_p(n^{1/2+\varepsilon}); \text{ and}$$

(b) for every $\gamma \in (0, 1)$ and $j = 0, \ldots, p$, there exists $\varepsilon_0 > 0$ such that

$$\sup_{\bar{Y}(t)\leq n^\gamma} \left| \sum_{i=1}^n H_{ji}(t) e^{\beta_0'\mathbf{Z}_i(t)} Y_i(t) - \sum_{i=1}^n h_j(t) e^{\beta_0'\mathbf{Z}_i(t)} Y_i(t) \right| = o_p(n^{1/2-\varepsilon_0}).$$

Lemma 3. Under Conditions A–C,

(a) for every $\varepsilon > 0$ and $k = 0, 1$,

$$\sup_t \|\mathbf{S}^{(k)}(\beta_0, t) - \mathbf{s}^{(k)}(\beta_0, t)\| = o_p(n^{-1/2+\varepsilon}); \text{ and}$$

(b) for every $\gamma \in (0, 1)$ and $k = 0, 1$, there exists $\varepsilon_0 > 0$ such that

$$\sup_{\bar{Y}(t)\leq n^\gamma} \|\mathbf{S}^{(k)}(\beta_0, t) - \mathbf{s}^{(k)}(\beta_0, t)\| = o_p(n^{-1/2-\varepsilon_0}).$$

Proof of Theorem 2.1 Because $\sum_i Y_i(t) H_{0i}(t) e^{\beta'\mathbf{Z}_i(t)} \{Z_{ji}(t) - E_j(\beta, t)\} = 0$, we can rewrite the jth component of $\tilde{\mathbf{U}}(\beta_0)$ as

$$\begin{aligned}
\tilde{U}_j(\beta_0) &= \sum_{i=1}^n \int_0^\infty H_{ji}(t)\{Z_{ji}(t) - E_j(\beta_0, t)\} \\
&\quad \times \{dN_i(t) - e^{\beta_0'\mathbf{Z}_i(t)} Y_i(t)\lambda_0(t)\,dt\} \\
&\quad + \sum_{i=1}^n \int_0^\infty \{Z_{ji}(t) - E_j(\beta_0, t)\} e^{\beta_0'\mathbf{Z}_i(t)} Y_i(t) \\
&\quad \times \left\{H_{ji}(t) - \frac{h_j(t)}{h_0(t)} H_{0i}(t)\right\}\lambda_0(t)\,dt \\
&= \tilde{U}_{j1}(\beta_0) + \tilde{U}_{j2}(\beta_0), \quad \text{say.}
\end{aligned} \tag{A.1}$$

Now $\tilde{U}_{j1}(\beta_0)$ is a martingale integral, and so is

$$\bar{U}_{j1}(\beta_0) = \sum_{i=1}^n \int_0^\infty H_{ji}(t)\{Z_{ji}(t) - e_j(\beta_0, t)\} \times \{dN_i(t) - e^{\beta_0'\mathbf{Z}_i(t)} Y_i(t)\lambda_0(t)\,dt\}.$$

It follows from a standard calculation of the predictable variation of $\{\tilde{U}_{j1}(\beta_0) - \bar{U}_{j1}(\beta_0)\}$ that

$$\tilde{U}_{j1}(\beta_0) = \bar{U}_{j1}(\beta_0) + o_p(n^{1/2}). \tag{A.2}$$

Next we show that E_j in $\tilde{U}_{j2}(\beta_0)$ can also be replaced by e_j. Splitting the integral into two parts at $\tau_n = \inf\{t : \bar{Y}(t) < n^\gamma\}$, where $\frac{1}{2} < \gamma < 1$, we have

$$\begin{aligned}
&\left| \int_{\tau_n}^\infty \{E_j(\beta_0, t) - e_j(\beta_0, t)\} \right. \\
&\left. \times \sum_{i=1}^n e^{\beta_0'\mathbf{Z}_i(t)} Y_i(t) \left\{H_{ji}(t) - \frac{h_j(t)}{h_0(t)} H_{0i}(t)\right\}\lambda_0(t)\,dt \right| \\
&\leq 2K \max_{1\leq j\leq n} \Lambda_0(X_j) \\
&\quad \times \sup_{\bar{Y}(t)<n^\gamma} \left| \sum_{i=1}^n e^{\beta_0'\mathbf{Z}_i(t)} Y_i(t) \left\{H_{ji}(t) - \frac{h_j(t)}{h_0(t)} H_{0i}(t)\right\} \right| \\
&\leq 2K \max_{1\leq j\leq n} \Lambda_0(X_j) \sup_{\bar{Y}(t)<n^\gamma} \left[\left| \sum_{i=1}^n e^{\beta_0'\mathbf{Z}_i(t)} Y_i(t)\{H_{ji}(t) - h_j(t)\} \right| \right. \\
&\quad \left. + \frac{h_j(t)}{h_0(t)} \left| \sum_{i=1}^n e^{\beta_0'\mathbf{Z}_i(t)} Y_i(t)\{H_{0i}(t) - h_0(t)\} \right| \right] \\
&= o_p(n^{1/2}),
\end{aligned} \tag{A.3}$$

where the last equality follows from Lemma 1 (a) and Lemma 2 (b). For $t \le \tau_n$, Lemma 3 entails

$$\sup_{t\le\tau_n} |E_j(\beta_0, t) - e_j(\beta_0, t)| = o_p(n^{-\epsilon_0}) \quad \text{for some} \quad \epsilon_0 > 0. \quad \text{(A.4)}$$

This, in conjunction with Lemma 2 (a), can be used to show that

$$\left| \int_0^{\tau_n} \{E_j(\beta_0, t) - e_j(\beta_0, t)\} \sum_{i=1}^n e^{\beta_0' \mathbf{Z}_i(t)} Y_i(t) \times \left\{ H_{ji}(t) - \frac{h_j(t)}{h_0(t)} H_{0i}(t) \right\} \lambda_0(t)\, dt \right| = o_p(n^{1/2}). \quad \text{(A.5)}$$

By virtue of (A.1)–(A.3) and (A.5), $\tilde{U}_j(\beta_0) = \sum_{i=1}^n W_{ji}(\beta_0) + o_p(n^{1/2})$. Note that $\{\mathbf{W}_i(\beta_0)\}$ are iid random vectors with mean $\mathbf{0}$ and that $\max_{1\le i\le n} \|\mathbf{W}_i(\beta_0)\| = O_p(\log n)$ due to Lemma 1. It then follows from the multivariate central limit theorem and the Cramer–Wold device that $n^{-1/2}\tilde{\mathbf{U}}(\beta_0)$ is asymptotically normal with mean $\mathbf{0}$ and covariance matrix $\mathbf{B}(\beta_0)$.

To prove the consistency of the variance–covariance estimator $\mathbf{B}_n(\beta_0)$, it suffices to show that, by the law of large numbers and the Cauchy–Schwarz inequality,

$$n^{-1} \sum_{i=1}^n \{\hat{W}_{ji}(\beta_0) - W_{ji}(\beta_0)\}^2 \to 0 \quad \text{in probability.} \quad \text{(A.6)}$$

In view of (5) and (6),

$$\begin{aligned} &\hat{W}_{ji}(\beta_0) - W_{ji}(\beta_0) \\ &= \int_0^\infty \{e_j(\beta_0, t) - E_j(\beta_0, t)\} H_{ji}(t)\, dN_i(t) \\ &\quad + \int_0^\infty Z_{ji}(t) e^{\beta_0' \mathbf{Z}_i(t)} Y_i(t) H_{0i}(t) \\ &\quad \times \left\{ \frac{h_j(t)}{h_0(t)} \lambda_0(t)\, dt - \frac{\sum_l H_{jl}(t) dN_l(t)}{nS^{(0)}(\beta_0, t)} \right\} \\ &\quad + \int_0^\infty e^{\beta_0' \mathbf{Z}_i(t)} Y_i(t) H_{0i}(t) \left\{ E_j(\beta_0, t) \frac{\sum_l H_{jl}(t)\, dN_l(t)}{nS^{(0)}(\beta_0, t)} \right. \\ &\quad \left. - e_j(\beta_0, t) \frac{h_j(t)}{h_0(t)} \lambda_0(t)\, dt \right\} \\ &= D_{ji1} + D_{ji2} + D_{ji3}, \quad \text{say.} \end{aligned} \quad \text{(A.7)}$$

From (A.4) and the fact that $\sum_i \left[\int_{\tau_n}^\infty \{e_j(\beta_0, t) - E_j(\beta_0, t)\} H_{ji}(t) \times dN_i(t)\right]^2 = O_p(\bar{Y}(\tau_n)) = O_p(n^\gamma)$ $(\gamma < 1)$ follows the result that $n^{-1} \sum_i D_{ji1}^2 \to 0$ in probability. By similar but somewhat more elaborate arguments, we can show that $n^{-1} \sum_i (D_{ji2}^2 + D_{ji3}^2) \to 0$ in probability as well. The details are again given in Lin and Ying (1992). Hence (A.6) holds.

Proof of Theorem 2.2 Applying Lemmas 2 and 3 as in the proof of Theorem 2.1, we can show that the convergence of $\mathbf{A}_n(\beta)$ to $\mathbf{A}(\beta)$ is uniform for β in any compact region. This entails the continuity of $\mathbf{A}(\beta)$. For β close to β_0, the Taylor expansion yields $n^{-1/2}\tilde{\mathbf{U}}(\beta) = n^{-1/2}\tilde{\mathbf{U}}(\beta_0) - \mathbf{A}_n(\beta^*) n^{1/2}(\beta - \beta_0)$, where β^* is on the line segment between β and β_0. Because $\mathbf{A}(\beta_0)$ is nonsingular, $\mathbf{A}(\beta)$ is also nonsingular for β in some neighborhood of β_0 and so is $\mathbf{A}_n(\beta^*)$ for all large n. Inverting $\mathbf{A}_n(\beta^*)$ and applying Theorem 2.1, we get the desired existence and asymptotic normality for $\tilde{\beta}$. Furthermore, if $\tilde{\mathbf{u}}$ has no roots other than β_0, then any root of $\tilde{\mathbf{U}}$ has to be consistent. But the nonsingularity of $\mathbf{A}(\beta_0)$ ensures that for large n there is a unique root of $\tilde{\mathbf{U}}$ in a small neighborhood of β_0. For the consistency of the variance estimator, all we need to do is to show that $\mathbf{B}_n(\tilde{\beta}) - \mathbf{B}_n(\beta_0) \to \mathbf{0}$ and $\mathbf{A}_n(\tilde{\beta}) - \mathbf{A}_n(\beta_0) \to \mathbf{0}$ in probability. These two convergence results can be easily verified by taking the Taylor expansions of $\mathbf{B}_n(\tilde{\beta})$ and $\mathbf{A}_n(\tilde{\beta})$ at β_0. The details are omitted.

Proof of Theorem 2.3 By the Taylor expansion of $\tilde{\Lambda}(\tilde{\beta}, t)$ at β_0,

$$\begin{aligned} &n^{1/2}\{\tilde{\Lambda}(\tilde{\beta}, t) - \Lambda_0(t)\} \\ &= n^{1/2}\{\tilde{\Lambda}(\beta_0, t) - \Lambda_0(t)\} + n^{1/2}\mathbf{J}'(t)(\tilde{\beta} - \beta_0) + o_p(1), \end{aligned} \quad \text{(A.8)}$$

which implies the tightness of $n^{1/2}\{\tilde{\Lambda}(\tilde{\beta}, \cdot) - \Lambda_0(\cdot)\}$. Because $n^{1/2}(\tilde{\beta} - \beta_0) = n^{-1/2}\mathbf{A}^{-1}(\beta_0) \sum_i \mathbf{W}_i(\beta_0) + o_p(1)$, the right side of (A.8) is asymptotically equivalent to

$$\begin{aligned} &n^{1/2}\{\tilde{\Lambda}(\beta_0, t) - \Lambda_0(t)\} + n^{-1/2}\mathbf{J}'(t)\mathbf{A}^{-1}(\beta_0) \\ &\quad \times \sum_{i=1}^n \int_0^\infty \mathbf{H}_i(u)\{\mathbf{Z}_i(u) - \mathbf{e}(\beta_0, u)\} \\ &\quad \times \{dN_i(u) - e^{\beta_0' \mathbf{Z}_i(u)} Y_i(u) \lambda_0(u)\, du\} \\ &\quad + n^{-1/2}\mathbf{J}'(t)\mathbf{A}^{-1}(\beta_0) \sum_{i=1}^n \int_0^\infty \left\{ \mathbf{H}_i(u) - \frac{H_{0i}(u)}{h_0(u)} \mathbf{h}(u) \right\} \\ &\quad \times \{\mathbf{Z}_i(u) - \mathbf{e}(\beta_0, u)\} e^{\beta_0' \mathbf{Z}_i(u)} Y_i(u) \lambda_0(u)\, du \\ &= R_1(t) + R_2(t) + R_3(t), \quad \text{say.} \end{aligned}$$

It follows from a simple calculation of predictable covariations that $R_1(\cdot)$ and $R_2(\cdot)$ are asymptotically uncorrelated. Thus the asymptotic covariance function $\psi(t, s)$ is the limit of

$$\begin{aligned} &\operatorname{cov}\{R_1(t), R_1(s)\} + \operatorname{cov}\{R_2(t) + R_3(t), R_2(s) + R_3(s)\} \\ &\qquad + \operatorname{cov}\{R_1(t), R_3(s)\} + \operatorname{cov}\{R_1(s), R_3(t)\}. \end{aligned}$$

Evaluations of the four covariances in the preceding expression lead to formula (8). The consistency of the covariance function estimator can be proved in the same way as in the proof of Theorem 2.1.

[Received March 1992. Revised August 1992.]

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.

Breslow, N. E., and Day, N. E. (1987), *Statistical Methods in Cancer Research, Vol. II: The Design and Analysis of Cohort Studies,* Lyon, France: IARC.

Chyou, P. H., Nomura, A. M., Hankin, J. H., and Stemmermann, G. N. (1990), "A Case-Cohort Study of Diet and Stomach Cancer," *Cancer Research,* 50, 7501–7504.

Cox, D. R. (1972), "Regression Models and Life-Tables" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 34, 187–220.

Fischl, M. A., et al. (1987), "The Efficacy of Azidothymidine (AZT) in the Treatment of Patients with AIDS and AIDS-Related Complex," *New England Journal of Medicine,* 317, 185–191.

Lee, E. W., Wei, L. J., and Amato, D. A. (1992), "Cox-Type Regression Analysis for Large Numbers of Small Groups of Correlated Failure Time Observations," in *Survival Analysis: State of the Art,* eds. J. P. Klein and P. K. Goel. pp. 237–247. Dordrecht: Kluwer Academic Publishers.

Lin, D. Y., and Wei, L. J. (1989), "The Robust Inference for the Cox Proportional Hazards Model," *Journal of the American Statistical Association,* 84, 1074–1078.

Lin, D. Y., and Ying, Z. (1992), "Cox Regression with Incomplete Covariate Measurements," Technical Report 112, University of Washington, Dept. of Biostatistics.

Link, W. A. (1989), "A Model for Informative Censoring," *Journal of the American Statistical Association,* 84, 749–752.

Overvad, K., et al. (1991), "Selenium in Human Mammary Carcinogenesis: A Case-Cohort Study," *European Journal of Cancer,* 27, 900–902.

Prentice, R. L. (1986), "A Case-Cohort Design for Epidemiologic Cohort Studies and Disease Prevention Trials," *Biometrika,* 73, 1–11.

Rubin, D. B. (1976), "Inference and Missing Data," *Biometrika,* 63, 581–592.

Self, S. G., and Prentice, R. L. (1988), "Asymptotic Distribution Theory and Efficiency Results for Case-Cohort Studies," *The Annals of Statistics,* 16, 64–81.

Sorensen, T. I., and Sonne-Holm, S. (1988), "Risk in Childhood of Development of Severe Adult Obesity: Retrospective, Population-Based Case-Cohort Study," *American Journal of Epidemiology,* 127, 104–113.
Struthers, C. A., and Kalbfleisch, J. D. (1986), "Misspecified Proportional Hazards Models," *Biometrika,* 73, 363–369.
Tsiatis, A. A. (1981), "A Large-Sample Study of Cox's Regression Model," *The Annals of Statistics,* 9, 93–108.
Wacholder, S., Gail, M. H., Pee, D., and Brookmeyer, R. (1989), "Alternative Variance and Efficiency Calculations for the Case-Cohort Design," *Biometrika,* 76, 117–123.
Wei, L. J., Lin, D. Y., and Weissfeld, L. (1989), "Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions," *Journal of the American Statistical Association,* 84, 1065–1073.
Wichmann, B. A., and Hill, I. D. (1982), "An Efficient and Portable Pseudo-Random Number Generator," *Applied Statistics,* 31, 188–190.