Statistica Sinica **23** (2013), 1155-1180
doi:http://dx.doi.org/10.5705/ss.2012.117

A GENERAL SEMIPARAMETRIC Z-ESTIMATION APPROACH FOR CASE-COHORT STUDIES

Bin Nan and Jon A. Wellner

University of Michigan and University of Washington

Abstract: Case-cohort design, an outcome-dependent sampling design for censored survival data, is increasingly used in biomedical research. The development of asymptotic theory for a case-cohort design in the current literature primarily relies on counting process stochastic integrals. Such an approach, however, is rather limited and lacks theoretical justification for outcome-dependent weighted methods due to non-predictability. Instead of stochastic integrals, we derive asymptotic properties for case-cohort studies based on a general Z-estimation theory for semiparametric models with bundled parameters using empirical process theory. Both the Cox model and the additive hazards model with time-dependent covariates are considered.

Key words and phrases: Additive hazards model, bundled parameters, case-cohort study, Cox model, Donsker class, empirical process, Glivenko-Cantelli class, missing covariates, semiparametric estimation function, Z-estimation.

1. Introduction

Case-cohort designs, originally proposed by Prentice (1986) for right-censored survival data, are very useful in large epidemiologic cohort studies, and their applications are increasingly common in biomedical research. In a case-cohort study, complete data are only obtained for all failures observed during follow-up and for a sub-sample, called the subcohort, of the entire cohort. The subcohort can be a simple random or stratified sub-sample. Such a design is cost-effective for studies of rare events, and has been extended to other models including the additive hazards model (Kulich and Lin (2000)), transformation models (Chen and Zucker (2009); Kong, Cai, and Sen (2004); Lu and Tsiatis (2006)), and the accelerated failure time model (Nan, Kalbfleisch, and Yu (2009); Nan, Yu, and Kalbfleisch (2006)), and also to other censoring mechanisms (Li, Gilbert, and Nan (2008); Li and Nan (2011)), among many others.

For right-censored data, the pseudo likelihood approach of Self and Prentice (1988) constructs risk sets from subcohort only, thus the counting process martingale theory is naturally applicable for deriving the asymptotic properties for the Cox-type regression models. This same strategy can be applied to some

other regression models for right-censored data, for example, the accelerated failure time model studied by Nan, Yu, and Kalbfleisch (2006). Since complete information is also observed for all the failures, constructing risk sets from all observed data including failures outside the subcohort would yield more efficient estimation. This has been observed by many authors particularly through extensive simulations, for example, Borgan et al. (2000); Chen and Lo (1999); Chen and Zucker (2009); Kalbfleisch and Lawless (1988); Kulich and Lin (2000, 2004). The development of corresponding asymptotic theories has been primarily based on calculations of counting process stochastic integrals. Such a method, however, lacks theoretical justification because the integrands of those stochastic integrals are not predicable, not even adapted with respect to any filtration generated from the history.

To overcome this technical hurdle, we consider a general semiparametric Z-estimation method for *bundled parameters* using empirical process theory, see e.g., van der Vaart and Wellner (1996, 2007). Our approach does not use the stochastic integral formulation, thus there is no predictability requirement. The main body of the article is as follows. In Section 2, we introduce a general asymptotic theory for semiparametric Z-estimation with bundled parameters. We then apply the Z-estimation theory to case-cohort studies in Section 3. Both the Cox model and the additive hazards model with time-dependent covariates will be considered. We make some concluding remarks in Section 4. Detailed proofs are provided in Appendices A and B.

2. Semiparametric Z-estimation for Bundled Parameters

Let $\theta \in \Theta \subset \mathbb{R}^d$ be the parameter of interest, and $\eta : \mathcal{X} \times \Theta \to \mathbb{R}^J$ be infinite dimensional nuisance parameter(s) in a Banach space $\mathcal{H} \equiv \{(x, \theta) \mapsto \eta(x, \theta) \in \mathbb{R}^J : x \in \mathcal{X}, \theta \in \Theta\}$. Such a parametrization allows the nuisance parameter to be a function of the parameter of interest, thus the two types of parameters are *bundled together*, a terminology originally used by Huang and Wellner (1997) and further studied by, for example, Ding and Nan (2011). Denote the random map $\mathcal{X}^n \mapsto \mathbb{R}^d$ with n observations $X_1, \ldots, X_n$ as

$$\Psi_n(\theta, \eta) \;\equiv\; \Psi_n(X_1, \ldots, X_n; \theta, \eta(\cdot; \theta)) \,. \tag{2.1}$$

This becomes an estimating function for θ when η is given or replaced by its estimator. For independent and identically distributed (i.i.d.) observations $X_1, \ldots, X_n$, very often $\Psi_n(\theta, \eta)$ takes the form:

$$\Psi_n(\theta, \eta) \;=\; \frac{1}{n} \sum_{i=1}^{n} \psi(X_i; \theta, \eta(\cdot; \theta)), \tag{2.2}$$

where $\psi(\theta,\eta) \equiv \psi(X;\theta,\eta(\cdot;\theta))$ is a random map $\mathcal{X} \mapsto \mathbb{R}^d$ with a single observation X.

Here we use the term "nuisance parameter" in a rather loose sense. It does not need to be an actual parameter (for example, the baseline hazard function in the Cox model) in the original parametrization of the distribution of X. Broadly speaking, it is an unknown quantity in the estimating function in addition to the parameter of interest. The unknown quantity η as a function of θ needs to be estimated prior to estimating θ. We call the solution to $\Psi_n(\theta,\hat{\eta}_n(\cdot;\theta))=0$ the Z-estimator for θ, where $\hat{\eta}_n$ is some estimator for η. This type of generalization has been considered in the econometrics literature; see for example, Newey (1994); Chen, Linton, and Van Keilegom (2003). We provide slightly modified results of Chen, Linton, and Van Keilegom (2003) with a focus on Z-estimation in the following lemmas, which are used for the estimates in the considered case-cohort studies. Proofs of the lemmas are provided in Appendix A.

Let θ_0 denote the true value of θ and η_0 be the true functional form of η. Let $\Psi(\theta,\eta)$ be a deterministic function, which usually denotes the limit of $\Psi_n(\theta,\eta)$ as $n\to\infty$. We use p^* to denote "in outer probability", where the *outer probability* P^* is defined as $P^*(B)=\inf\{P(A): A\supset B, A\in\mathcal{A}\}$ for any subset B of Ω in a probability space $(\Omega,\mathcal{A},P)$. We refer its detailed discussion to van der Vaart and Wellner (1996). Note that the lemmas in this section do not require i.i.d. data, though data in the case-cohort studies we consider are assumed to be i.i.d. Let $|\cdot|$ be the Euclidian norm and let $\|\eta\|=\sup_{\theta\in\Theta}\rho(\eta(\cdot;\theta))$ for some norm or semi-norm ρ; for example, $\rho(\eta(\cdot;\theta))=\sup_{x\in\mathcal{X}}|\eta(x;\theta)|$, which gives $\|\eta\|=\sup_{\theta\in\Theta}\sup_{x\in\mathcal{X}}|\eta(x;\theta)|$.

Lemma 1 (Consistency)**.** *Suppose θ_0 is the unique solution to $\Psi(\theta,\eta_0(\cdot;\theta))=0$ in the parameter space Θ, and $\hat{\eta}_n$ is an estimator of η_0 such that $\|\hat{\eta}_n-\eta_0\|=o_{p^*}(1)$. If*

$$\sup_{\theta\in\Theta,\|\eta-\eta_0\|\le\delta_n}\frac{|\Psi_n(\theta,\eta(\cdot;\theta))-\Psi(\theta,\eta_0(\cdot;\theta))|}{1+|\Psi_n(\theta,\eta(\cdot;\theta))|+|\Psi(\theta,\eta_0(\cdot;\theta))|}=o_{p^*}(1) \tag{2.3}$$

for every sequence $\{\delta_n\}\downarrow 0$, then $\hat{\theta}_n$ satisfying $\Psi_n(\hat{\theta}_n,\hat{\eta}_n(\cdot;\hat{\theta}_n))=o_{p^*}(1)$ converges in outer probability to θ_0.

Consistency is a global property. Our main condition (2.3) is therefore necessarily global. The p^* in (2.3) indicates that the left-hand side converges to 0 in outer probability in case the term on the left is not Borel measurable. It is a stronger condition to require that the convergence holds when the denominator is replaced by 1. The purpose of adding an extra term in the denominator is to control the numerator when it blows up to infinity for some $\theta\in\Theta$.

Lemma 2 (Rate of convergence and asymptotic representation)**.** *Let* $\mathcal{H}_0 = \{\eta(x;\theta) : x \in \mathcal{X}, \theta \in \Theta_0\}$ *be a collection of sets of* J *functions that are continuously differentiable in* θ *for all* $x \in \mathcal{X}$ *with bounded derivative matrices* $\{\dot\eta(\cdot;\theta)\}$*, where* $\Theta_0 \subset \Theta$ *is a neighborhood of* θ_0*. Suppose that* $\hat\theta_n$ *satisfying* $\Psi_n(\hat\theta_n, \hat\eta_n(\cdot;\hat\theta_n)) = o_{p^*}(n^{-1/2})$ *is a consistent estimator of* θ_0 *that is the unique solution to* $\Psi(\theta, \eta_0(\cdot;\theta)) = 0$ *in* Θ*, and that* $\hat\eta_n \in \mathcal{H}_0$ *is an estimator of* $\eta_0 \in \mathcal{H}_0$ *satisfying* $\|\hat\eta_n - \eta_0\| = O_{p^*}(n^{-\beta})$ *for some* $\beta > 0$*. Suppose the following four conditions are satisfied.*

(i) (*Stochastic equicontinuity.*)

$$\frac{|n^{1/2}(\Psi_n - \Psi)(\hat\theta_n, \hat\eta_n(\cdot;\hat\theta_n)) - n^{1/2}(\Psi_n - \Psi)(\theta_0, \eta_0(\cdot;\theta_0))|}{1 + n^{1/2}|\Psi_n(\hat\theta_n, \hat\eta_n(\cdot;\hat\theta_n))| + n^{1/2}|\Psi(\hat\theta_n, \hat\eta_n(\cdot;\hat\theta_n))|} = o_{p^*}(1)\ .$$

(ii) $n^{1/2}\Psi_n(\theta_0, \eta_0(\cdot;\theta_0)) = O_{p^*}(1)$.

(iii) (*Smoothness.*) (a) *If* $\beta = 1/2$*, the function* $\Psi(\theta, \eta(\cdot;\theta)) : \Theta_0 \times \mathcal{H}_0 \to \mathbb{R}^d$ *is Fréchet differentiable at* $(\theta_0, \eta_0(\cdot;\theta_0))$*, i.e., there exists a continuous* $d \times d$ *matrix* $\dot\Psi_1(\theta_0, \eta_0(\cdot;\theta_0))$ *and continuous linear functionals* $\dot\Psi_2(\theta_0, \eta_0(\cdot;\theta_0))$ *with* $\dot\Psi_2[a] = \sum_{j=1}^J \dot\Psi_{2_j}[a]$ *such that*

$$\begin{aligned} &|\Psi(\theta, \eta(\cdot;\theta)) - \Psi(\theta_0, \eta_0(\cdot;\theta_0)) \\ &\qquad - \{\dot\Psi_1(\theta_0, \eta_0(\cdot;\theta_0)) + \dot\Psi_2(\theta_0, \eta_0(\cdot;\theta_0))[\dot\eta_0(\cdot;\theta_0)]\}(\theta - \theta_0) \\ &\qquad - \dot\Psi_2(\theta_0, \eta_0(\cdot;\theta_0))[(\eta - \eta_0)(\cdot;\theta_0)]| \\ &\quad = o(|\theta - \theta_0|) + o(\|\eta - \eta_0\|)\,; \end{aligned} \tag{2.4}$$

or (b) *if* $0 < \beta < 1/2$*, for some* $\alpha > 1$ *satisfying* $\alpha\beta > 1/2$ *we have*

$$\begin{aligned} &|\Psi(\theta, \eta(\cdot;\theta)) - \Psi(\theta_0, \eta_0(\cdot;\theta_0)) \\ &\qquad - \{\dot\Psi_1(\theta_0, \eta_0(\cdot;\theta_0)) + \dot\Psi_2(\theta_0, \eta_0(\cdot;\theta_0))[\dot\eta_0(\cdot;\theta_0)]\}(\theta - \theta_0) \\ &\qquad - \dot\Psi_2(\theta_0, \eta_0(\cdot;\theta_0))[(\eta - \eta_0)(\cdot;\theta_0)]| \\ &\quad = o(|\theta - \theta_0|) + O(\|\eta - \eta_0\|^\alpha)\,. \end{aligned} \tag{2.5}$$

Here the subscripts 1 and 2 correspond to the first and the second arguments in $\Psi(\cdot,\cdot)$*, respectively, and we assume that the matrix* $A = -\dot\Psi_1(\theta_0, \eta_0(\cdot;\theta_0)) - \dot\Psi_2(\theta_0, \eta_0(\cdot;\theta_0))[\dot\eta_0(\cdot;\theta_0)]$ *is nonsingular.*

(iv) $n^{1/2}\dot\Psi_2(\theta_0, \eta_0(\cdot;\theta_0))[(\hat\eta_n - \eta_0)(\cdot;\theta_0)] = O_{p^*}(1)$.

Then $\hat\theta_n$ *is* $n^{1/2}$*-consistent, and further we have*

$$\begin{aligned} n^{1/2}(\hat\theta_n - \theta_0) = A^{-1}n^{1/2}\Big\{&(\Psi_n - \Psi)(\theta_0, \eta_0(\cdot;\theta_0)) \\ &+ \dot\Psi_2(\theta_0, \eta_0(\cdot;\theta_0))[(\hat\eta_n - \eta_0)(\cdot;\theta_0)]\Big\} + o_{p^*}(1). \end{aligned} \tag{2.6}$$

Remark. For i.i.d. data, Condition (i) in Lemma 2 holds if the class of functions $\{\psi(\theta,\eta) : |\theta-\theta_0| < \delta, \|\eta-\eta_0\| < \delta\}$ is Donsker for some $\delta > 0$, and satisfies $E_0|\psi(X;\theta,\eta) - \psi(X;\theta_0,\eta_0)|^2 \to 0$ as $|\theta-\theta_0| \to 0$ and $\|\eta-\eta_0\| \to 0$ (see e.g., Corollary 2.3.12 of (van der Vaart and Wellner, 1996, p.115)). Though simpler, this is stronger than Condition (i). Condition (ii) holds automatically for i.i.d. data if $E_0|\psi(\theta_0,\eta_0)|^2 < \infty$ and Ψ_n takes the form (2.2). In Condition (iii), $\{\dot\Psi_1(\theta_0,\eta_0(\cdot;\theta_0)) + \dot\Psi_2(\theta_0,\eta_0(\cdot;\theta_0))[\dot\eta_0(\cdot;\theta_0)]\}(\theta-\theta_0)$ is obtained by the chain rule, which is the usual inner product of a $d\times d$ matrix and a $d\times 1$ vector; whereas $\dot\Psi_2(\theta_0,\eta_0(\cdot;\theta_0))[(\eta-\eta_0)(\cdot;\theta_0)] = \sum_{j=1}^J \dot\Psi_{2_j}(\theta_0,\eta_0(\cdot;\theta_0))[(\eta_j-\eta_{0j})(\cdot;\theta_0)]$, here J is the number of infinite dimensional parameters contained in η, is the sum of separate terms with each $\dot\Psi_{2_j}$ being a bounded linear functional that brings $\eta-\eta_0$ to a real number, where η is close to η_0 in n^β-rate for some $\beta > 0$. Proposition 1 of Bickel et al. (1993), page 455, provides useful tools for checking Fréchet differentiability for infinite-dimensional parameters. Condition (iv) holds automatically under (i)−(iii) if $\hat\eta_n$ is $n^{1/2}$-consistent, but may require extensive work for slower than root-n convergence rate, see e.g., Wong and Severini (1991) and Huang and Wellner (1995). In view of the structure of (2.6), the asymptotic distribution of $n^{1/2}(\hat\theta_n-\theta_0)$ is determined by the asymptotic joint distribution of the random variables $n^{1/2}(\Psi_n-\Psi)(\theta_0,\eta_0(\cdot;\theta_0))$ and $n^{1/2}\dot\Psi_2(\theta_0,\eta_0(\cdot;\theta_0))[(\hat\eta_n-\eta_0)(\cdot;\theta_0)]$, particularly if the asymptotic joint distribution is multivariate Gaussian.

In the case that η is free of θ, we have $\dot\eta = 0$. Then Lemma 2 reduces to the following corollary that was studied by Hu (1998). The corollary is particularly useful for the case-cohort additive hazards model in the next section. Now we replace $\dot\Psi_1$ by $\dot\Psi_\theta$ and $\dot\Psi_2$ by $\dot\Psi_\eta$ without causing any confusion, and the notation $\|\cdot\|$ becomes a norm.

Corollary 1 (Rate of convergence and asymptotic representation)**.** *Suppose that $\hat\theta_n$ satisfying $\Psi_n(\hat\theta_n,\hat\eta_n) = o_{p^*}(n^{-1/2})$ is a consistent estimator of θ_0 that is the unique solution to $\Psi(\theta,\eta_0) = 0$ in Θ, and that $\hat\eta_n$ is an estimator of η_0 satisfying $\|\hat\eta_n-\eta_0\| = O_{p^*}(n^{-\beta})$ for some $\beta > 0$. Suppose the following four conditions are satisfied.*

(i) (*Stochastic equicontinuity.*)

$$\frac{|n^{1/2}(\Psi_n-\Psi)(\hat\theta_n,\hat\eta_n) - n^{1/2}(\Psi_n-\Psi)(\theta_0,\eta_0)|}{1+n^{1/2}|\Psi_n(\hat\theta_n,\hat\eta_n)| + n^{1/2}|\Psi(\hat\theta_n,\hat\eta_n)|} = o_{p^*}(1)\ .$$

(ii) $n^{1/2}\Psi_n(\theta_0,\eta_0) = O_{p^*}(1)$.

(iii) (*Smoothness.*) (a) *If $\beta = 1/2$, function $\Psi(\theta,\eta)$ is Fréchet differentiable at (θ_0,η_0), i.e., there exists a continuous and nonsingular $d\times d$ matrix $\dot\Psi_\theta(\theta_0,\eta_0)$*

and a continuous linear functional $\dot{\Psi}_\eta(\theta_0, \eta_0)$ *such that*

$$\begin{aligned} &|\Psi(\theta,\eta) - \Psi(\theta_0,\eta_0) - \dot{\Psi}_\theta(\theta-\theta_0) - \dot{\Psi}_\eta(\theta_0,\eta_0)[\eta-\eta_0]| \\ &\quad = o(|\theta-\theta_0|) + o(\|\eta-\eta_0\|); \end{aligned} \tag{2.7}$$

or (b) *if* $0 < \beta < 1/2$, *for some* $\alpha > 1$ *satisfying* $\alpha\beta > 1/2$ *we have*

$$\begin{aligned} &|\Psi(\theta,\eta) - \Psi(\theta_0,\eta_0) - \dot{\Psi}_\theta(\theta-\theta_0) - \dot{\Psi}_\eta(\theta_0,\eta_0)[\eta-\eta_0]| \\ &\quad = o(|\theta-\theta_0|) + O(\|\eta-\eta_0\|^\alpha). \end{aligned} \tag{2.8}$$

(iv) $n^{1/2}\dot{\Psi}_\eta(\theta_0,\eta_0)[\hat{\eta}_n - \eta_0] = O_{p^*}(1)$.

Then $\hat{\theta}_n$ *is* $n^{1/2}$*-consistent, and further we have*

$$\begin{aligned} n^{1/2}(\hat{\theta}_n - \theta_0) = &\Big\{ -\dot{\Psi}_\theta(\theta_0,\eta_0)\Big\}^{-1} n^{1/2}\Big\{(\Psi_n - \Psi)(\theta_0,\eta_0) + \dot{\Psi}_\eta(\theta_0,\eta_0)[\hat{\eta}_n - \eta_0]\Big\} \\ &+ o_{p^*}(1). \end{aligned} \tag{2.9}$$

3. Case-Cohort Studies

We consider two models that are used for analyzing case-cohort data: the Cox model and the additive hazards model. Let X be the generic random variable that consists of several random variables. Let T be the failure time and C the censoring time; we only observe $Y = \min(T, C)$ and the failure indicator $\Delta = 1(T \le C)$. Let $Z(\cdot)$ be the d-dimensional covariate process and $\bar{Z}(t)$ be the covariate history up to time t. We assume that for all t, events $\{T \ge t\}$ and $\{C \ge t\}$ are conditionally independent given $\bar{Z}(t)$, and both are independent of $\{\bar{Z}(s) : s > t\}$. In other words, $Z(\cdot)$ is an external covariate, see Kalbfleisch and Prentice (2002). Suppose potentially we would have n i.i.d. copies of $(Y, \Delta, \bar{Z}(Y))$ in the full cohort, but we only observe $\bar{Z}(Y)$ for all failures and subjects in the subcohort that is a sub-sample of the entire cohort. The subcohort may be selected using a variety of sampling schemes including simple random sampling and stratified sampling based on some auxiliary variable $Z^*(\cdot)$ that can be a subset of $Z(\cdot)$, may or may not be time-dependent, and is available to everyone in the cohort. We focus on the independent Bernoulli sampling method for selecting the subcohort, by which a coin is flipped for each subject i in the cohort with a given success probability π_i that may depend on Z_i^*. For finite population sampling methods, as applied in Breslow and Wellner (2007), we expect the weighted bootstrap empirical process theory of Præstgaard and Wellner (1993) to be a useful tool to verify conditions in Lemmas 1 and 2. See Saegusa and Wellner (2012) for a related problem using weighted bootstrap empirical process theory.

Let R_i be the subcohort indicator that equals 1 if the ith subject is selected into the subcohort and 0 otherwise. Then $\pi_i = P(R_i = 1|Z_i^*)$. Thus the observed data in such a case-cohort study are i.i.d. and the missing data mechanism is missing at random (Little and Rubin (2002)). The following is a set of common regularity conditions for both the Cox model and the additive hazards model.

Assumption (A): The sample paths of $Z(\cdot) \in \mathcal{Z}$ are bounded with bounded variation, and the parameter space Θ is compact.

Assumption (B): The conditional distribution of T given $\bar{Z}(\cdot)$ possesses a continuous Lebesgue density.

Assumption (C): The study stops at a finite time $\tau > 0$ such that, for constants σ_1 and σ_2, $\inf_{z\in\mathcal{Z}} P(C \geq \tau|\bar{Z}(\tau) = \bar{z}(\tau)) = \sigma_1 > 0$ and $\inf_{z\in\mathcal{Z}} P(T > \tau|\bar{Z}(\tau) = \bar{z}(\tau)) = \sigma_2 \in (0, 1)$.

Assumption (D): The map $\Psi(\theta, \eta(\cdot;\theta)) = P\psi(\theta, \eta(\cdot;\theta))$ has a nonsingular partial derivative with respect to θ at $(\theta_0, \eta_0(\cdot;\theta_0))$, where $\psi(\theta, \eta(\cdot;\theta))$ is given in (3.1) for the Cox model and in (3.7) for the additive hazards model.

Assumption (E): In case-cohort studies, data are missing at random with $\pi_i \geq \sigma_3 > 0$ for all i and a constant σ_3.

Note that the assumption of compact Θ is only for technical convenience, which may be unnecessarily strong. Later we will see that for the additive hazards model, η is free of θ. Also note that $\psi(\theta, \eta(\cdot;\theta))$ reduces to $\psi(\eta(\cdot;\theta))$ for the Cox model, but the first argument still stays in order to keep a clear and consistent notation. The following are some standard empirical process notations that we use in the rest of the paper. Suppose $X_1, \ldots, X_n$ are i.i.d. p-dimensional random variables that follow the distribution P on a measurable space $(\mathcal{X}, \mathcal{A})$. For a measurable function $f : \mathcal{X} \mapsto \mathbb{R}$, let

$$\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^{n} f(X_i)\,, \quad Pf = \int f dP\,,$$

and

$$\mathbb{G}_n f = n^{-1/2}\sum_{i=1}^{n}\{f(X_i) - Pf\} = n^{1/2}(\mathbb{P}_n - P)f\,.$$

Function f can be replaced by a random function $x \mapsto \hat{f}_n(x; X_1, \ldots, X_n)$. Thus,

$$\mathbb{P}_n\hat{f}_n = \frac{1}{n}\sum_{i=1}^{n}\hat{f}_n(X_i; X_1, \ldots, X_n)\,, \quad P\hat{f}_n = \int \hat{f}_n(x; X_1, \ldots, X_n)dP(x)\,,$$

and

$$\mathbb{G}_n\hat{f}_n = n^{-1/2}\sum_{i=1}^{n}\{\hat{f}_n(X_i; X_1, \ldots, X_n) - P\hat{f}_n\} = n^{1/2}(\mathbb{P}_n - P)\hat{f}_n\,.$$

3.1. Case-cohort study: the Cox model

For the Cox model with external time-dependent covariates, we have

$$\lambda(t|\bar{Z}(t)) = \lambda_0(t)e^{\theta_0' Z(t)}\,,$$
$$1 - F_{T|\bar{Z}(\tau)}(t|\bar{z}(\tau)) = 1 - F_{T|\bar{Z}(t)}(t|\bar{z}(t)) = \exp\left\{-\int_0^t e^{\theta_0' z(s)} d\Lambda_0(s)\right\},$$

where $F_{T|\bar{Z}(\tau)}$ is the conditional distribution function of T given $\bar{Z}(\tau)$, Λ_0 is the baseline cumulative hazard function, and θ_0 is the parameter of interest. Let X_i be the i-th observation in the case-cohort study. We define the following random map

$$\Psi_n(\theta,\eta) = \frac{1}{n}\sum_{i=1}^n \psi(X_i;\theta,\eta) = \frac{1}{n}\sum_{i=1}^n \Omega_i(Y_i)\{Z_i(Y_i) - \eta(Y_i;\theta)\}\Delta_i\,, \tag{3.1}$$

with true η given by

$$\eta_0(t;\theta) = \frac{E\{Z(t)e^{\theta' Z(t)}1(Y \geq t)\}}{E\{e^{\theta' Z(t)}1(Y \geq t)\}}\,,$$

where Ω_i's are diagonal weight matrices with subject and covariate specific random weights on the diag that are allowed to depend on time and have expectation 1 given underlying complete data $(Y_i, \Delta_i, \bar{Z}_i(Y_i), \bar{Z}_i^*(Y_i))$. By choosing a weight matrix, we are allowed to weight each component of $\psi(X_i;\theta,\eta)$ differently, as in Kulich and Lin (2004). For notational simplicity, we consider a scalar weight Ω_i in the rest of the article. The proof for matrices Ω_i's is almost identical. It has been shown by Andersen and Gill (1982) that $E\psi(\theta_0, \eta_0(\cdot;\theta_0)) = E[\{Z(Y) - \eta_0(Y;\theta_0)\}\Delta] = 0$. The explicit functional form of η_0 is unknown and needs to be estimated first in order to estimate θ from (3.1).

For full-cohort data, $\Omega_i = 1$, and the partial likelihood estimating function is

$$\Psi_n(\theta, \hat{\eta}_n^F) = \frac{1}{n}\sum_{i=1}^n \{Z_i(Y_i) - \hat{\eta}_n^F(Y_i;\theta)\}\Delta_i\,, \tag{3.2}$$

where $\hat{\eta}_n^F$ is an estimator of η_0 using full data, of the form

$$\hat{\eta}_n^F(t;\theta) = \frac{\sum_{j=1}^n Z_j(t)e^{\theta' Z_j(t)}1(Y_j \geq t)}{\sum_{j=1}^n e^{\theta' Z_j(t)}1(Y_j \geq t)}\,.$$

For case-cohort data where the subcohort is a sub-sample of the entire cohort selected with a constant probability π_i for all i, and with $\Omega_i = 1$, the pseudo-likelihood estimating function of Self and Prentice (1988) is

$$\Psi_n(\theta, \hat{\eta}_n^{SP}) = \frac{1}{n}\sum_{i=1}^n \{Z_i(Y_i) - \hat{\eta}_n^{SP}(Y_i;\theta)\}\Delta_i\,, \tag{3.3}$$

where $\hat{\eta}_n^{SP}$ is the estimator of η_0 considered by Self and Prentice (1988) using the subcohort data only, of the form

$$\hat{\eta}_n^{SP}(t;\theta) = \frac{\sum_{j\in\mathcal{SC}} Z_j(t)e^{\theta' Z_j(t)}1(Y_j \geq t)}{\sum_{j\in\mathcal{SC}} e^{\theta' Z_j(t)}1(Y_j \geq t)}\,.$$

Here $\mathcal{SC}$ denotes the set of subjects in the subcohort.

In order to improve efficiency, the subcohort can be chosen by stratified sampling and, furthermore, it is tempting to include failures outside the subcohort to estimate η_0, see e.g., Kalbfleisch and Lawless (1988). The corresponding estimating function then becomes

$$\Psi_n(\theta, \hat{\eta}_n^W) = \frac{1}{n}\sum_{i=1}^{n} \Omega_i(Y_i)\{Z_i(Y_i) - \hat{\eta}_n^W(Y_i;\theta)\}\Delta_i\,, \tag{3.4}$$

where $\hat{\eta}_n^W$ is a weighted estimator of η_0 of the form

$$\hat{\eta}_n^W(t;\theta) = \frac{\sum_{j=1}^{n} W_j(t)Z_j(t)e^{\theta' Z_j(t)}1(Y_j \geq t)}{\sum_{j=1}^{n} W_j(t)e^{\theta' Z_j(t)}1(Y_j \geq t)}\,.$$

Here W_i could also be diagonal weight matrices with subject and covariate specific random weights on the diag. Again for notational simplicity, we consider scalar W_i, which may or may not equal to Ω_i. We also require that W_i have expectation 1 given the complete data $(Y_i, \Delta_i, Z_i(\cdot), Z_i^*(\cdot))$. We consider a broad class of weighted problems by allowing both weights Ω and W to be time-dependent. The commonly used weights, originally proposed by Kalbfleisch and Lawless (1988), are the inverse-probability weights

$$W_i = \Delta_i + \frac{R_i}{\pi_i}(1 - \Delta_i)\,, \tag{3.5}$$

where π_i can be time-dependent, see Kulich and Lin (2004) for example.

Note that the estimating functions in (3.2) and (3.3) can be expressed by using counting process stochastic integrals, and martingale theory applies in deriving asymptotic properties of the corresponding estimators, see e.g., Andersen and Gill (1982) and Self and Prentice (1988). Using a similar stochastic integral for the estimating function (3.4) with weights (3.5), however, creates a measurability problem because the integrand is no longer adapted to any meaningful filtration (and hence not predictable). See e.g., Chung and Williams (1990) and Protter (2004) for detailed discussions on stochastic integration. In this article, instead of using stochastic integrals, we give a rigorous proof of asymptotic properties of the estimators obtained from the estimating function (3.4) using the general Z-estimation theory provided in Section 2.

It grants great flexibility in estimating θ from (3.4) to use two possibly different weights Ω_i and W_i . When $\Omega_i = W_i = 1$, the estimating function $\Psi_n(\theta, \hat{\eta}_n^W(\cdot;\theta))$ reduces to (3.2), the partial likelihood estimating function of Cox (1972) for full-cohort data. When $\Omega_i = 1$ and $W_i = R_i/\pi_i$ with constant $\pi_i = \pi > 0$ for all i, $\Psi_n(\theta, \hat{\eta}_n^W(\cdot;\theta))$ is (3.3), the pseudo-likelihood estimating function of Self and Prentice (1988). When $\Omega_i = W_i$ as in (3.5), $\Psi_n(\theta, \hat{\eta}_n^W(\cdot;\theta))$ is equivalent to the weighted estimating function of Kalbfleisch and Lawless (1988). When $\Omega_i = W_i = R_i^*/\pi_i^*$, with R_i^* being 1 if subject i has complete data and 0 otherwise, and $\pi_i^* = P(R_i^* = 1|Y_i, \Delta_i, \bar{Z}_i(Y_i), \bar{Z}_i^*(Y_i))$, $\Psi_n(\theta, \hat{\eta}_n^W(\cdot;\theta))$ is the estimating function proposed by Pugh et al. (1992). The corresponding asymptotic properties have been studied by Breslow and Wellner (2007) for both independent stratified Bernoulli sampling and finite population stratified sampling when covariates are time-independent. To improve efficiency, Kulich and Lin (2004) considered the estimating function $\Psi_n(\theta, \hat{\eta}_n^W(\cdot;\theta))$ with $\Omega_i = 1$ and W_i being time-dependent weights. A clear advantage of introducing weights Ω_i in $\Psi_n(\theta, \hat{\eta}_n^W(\cdot;\theta))$ is that it allows one to estimate θ from a data set in which some failures may have missing data, e.g., the two-phase design studied by Breslow and Wellner (2007). This is more general than a traditional case-cohort study that requires all failures to be completely observed. It is obvious that all the above weights are nonnegative and bounded, have unit conditional expectation given complete data by Assumption (E), and are zero if corresponding covariates are missing. We assume this holds throughout the rest of the paper.

The main results are as follows.

Proposition 1. *Let $\hat{\eta}_n(t;\theta) = \hat{\eta}_n^W(t;\theta)$ as in* (3.4)*, and suppose the weight process $W(t)$ has bounded sample paths of bounded variation. Then both $\hat{\eta}_n(t;\theta)$ and $\eta_0(t;\theta)$ belong to a Donsker class, and further we have $\|\hat{\eta}_n - \eta_0\| = O_{p^*}(n^{-1/2})$.*

Proposition 1 plays an important role in the proofs of consistency and asymptotic normality provided in Propositions 2 and 3, respectively. The proof of Proposition 1 is deferred to Appendix B.

Proposition 2. *If the conditions in Proposition* 1 *hold and the weight process $\Omega(t)$ has bounded sample paths of bounded variation, then the root of* (3.4)*, denoted as $\hat{\theta}_n$, is a consistent estimator of θ_0.*

Proposition 2 can be proved by verifying conditions in Lemma 1. Details are given in Appendix B.

Proposition 3. *If the conditions in Propositions* 3.1 *and* 3.2 *hold, then the root of* (3.4) *has the asymptotic representation*

$$n^{1/2}(\hat{\theta}_n - \theta_0) = \Big\{ -\frac{\partial}{\partial\theta}\Psi(\theta_0, \eta_0(\cdot;\theta))\Big|_{\theta=\theta_0} \Big\}^{-1} \mathbb{G}_n \Big[\Omega(Y)\{Z(Y) - \eta_0(Y;\theta_0)\}\Delta - \int W(t)\{Z(t) - \eta_0(t;\theta_0)\}e^{\theta_0' Z(t)} 1(Y \geq t) d\Lambda_0(t) \Big] + o_{p^*}(1). \quad (3.6)$$

The limit distribution of $n^{1/2}(\hat{\theta}_n - \theta_0)$ *is Gaussian with zero mean and variance* $A^{-1}B\left(A^{-1}\right)'$, *where*

$$A = -\left.\frac{\partial}{\partial\theta}\Psi(\theta, \eta_0(\cdot;\theta))\right|_{\theta=\theta_0},$$
$$B = P\Big[\Omega(Y)\{Z(Y) - \eta_0(Y;\theta_0)\}\Delta - \int W(t)\{Z(t) - \eta_0(t;\theta_0)\}e^{\theta_0' Z(t)}1(Y \geq t)d\Lambda_0(t)\Big]^{\otimes 2},$$

where $a^{\otimes 2} = aa'$.

Equation (3.6) can be derived following Lemma 2, and the asymptotic normality follows directly from the Central Limit Theorem. The proof of Proposition 3 is in Appendix B.

It is worth noting that (3.6) reduces to the asymptotic representation of the partial likelihood estimator of Cox (1972) when $\Omega_i = W_i = 1$ for all i. It also reduces to the asymptotic representation of Self and Prentice (1988) when $\Omega_i = 1$ and W_i is the inverse selection probability weight of subject i into the subcohort, and of Breslow and Wellner (2007) when Ω_i and W_i are the inverse selection probability weight in a two-phase sampling design. The estimators here are generally not semiparametric efficient except for the case of full-cohort data where $\Omega_i = W_i = 1$ for all i. Efficient estimation is not our focus here. See Nan, Emond, and Wellner (2004) for calculations of information bounds, and Nan (2004) for an efficient estimator when covariates are discrete.

We have taken the weights Ω_i and W_i as given for each i. It has been shown in the missing data literature that using estimated rather than known weights can improve efficiency, see e.g., Robins, Rotnitzky, and Zhao (1994), Breslow and Wellner (2007), and Li and Nan (2011). In particular, Breslow and Wellner (2007) showed that, for the Cox model with time-independent covariates, the weighted estimator from a finite population sampling has the same asymptotic distribution as the weighted estimator from an i.i.d. Bernoulli sampling with the same selection probability but using the estimated weights. The asymptotic variance is smaller than that obtained using the true weights for the case of i.i.d. sampling. The same property holds for the Cox model with time-dependent covariates and time-dependent weights in the case of i.i.d. sampling.

3.2. Case-cohort study: the additive hazards model

Lin and Ying (1994) proposed an additive hazards model in which the hazard function given covariate history $\bar{Z}(\cdot)$ is

$$\lambda(t|\bar{Z}(t)) = \lambda_0(t) + \theta_0' Z(t),$$

where λ_0 is the baseline hazard and θ_0 is the parameter of interest. This model allows one to estimate the covariate effect on the absolute risk. Define the random map

$$\begin{aligned}\Psi_n(\theta,\eta) &= \frac{1}{n}\sum_{i=1}^n \psi(X_i;\theta,\eta)\\ &= \frac{1}{n}\sum_{i=1}^n \Big\{\Omega_i(Y_i)\{Z_i(Y_i)-\eta(Y_i)\}\Delta_i\\ &\quad - \int \Omega_i(t)\{Z_i(t)-\eta(t)\}1(Y_i\ge t)\theta' Z_i(t)dt\Big\} \qquad (3.7)\end{aligned}$$

with

$$\eta_0(t) = \frac{E\{Z(t)1(Y\ge t)\}}{E\{1(Y\ge t)\}},$$

where the Ω_i are defined in the same way as done for the Cox model. Then the estimating function proposed by Lin and Ying (1994) can be viewed as (3.7) with $\Omega_i = 1$ and η_0 being estimated empirically:

$$\begin{aligned}\Psi_n(\theta,\tilde\eta_n^F) &= \frac{1}{n}\sum_{i=1}^n \Big\{\{Z_i(Y_i)-\tilde\eta_n^F(Y_i)\}\Delta_i\\ &\quad - \int \{Z_i(t)-\tilde\eta_n^F(t)\}1(Y_i\ge t)\theta' Z_i(t)\,dt\Big\} \qquad (3.8)\end{aligned}$$

with

$$\tilde\eta_n^F(t) = \frac{\sum_{j=1}^n Z_j(t)1(Y_j\ge t)}{\sum_{j=1}^n 1(Y_j\ge t)}.$$

Note that both η_0 and $\tilde\eta_n^F$ do not involve θ. The estimator of θ is

$$\tilde\theta_n = \Big[\frac{1}{n}\sum_{i=1}^n \int \{Z_i(t)-\tilde\eta_n^F(t)\}^{\otimes 2}1(Y_i\ge t)dt\Big]^{-1}\frac{1}{n}\sum_{i=1}^n\{Z_i(Y_i)-\tilde\eta_n^F(Y_i)\}\Delta_i\,. \qquad (3.9)$$

Lin and Ying (1994) defined $\Psi_n(\theta,\tilde\eta_n^F)$ and $\tilde\theta_n$ using the stochastic integral formulation and studied their asymptotic properties using martingale theory.

For case-cohort studies, Kulich and Lin (2000) modified the estimating function (3.8) and proposed the estimating function (with $\Omega_i = W_i$)

$$\begin{aligned}\Psi_n(\theta,\tilde\eta_n^W) &= \frac{1}{n}\sum_{i=1}^n \Big\{\Omega_i(Y_i)\{Z_i(Y_i)-\tilde\eta_n^W(Y_i)\}\Delta_i\\ &\quad - \int \Omega_i(t)\{Z_i(t)-\tilde\eta_n^W(t)\}1(Y_i\ge t)\theta' Z_i(t)\,dt\Big\} \qquad (3.10)\end{aligned}$$

with

$$\tilde{\eta}_n^W(t) = \frac{\sum_{j=1}^n W_j(t) Z_j(t) 1(Y_j \geq t)}{\sum_{j=1}^n W_j(t) 1(Y_j \geq t)}. \tag{3.11}$$

The estimator is

$$\tilde{\theta}_n = \Big[\frac{1}{n}\sum_{i=1}^n \int \Omega_i(t)\{Z_i(t) - \tilde{\eta}_n^W(t)\} Z_i(t)' 1(Y_i \geq t) dt\Big]^{-1} \times \frac{1}{n}\sum_{i=1}^n \Omega_i(Y_i)\{Z_i(Y_i) - \tilde{\eta}_n^W(Y_i)\}\Delta_i \ . \tag{3.12}$$

Here we have extended the method of Kulich and Lin (2000) by introducing two weight matrices Ω and W in (3.7) and (3.11), respectively, as in the previous subsection.

When weights W_i or Ω_i depend on Δ_i as in (3.5), for the same reason as that in the previous example, martingale theory does not apply. Here we provide a proof without using stochastic integrals. As we assumed for the Cox model, Ω_i and W_i are nonnegative with unit conditional expectation given complete data.

We consider the weighted estimating function (3.10) that reduces to (3.8) when $\Omega_i = W_i = 1$ for all i. We assume a one-dimensional covariate Z and thus a one-dimensional θ in the following, the multi-dimensional case is a straightforward extension. The main results for the additive hazards model are the following.

Proposition 4. *Let $\tilde{\eta}_n(t) = \hat{\eta}_n^W(t)$ as in equation (3.10). If the weight process $W(t)$ has bounded sample paths of bounded variation, then both $\tilde{\eta}_n(t)$ and $\eta_0(t)$ belong to a Donsker class, and furthermore, $\|\tilde{\eta}_n - \eta_0\| = O_{p^*}(n^{-1/2})$.*

This is a special case of Proposition 1.

Proposition 5. *If the conditions in Proposition 3.4 hold and the weight process $\Omega(t)$ has bounded sample paths of bounded variation, then the root of* (3.10) *is a consistent estimator of θ_0.*

Similar to the proof of Proposition 2, we only need to verify the conditions in Lemma 1. Details are given in Appendix B.

Proposition 6. *If the conditions in Propositions* 4 *and* 5 *hold, then $n^{1/2}(\tilde{\theta}_n - \theta_0)$ converges in distribution to a zero mean Gaussian random variable.*

The proof is in Appendix B.

4. Discussion

We have discussed the proportional hazards model and the additive hazards

model in case-cohort studies, though our method applies to a much broader range of semiparametric estimation problems. The parameter estimation in the case-cohort studies is hard to handle by traditional martingale-based methods when certain more efficient but unpredictable weights are considered, but becomes straightforward by using the general Z-estimation theory.

For case-cohort studies, Breslow and Wellner (2007) considered finite population stratified sampling and applied the exchangeably weighted bootstrap empirical process theory of Præstgaard and Wellner (1993) for the Cox model with *time-independent* covariates. The general Z-estimation theory in Section 2 is likely to be applicable to the finite population stratified sampling designs for *time-dependent* covariates.

Another widely used cost-effective design for censored survival data, the nested case-control study, samples controls from risk populations at different observed failure times. The sampling probabilities for controls are thus dependent upon information from other subjects (observed failures in this case), creating a complicated dependent sampling. In such a design, the constructed risk sets and the sampling probabilities are predictable, allowing the standard counting process martingale theory to be applied to the proofs of asymptotic properties, see e.g., Goldstein and Langholz (1992) and Borgan, Goldstein, and Langholz (1995). The general Z-estimation theory, however, may provide alternative (potentially more concise) proofs for the nested case-control study. The key is to establish desirable properties for $\hat{\eta}_n$, which highly depend on the sampling scheme for selecting controls at each failure time point and the construction of estimating method.

For missing data problems, the estimated likelihood method of Pepe and Fleming (1991), the mean score method of Reilly and Pepe (1995), and the pseudoscore method of Chatterjee, Chen, and Breslow (2003), among others, also fit into the general Z-estimation framework nicely. Let Y be the response variable and (Z, V) be covariates where Z is sometimes missing. Let R be 1 if Z is observed and 0 otherwise, and let X denote the observed data. Suppose that the parameter of interest $\theta \in \Theta \subset \mathbb{R}^d$ could be estimated by using the complete data score function $\dot{l}^0_\theta(\cdot;\theta)$ as the estimating function if there were no missing data. When Z is sometimes missing at random (Little and Rubin (2002)), then the observed data score function for θ is

$$\dot{l}_\theta(X;\theta,\eta_0(\cdot;\theta)) = R\dot{l}^0_\theta(Y,Z,V;\theta) + (1-R)\eta_0(Y,V;\theta)\,,$$

where $\eta_0(Y,V;\theta) = E\{\dot{l}^0_\theta(Y,Z,V;\theta)|Y,V\}$ has unknown functional form. If $\psi(\cdot;\theta,\eta(\cdot;\theta)) = \dot{l}_\theta(\cdot;\theta,\eta(\cdot;\theta))$, then $\psi(\cdot;\theta,\hat{\eta}_n(\cdot;\theta))$ is an estimating function for θ where $\hat{\eta}_n(\cdot;\theta))$ is an estimator of $\eta_0(\cdot;\theta)$. The asymptotic properties of the Z-estimator for θ depend on the behavior of $\hat{\eta}_n$ and can be derived from the

theorems given in Section 2. Nonparametric methods have been proposed to estimate $\eta_0(\cdot;\theta)$. Apparently efficiency can be improved by using the weighted estimating function of Robins, Rotnitzky, and Zhao (1994). This may also apply to the composite likelihoods for semiparametric models, see e.g., Lindsay (1987) and Varin, Reid, and Firth (2011), particularly for missing data problems.

The theory in Section 2 requires smooth η with respect to θ, according to (2.4) or (2.5). For the rank-based estimating function for the accelerated failure time model, the smoothness condition does not hold. Nan, Kalbfleisch, and Yu (2009) have shown that a similar idea for bundled parameters with missing data is applicable to the rank-based estimator for the accelerated failure time model. For models with bundled parameters in the original parameterization, Ding and Nan (2011) have proposed a sieve maximum likelihood estimating method and applied the method to the efficient estimation of the accelerated failure time model.

Acknowledgement

The first author was supported in part by NSF grant DMS–1007590 and NIH grant R01 AG036802. The second author was supported in part by NSF grant DMS–1104832, NIAID grant 2R01 AI291968-04, and the Alexander von Humboldt Foundation.

Appendix A: Proofs of Lemmas 1 and 2

A.1. Proof of Lemma 1

Since θ_0 is the unique solution to $\Psi(\theta, \eta_0(\cdot;\theta)) = 0$, for any fixed $\epsilon > 0$ there exists a $\delta > 0$ such that

$$P\left[|\hat{\theta}_n - \theta_0| > \epsilon\right] \leq P\left[|\Psi(\hat{\theta}_n, \eta_0(\cdot;\hat{\theta}_n))| > \delta\right].$$

If we can prove $|\Psi(\hat{\theta}_n, \eta_0(\cdot;\hat{\theta}_n))| \to_{p^*} 0$, then the consistency of $\hat{\theta}_n$ follows immediately.

Now, since $||\hat{\eta}_n - \eta_0|| = o_{p^*}(1)$, there exists a sequence $\{\delta_n\} \downarrow 0$ such that $||\hat{\eta}_n - \eta_0|| \leq \delta_n$ with probability tending to one. Hence taking $\eta = \hat{\eta}_n$ in (2.3), we have

$$\begin{aligned}
&|\Psi(\hat{\theta}_n, \eta_0(\cdot;\hat{\theta}_n))| \\
&\quad \leq |\Psi_n(\hat{\theta}_n, \hat{\eta}_n(\cdot;\hat{\theta}_n))| + |\Psi(\hat{\theta}_n, \eta_0(\cdot;\hat{\theta}_n)) - \Psi_n(\hat{\theta}_n, \hat{\eta}_n(\cdot;\hat{\theta}_n))| \\
&\quad \leq |\Psi_n(\hat{\theta}_n, \hat{\eta}_n(\cdot;\hat{\theta}_n))| + o_{p^*}\Big(1 + |\Psi_n(\hat{\theta}_n, \hat{\eta}_n(\cdot;\hat{\theta}_n))| + |\Psi(\hat{\theta}_n, \eta_0(\cdot;\hat{\theta}_n))|\Big) \\
&\quad \leq o_{p^*}(1) + o_{p^*}\left(1 + o_{p^*}(1) + |\Psi(\hat{\theta}_n, \eta_0(\cdot;\hat{\theta}_n))|\right),
\end{aligned}$$

which implies $|\Psi(\hat{\theta}_n, \eta_0(\cdot;\hat{\theta}_n))| = o_{p^*}(1)$.

A.2. Proof of Lemma 2

We first show the following result that will be used later:

$$n^{1/2}\left|\Psi(\hat{\theta}_n,\hat{\eta}_n(\cdot,\hat{\theta}_n))\right| = O_{p^*}(1). \tag{A.1}$$

By Condition (i),

$$\begin{aligned}
&n^{1/2}\left|(\Psi_n-\Psi)(\hat{\theta}_n,\hat{\eta}_n(\cdot;\hat{\theta}_n))-(\Psi_n-\Psi)(\theta_0,\eta_0(\cdot;\theta_0))\right| \\
&\quad = o_{p^*}(1)+o_{p^*}\left(n^{1/2}\left|\Psi_n(\hat{\theta}_n,\hat{\eta}_n(\cdot;\hat{\theta}_n))\right|\right)+o_{p^*}\left(n^{1/2}\left|\Psi(\hat{\theta}_n,\hat{\eta}_n(\cdot;\hat{\theta}_n))\right|\right).
\end{aligned}$$

By the triangle inequality $-|a|+|b|-|c| \le |a-b-c|$ and the fact that $\Psi(\theta_0,\eta_0(\cdot;\theta_0)) = 0$,

$$\begin{aligned}
&n^{1/2}\left|\Psi(\hat{\theta}_n,\hat{\eta}_n(\cdot;\hat{\theta}_n))\right| - n^{1/2}\left|\Psi_n(\hat{\theta}_n,\hat{\eta}_n(\cdot;\hat{\theta}_n))\right| - n^{1/2}\left|\Psi_n(\theta_0,\eta_0(\cdot;\theta_0))\right| \\
&\quad \le n^{1/2}\left|(\Psi_n-\Psi)(\hat{\theta}_n,\hat{\eta}_n(\cdot;\hat{\theta}_n))-(\Psi_n-\Psi)(\theta_0,\eta_0(\cdot;\theta_0))\right| \\
&\quad = o_{p^*}(1)+o_{p^*}\left(n^{1/2}\left|\Psi_n(\hat{\theta}_n,\hat{\eta}_n(\cdot;\hat{\theta}_n))\right|\right)+o_{p^*}\left(n^{1/2}\left|\Psi(\hat{\theta}_n,\hat{\eta}_n(\cdot;\hat{\theta}_n))\right|\right),
\end{aligned}$$

which implies

$$\begin{aligned}
&n^{1/2}\left|\Psi(\hat{\theta}_n,\hat{\eta}_n(\cdot;\hat{\theta}_n))\right|[1-o_{p^*}(1)] \\
&\quad \le o_{p^*}(1)+n^{1/2}\left|\Psi_n(\hat{\theta}_n,\hat{\eta}_n(\cdot;\hat{\theta}_n))\right|[1+o_{p^*}(1)]+n^{1/2}\left|\Psi_n(\theta_0,\eta_0(\cdot;\theta_0))\right| \\
&\quad = o_{p^*}(1)+o_{p^*}(1)+O_{p^*}(1).
\end{aligned}$$

Hence (A.1) holds.

We then show the root-n consistency of $\hat{\theta}_n$. Since $|\hat{\theta}_n-\theta_0| = o_{p^*}(1)$ and $||\hat{\eta}_n-\eta_0|| = O_{p^*}(n^{-\beta})$ with $\beta>0$, there exists a sequence $\{\delta_n\}\downarrow 0$ and $c>0$ such that $|\hat{\theta}_n-\theta_0|\le\delta_n$ and $||\hat{\eta}_n-\eta_0||\le cn^{-\beta}$ with probability approaching one. Hence, taking $(\theta,\eta)=(\hat{\theta}_n,\hat{\eta}_n)$ in the smoothness condition (2.5),

$$\begin{aligned}
&\Big|n^{1/2}\left\{\Psi(\hat{\theta}_n,\hat{\eta}_n(\cdot;\hat{\theta}_n))-\Psi(\theta_0,\eta_0(\cdot;\theta_0))\right\} \\
&\qquad -n^{1/2}\left\{\dot{\Psi}_1(\theta_0,\eta_0(\cdot;\theta_0))+\dot{\Psi}_2(\theta_0,\eta_0(\cdot;\theta_0))[\dot{\eta}_0(\cdot;\theta_0)]\right\}(\hat{\theta}_n-\theta_0) \\
&\qquad -n^{1/2}\dot{\Psi}_2(\theta_0,\eta_0(\cdot;\theta_0))[(\hat{\eta}_n-\eta_0)(\cdot;\theta_0)]\Big| \\
&= o_{p^*}\left(n^{1/2}|\hat{\theta}_n-\theta_0|\right)+O_{p^*}\left(n^{1/2}\|\hat{\eta}_n-\eta_0\|^{\alpha}\right) \\
&= o_{p^*}\left(1+n^{1/2}|\hat{\theta}_n-\theta_0|\right),
\end{aligned} \tag{A.2}$$

since $n^{1/2}O_{p^*}(||\hat{\eta}_n-\eta_0||^{\alpha}) = o_{p^*}(1)$ by $\alpha\beta>1/2$. The same result can be obtained by using the smoothness condition (2.4) for $\beta=1/2$. By (A.1), the fact that

$\Psi(\theta_0, \eta_0(\cdot;\theta_0)) = 0$, and the triangle inequality $-|a| + |b| - |c| \le |a - b - c|$, (A.2) implies

$$\begin{aligned} -O_{p^*}(1) + &\left| n^{1/2} \left\{ \dot{\Psi}_1(\theta_0, \eta_0(\cdot;\theta_0)) + \dot{\Psi}_2(\theta_0, \eta_0(\cdot;\theta_0))[\dot{\eta}_0(\cdot;\theta_0)] \right\} (\hat{\theta}_n - \theta_0) \right| \\ &- \left| n^{1/2} \dot{\Psi}_2(\theta_0, \eta_0(\cdot;\theta_0))[(\hat{\eta}_n - \eta_0)(\cdot;\theta_0)] \right| \\ &\le o_{p^*}\left(1 + n^{1/2} \left| \hat{\theta}_n - \theta_0 \right| \right). \end{aligned} \tag{A.3}$$

Since the $d \times d$ matrix $\dot{\Psi}_1(\theta_0, \eta_0(\cdot;\theta_0)) + \dot{\Psi}_2(\theta_0, \eta_0(\cdot;\theta_0))[\dot{\eta}_0(\cdot;\theta_0)]$ is nonsingular, there exist a constant $c_1 > 0$ such that

$$\left| \left\{ \dot{\Psi}_1(\theta_0, \eta_0(\cdot;\theta_0)) + \dot{\Psi}_2(\theta_0, \eta_0(\cdot;\theta_0))[\dot{\eta}_0(\cdot;\theta_0)] \right\} (\theta - \theta_0) \right| \ge c_1 |\theta - \theta_0|$$

for $|\theta - \theta_0| \to 0$. On the other hand, by Condition (iv), in combination with inequality (A.3),

$$\begin{aligned} O_{p^*}(1) &\ge \left| n^{1/2} \left\{ \dot{\Psi}_1(\theta_0, \eta_0(\cdot;\theta_0)) + \dot{\Psi}_2(\theta_0, \eta_0(\cdot;\theta_0))[\dot{\eta}_0(\cdot;\theta_0)] \right\} (\hat{\theta}_n - \theta_0) \right| \\ &\quad - \left| n^{1/2} \dot{\Psi}_2(\theta_0, \eta_0(\cdot;\theta_0))[(\hat{\eta}_n - \eta_0)(\cdot;\theta_0)] \right| - o_{p^*}\left(1 + n^{1/2} \left| \hat{\theta}_n - \theta_0 \right| \right) \\ &\ge c_1 n^{1/2} \left| \hat{\theta}_n - \theta_0 \right| - O_{p^*}(1) - o_{p^*}\left(1 + n^{1/2} \left| \hat{\theta}_n - \theta_0 \right| \right) \\ &= \{ O_{p^*}(1) - o_{p^*}(1) \} \, n^{1/2} \left| \hat{\theta}_n - \theta_0 \right| - O_{p^*}(1). \end{aligned}$$

Hence the sequence $n^{1/2} \left| \hat{\theta}_n - \theta_0 \right|$ must be bounded in outer probability.

Now we are ready to prove (2.6). Because

$$\begin{aligned} & n^{1/2} \left[\Psi(\hat{\theta}_n, \hat{\eta}_n(\cdot;\hat{\theta}_n)) - \Psi(\theta_0, \eta_0(\cdot;\theta_0)) \right] \\ &= n^{1/2} \left[\Psi(\hat{\theta}_n, \hat{\eta}_n(\cdot;\hat{\theta}_n)) - \Psi_n(\hat{\theta}_n, \hat{\eta}_n(\cdot;\hat{\theta}_n)) + \Psi_n(\hat{\theta}_n, \hat{\eta}_n(\cdot;\hat{\theta}_n)) - \Psi(\theta_0, \eta_0(\cdot;\theta_0)) \right] \\ &= n^{1/2} (\Psi - \Psi_n)(\hat{\theta}_n, \hat{\eta}_n(\cdot;\hat{\theta}_n)) + o_{p^*}(1) - 0 \\ &= -n^{1/2} (\Psi_n - \Psi)(\theta_0, \eta_0(\cdot;\theta_0)) \pm o_{p^*}\Big(1 + n^{1/2} \left| \Psi_n(\hat{\theta}_n, \hat{\eta}_n(\cdot;\hat{\theta}_n)) \right| \\ &\quad + n^{1/2} \left| \Psi(\hat{\theta}_n, \hat{\eta}_n(\cdot;\hat{\theta}_n)) \right| \Big) \quad \text{(by Condition (i))} \\ &= -n^{1/2} (\Psi_n - \Psi)(\theta_0, \eta_0(\cdot;\theta_0)) \pm o_{p^*}(1) \quad \text{(by equation (A.1))}, \end{aligned} \tag{A.4}$$

after replacing (A.4) into the first term in the first line of (A.2) we obtain

$$\begin{aligned} \Big| & -n^{1/2} (\Psi_n - \Psi)(\theta_0, \eta_0(\cdot;\theta_0)) \pm o_{p^*}(1) \\ & -n^{1/2} \left\{ \dot{\Psi}_1(\theta_0, \eta_0(\cdot;\theta_0)) + \dot{\Psi}_2(\theta_0, \eta_0(\cdot;\theta_0))[\dot{\eta}_0(\cdot;\theta_0)] \right\} (\hat{\theta}_n - \theta_0) \end{aligned}$$

$$
\begin{aligned}
&\quad -n^{1/2}\dot{\Psi}_2(\theta_0,\eta_0(\cdot;\theta_0))[(\hat{\eta}_n-\eta_0)(\cdot;\theta_0)]\Big| \\
&= o_{p^*}\left(1+n^{1/2}\left|\hat{\theta}_n-\theta_0\right|\right) \\
&= o_{p^*}(1),
\end{aligned}
$$

which implies

$$
\begin{aligned}
n^{1/2}(\hat{\theta}_n-\theta_0) = \Big\{ &-\dot{\Psi}_1(\theta_0,\eta_0(\cdot;\theta_0)) - \dot{\Psi}_2(\theta_0,\eta_0(\cdot;\theta_0))[\dot{\eta}_0(\cdot;\theta_0)]\Big\}^{-1} \\
&\times n^{1/2}\Big\{(\Psi_n-\Psi)(\theta_0,\eta_0(\cdot;\theta_0)) \\
&+\dot{\Psi}_2(\theta_0,\eta_0(\cdot;\theta_0))[(\hat{\eta}_n-\eta_0)(\cdot;\theta_0)]\Big\} + o_{p^*}(1)\,.
\end{aligned}
$$

Appendix B: Proofs of Propositions 1 to 6

B.1. Proof of Proposition 1

We consider one nuisance parameter η for simplicity. The vector η can be dealt with by examining each of its components. Define

$$
\begin{aligned}
D_n^{(0)}(t,\theta) &\equiv \mathbb{P}_n\Big\{W(t)e^{\theta' Z(t)}1(Y\geq t)\Big\}, \\
d^{(0)}(t,\theta) &\equiv P\Big\{W(t)e^{\theta' Z(t)}1(Y\geq t)\Big\} = P\Big\{e^{\theta' Z(t)}1(Y\geq t)\Big\}; \\
D_n^{(1)}(t,\theta) &\equiv \mathbb{P}_n\Big\{W(t)Z(t)e^{\theta' Z(t)}1(Y\geq t)\Big\}, \\
d^{(1)}(t,\theta) &\equiv P\Big\{W(t)Z(t)e^{\theta' Z(t)}1(Y\geq t)\Big\} = P\Big\{Z(t)e^{\theta' Z(t)}1(Y\geq t)\Big\}.
\end{aligned}
$$

Then we have

$$
\hat{\eta}_n(t;\theta) = \frac{D_n^{(1)}(t,\theta)}{D_n^{(0)}(t,\theta)}\,, \quad \eta_0(t;\theta) = \frac{d^{(1)}(t,\theta)}{d^{(0)}(t,\theta)}\,.
$$

Apparently the sets of functions $\mathcal{F}_0 = \{W(t)1(Y\geq t)e^{\theta' Z(t)} : 0\leq t\leq\tau, \theta\in\Theta\}$ and $\mathcal{F}_1 = \{W(t)1(Y\geq t)Z(t)e^{\theta' Z(t)} : 0\leq t\leq\tau, \theta\in\Theta\}$ are well-behaved and belong to Donsker classes, see e.g., van der Vaart and Wellner (1996), Section 2.10. Hence we have that $n^{1/2}\{D_n^{(k)}(t,\theta)-d^{(k)}(t,\theta)\}$ converge weakly to zero mean Gaussian processes, and $\|D_n^{(k)}-d^{(k)}\| = O_{p^*}(n^{-1/2})$, $k=0,1$. Let $\bar{\mathcal{F}}_k$ be the closure of $\mathcal{F}_k$, $k=0,1$, respectively, in which the convergence is both pointwise and in $L_2(P)$. Then $D_n^{(k)}(t,\theta)$ and $d^{(k)}(t,\theta)$ are in the convex hull of $\bar{\mathcal{F}}_k$, $k=0,1$, and thus Donsker. See e.g., van der Vaart and Wellner (1996), Theorems 2.10.2 and 2.10.3. Hence both $\{\hat{\eta}_n(t;\theta)\}$ and $\{\eta_0(t;\theta)\}$ are Donsker by van der Vaart and Wellner (1996), Example 2.10.9, where $D_n^{(0)}$ and $d^{(0)}$ are bounded away (almost surely) from zero by Assumption (C).

Now we verify that $\hat{\eta}_n$ is $n^{1/2}$-consistent by the calculation

$$\begin{aligned}
&n^{1/2}\{\hat{\eta}_n(t;\theta) - \eta_0(t;\theta)\} \\
&\quad = n^{1/2}\Big[\frac{1}{d^{(0)}(t,\theta)}\{D_n^{(1)}(t,\theta) - d^{(1)}(t,\theta)\} \\
&\qquad - \frac{D_n^{(1)}(t,\theta)}{D_n^{(0)}(t,\theta)d^{(0)}(t,\theta)}\{D_n^{(0)}(t,\theta) - d^{(0)}(t,\theta)\}\Big] \\
&\quad = n^{1/2}\Big[\frac{1}{d^{(0)}(t,\theta)}\{D_n^{(1)}(t,\theta) - d^{(1)}(t,\theta)\} \\
&\qquad - \frac{d^{(1)}(t,\theta)}{d^{(0)}(t,\theta)^2}\{D_n^{(0)}(t,\theta) - d^{(0)}(t,\theta)\}\Big] + o_{p^*}(1) \\
&\quad = d^{(0)}(t,\theta)^{-1}\mathbb{G}_n\Big[W(t)\{Z(t) - \eta_0(t;\theta)\}e^{\theta' Z(t)}1(Y \geq t)\Big] + o_{p^*}(1)\,.
\end{aligned}$$

Since the classes of functions $\{W(t)\}$, $\{1(Y \geq t)\}$, $\{Z(t)\}$, and $\{e^{\theta' Z(t)}\}$ are all Donsker, and η_0 is a bounded deterministic function, we know that the class $\{W(t)\{Z(t) - \eta_0(t;\theta)\}e^{\theta' Z(t)}1(Y \geq t)\}$ is Donsker (see e.g., van der Vaart and Wellner (1996), Section 2.10). We then obtain the desired result.

B.2. Proof of Proposition 2

The uniqueness of θ_0 as a root of $\Psi(\theta, \eta_0(\cdot;\theta))$ is proved by Andersen and Gill (1982), here $\Psi(\theta, \eta_0(\cdot;\theta))$ corresponds to the derivative of the limit of their function (2.7). The uniform consistency of $\hat{\eta}_n$ is given by Proposition 3.1. Now we verify condition (2.3). We consider one-dimensional θ for simplicity. Suppose that $\Omega_i < K < \infty$ for all i for a constant K. Let $\|\eta - \eta_0\| \leq \delta_n \downarrow 0$. Then we have

$$\begin{aligned}
&|\Psi_n(\theta, \eta(\cdot;\theta)) - \Psi(\theta, \eta_0(\cdot;\theta))| \\
&\quad = \Big|\mathbb{P}_n[\Omega(Y)\{Z(Y) - \eta(Y;\theta)\}\Delta] - P[\Omega(Y)\{Z(Y) - \eta_0(Y;\theta)\}\Delta]\Big| \\
&\quad \leq \Big|\mathbb{P}_n[\Omega(Y)Z(Y)\Delta] - P[\Omega(Y)Z(Y)\Delta]\Big| + \Big|\mathbb{P}_n\left[\Omega(Y)\{\eta(Y;\theta) - \eta_0(Y;\theta)\}\Delta\right]\Big| \\
&\qquad + \Big|(\mathbb{P}_n - P)[\Omega(Y)\eta_0(Y;\theta)\Delta]\Big|\,.
\end{aligned}$$

The first term on the right side of the above inequality converges to zero in probability by the Weak Law of Large Numbers, while the second term

$$\Big|\mathbb{P}_n[\Omega(Y)\{\eta(Y;\theta) - \eta_0(Y;\theta)\}\Delta]\Big| \;\leq\; \mathbb{P}_n[\Omega(Y)\|\eta - \eta_0\|\Delta] \;\leq\; K\delta_n \;\to\; 0$$

uniformly over θ. The last term converges uniformly to zero in outer probability because $\{\eta_0(t;\theta) : 0 \leq t \leq \tau, \theta \in \Theta\}$ is a Donsker class, and $\{\Omega(t)\}$ and $\{\Delta\}$ are also Donsker. Thus $\{\Omega(t)\eta_0(t;\theta)\Delta\}$ is Donsker and hence a Glivenko-Cantelli class.

B.3. Proof of Proposition 3

Let $\mathcal{H}_0$ defined in Lemma 2 consist of functions of η_0 and $\hat{\eta}_n = \hat{\eta}_n^W$, thus a Donsker class. Obviously the class of functions $\{\psi(\theta, \eta(t;\theta)) = \Omega(t)\{Z(t) - \eta(t;\theta)\}\Delta : \theta \in \Theta_0, \eta \in \mathcal{H}_0, 0 \le t \le \tau\}$ is a Donsker class that satisfies $P_0|\psi(\theta,\eta) - \psi(\theta_0,\eta_0)|^2 \to 0$ as $|\theta - \theta_0| \to 0$ and $\|\eta - \eta_0\| \to 0$ by the Dominated Convergence Theorem. The Fréchet differentiability of $\{\Psi(\theta, \eta(\cdot;\theta)) : \theta \in \Theta_0, \eta \in \mathcal{H}_0\}$ can be verified via direct calculation. Thus from Propositions 1, 2, and the remark following Lemma 2 together with Assumption (D), we have the conditions in Lemma 2 satisfied and thus (2.6) holds.

Now we calculate the right side of (2.6) for the Cox model. Interchanging differentiation and integration yields

$$
\begin{aligned}
&n^{1/2}\dot{\Psi}_2(\theta_0, \eta_0(\cdot;\theta_0))[(\hat{\eta}_n - \eta_0)(\cdot;\theta_0)] \\
&\quad = -\, n^{1/2} P[\Omega(Y)\{\hat{\eta}_n(Y;\theta_0) - \eta_0(Y;\theta_0)\}\Delta] \\
&\quad = -\, n^{1/2} \int \Big[\frac{1}{d^{(0)}(t,\theta_0)}\{D_n^{(1)}(t,\theta_0) - d^{(1)}(t,\theta_0)\} \\
&\qquad - \frac{D_n^{(1)}(t,\theta_0)}{D_n^{(0)}(t,\theta_0) d^{(0)}(t,\theta_0)}\{D_n^{(0)}(t,\theta_0) - d^{(0)}(t,\theta_0)\}\Big] \delta dP_{Y,\Delta}(t,\delta) \\
&\quad = -\, n^{1/2} \int \Big[\frac{1}{d^{(0)}(t,\theta_0)}\{D_n^{(1)}(t,\theta_0) - d^{(1)}(t,\theta_0)\} \\
&\qquad - \frac{d^{(1)}(t,\theta_0)}{d^{(0)}(t,\theta_0)^2}\{D_n^{(0)}(t,\theta_0) - d^{(0)}(t,\theta_0)\}\Big] \delta dP_{Y,\Delta}(t,\delta) + o_{p^*}(1) \\
&\quad = -\, \mathbb{G}_n\Big\{ \int \Big\{W(t)\{Z(t) - \eta_0(t;\theta_0)\} e^{\theta_0' Z(t)} 1(Y \ge t)\Big\} \\
&\qquad \times \Big\{d^{(0)}(t,\theta_0)\Big\}^{-1} dP_{Y,\Delta}(t,1)\Big\} + o_{p^*}(1)\,.
\end{aligned}
$$

The second equality in the above holds because $E(\Omega|X) = 1$, and the third equality holds because the absolute difference between the two sides, except the term $o_{p^*}(1)$, is

$$
\begin{aligned}
&\Big|\int \Big\{\frac{d^{(1)}(t,\theta_0)}{d^{(0)}(t,\theta_0)^2} - \frac{D_n^{(1)}(t,\theta_0)}{D_n^{(0)}(t,\theta_0) d^{(0)}(t,\theta_0)}\Big\} n^{1/2}\Big\{D_n^{(0)}(t,\theta_0) - d^{(0)}(t,\theta_0)\Big\} \delta dP_{Y,\Delta}(t,\delta)\Big| \\
&\le \sup_{t\le\tau}\Big|\frac{d^{(1)}(t,\theta_0)}{d^{(0)}(t,\theta_0)^2} - \frac{D_n^{(1)}(t,\theta_0)}{D_n^{(0)}(t,\theta_0) d^{(0)}(t,\theta_0)}\Big| \sup_{t\le\tau}\Big|n^{1/2}\Big\{D_n^{(0)}(t,\theta_0) - d^{(0)}(t,\theta_0)\Big\}\Big| \\
&\qquad = \; o_{p^*}(1)\cdot O_{p^*}(1) = o_{p^*}(1)
\end{aligned}
$$

by Proposition 1 and tail bounds for the supremum of empirical processes in van der Vaart and Wellner (1996), Section 2.14.

Let $G(t|\bar{z}(t))$ be the conditional distribution function of the censoring time C at t given $\bar{Z}(t) = \bar{z}(t)$, or equivalently given $\bar{Z}(\tau) = \bar{z}(\tau)$ where $t \le \tau$, and H_t be the joint distribution function of $\bar{Z}(t)$. Then

$$\begin{aligned} d^{(0)}(t,\theta_0) &= P\Big\{W(t)1(Y \ge t)e^{\theta_0' Z(t)}\Big\} \\ &= P\Big\{1(Y \ge t)e^{\theta_0' Z(t)}\Big\} \\ &= E\Big[e^{\theta_0' Z(t)} E\{1(Y \ge t)|\bar{Z}(t)\}\Big] \\ &= E\Big[e^{\theta_0' Z(t)} P(T \ge t|\bar{Z}(t)) P(C \ge t|\bar{Z}(t))\Big] \\ &= \int e^{\theta_0' z(t)} \exp\Big\{-\int_0^t e^{\theta_0' z(s)} d\Lambda_0(s)\Big\}\{1 - G(t^-|\bar{z}(t))\} dH_t(\bar{z}(t))\,. \end{aligned}$$

On the other hand, from the joint distribution of $(Y, \Delta, \bar{Z}(Y))$, or equivalently of $(Y, \Delta, \bar{Z}(\tau))$, we obtain

$$\begin{aligned} dP_{Y,\Delta}(t,1) &= \Big[\int e^{\theta_0' z(t)} \exp\Big\{-\int_0^t e^{\theta_0' z(s)} d\Lambda_0(s)\Big\} \\ &\qquad \times \{1 - G(t^-|\bar{z}(t))\} dH_t(\bar{z}(t))\Big] d\Lambda_0(t) \\ &= d^{(0)}(t,\theta_0) d\Lambda_0(t). \end{aligned}$$

Thus we have

$$\begin{aligned} &n^{1/2} \dot{\Psi}_2(\theta_0, \eta_0(\cdot;\theta_0))[(\hat{\eta}_n - \eta_0)(\cdot;\theta_0)] \\ &\quad = -\mathbb{G}_n\Big[\int \Big\{W(t)\{Z(t) - \eta_0(t;\theta_0)\} e^{\theta_0' Z(t)} 1(Y \ge t)\Big\} d\Lambda_0(t)\Big] + o_{p^*}(1). \end{aligned}$$

It is obvious that $\dot{\Psi}_1 = 0$ and, by interchanging differentiation and integration, we have

$$\begin{aligned} \dot{\Psi}_2(\theta_0, \eta_0(\cdot;\theta_0))[\dot{\eta}_0(\cdot;\theta_0)] &= -P\Delta \dot{\eta}_0(Y;\theta_0) \\ &= P\Big[\frac{\partial}{\partial\theta}\psi(\theta_0, \eta_0(\cdot;\theta))\Big]_{\theta=\theta_0} = \frac{\partial}{\partial\theta}\Psi(\theta_0, \eta_0(\cdot;\theta))\Big|_{\theta=\theta_0}. \end{aligned}$$

Then by (2.6) we have

$$\begin{aligned} n^{1/2}(\hat{\theta}_n - \theta_0) = \Big\{-\frac{\partial}{\partial\theta}\Psi(\theta_0, \eta_0(\cdot;\theta))\Big|_{\theta=\theta_0}\Big\}^{-1} \mathbb{G}_n\Big[&\Omega(Y)\{Z(Y) - \eta_0(Y;\theta_0)\}\Delta \\ &- \int W(t)\{Z(t) - \eta_0(t;\theta_0)\} e^{\theta_0' Z(t)} 1(Y \ge t) d\Lambda_0(t)\Big] + o_{p^*}(1)\,, \end{aligned}$$

which converges in distribution to a zero mean Gaussian random variable by the Central Limit Theorem for i.i.d. data.

B.4. Proof of Proposition 4

The function $\tilde{\eta}_n^W(t)$ is the same as that in (3.4) at $\theta = 0$, and $\eta_0(t)$ is the same as its counterpart in the Cox model at $\theta = 0$. Hence the stated result is just a consequence of Proposition 1.

B.5. Proof of Proposition 5

Obviously $\Psi(\theta, \eta_0) = P\{\psi(\theta, \eta_0)\}$ is a linear function for θ with a non-zero slope by Assumption (D), hence θ_0 is the unique solution of $\Psi(\theta, \eta_0) = 0$. Proposition 4 provides the uniform consistency of $\hat{\eta}_n$. We now verify (2.3). Let $\|\eta - \eta_0\| \downarrow 0$. We have

$$
\begin{aligned}
&|\Psi_n(\theta, \eta) - \Psi(\theta, \eta_0)| \\
&\quad \le \left|\mathbb{P}_n[\Omega(Y)\{Z(Y) - \eta(Y)\}\Delta] - P[\Omega(Y)\{Z(Y) - \eta_0(Y)\}\Delta]\right| \\
&\qquad + \left|\mathbb{P}_n \int \Omega(t)\{Z(t) - \eta(t)\}1(Y \ge t)\theta Z(t)\, dt \right. \\
&\qquad \left. - P \int \Omega(t)\{Z(t) - \eta_0(t)\}1(Y \ge t)\theta Z(t)\, dt\right| \\
&\quad \le \left|(\mathbb{P}_n - P)[\Omega(Y)\{Z(Y) - \eta_0(Y)\}\Delta]\right| + \left|\mathbb{P}_n[\Omega(Y)\{\eta(Y) - \eta_0(Y)\}\Delta]\right| \\
&\qquad + \left|(\mathbb{P}_n - P) \int \Omega(t)\{Z(t) - \eta_0(t)\}1(Y \ge t)\theta Z(t)\, dt\right| \\
&\qquad + \left|\mathbb{P}_n \int \Omega(t)\{\eta(t) - \eta_0(t)\}1(Y \ge t)\theta Z(t)\, dt\right| \\
&\quad \le \left|(\mathbb{P}_n - P)[\Omega(Y)\{Z(Y) - \eta_0(Y)\}\Delta]\right| \\
&\qquad + \left|(\mathbb{P}_n - P) \int \Omega(t)\{Z(t) - \eta_0(t)\}1(Y \ge t)\theta Z(t)\, dt\right| \\
&\qquad + \delta_n \mathbb{P}_n \left\{\Omega(Y)\Delta + \int \Omega(t)1(Y \ge t)|\theta Z(t)|\, dt\right\},
\end{aligned}
$$

in which the first two terms on the right hand side of the last inequality converge to zero in probability by the Weak Law of Large Numbers, and the third term converges to zero because $\delta_n \to 0$. We then have the desired result by Lemma 1.

B.6. Proof of Proposition 6

As in the proof of Proposition 3.3, the Fréchet differentiability of $\{\Psi(\theta, \eta) : \theta \in \Theta_0, \eta \in \mathcal{H}_0\}$ can be verified via direct calculation. The set $\{\Omega(t)\Delta\{Z(t) - \eta(t)\} : \eta \in \mathcal{H}_0, 0 \le t \le \tau\}$ is Donsker, thus we only need to show the class of functions $\{\int_0^\tau \Omega(t)\{Z(t) - \eta(t)\}1(Y \ge t)\theta Z(t)dt : \theta \in \Theta_0, \eta \in \mathcal{H}_0\}$ is Donsker, here $\mathcal{H}_0$ is reduced from that in the proof of Proposition 3.3. Let $f = \int_0^\tau \Omega(t)\{Z(t) -$

$\eta(t)\}Z(t)1(Y \geq t)dt$ and

$$f^m = \sum_{i=1}^{m} \Omega(t_i)\{Z(t_i) - \eta(t_i)\}Z(t_i)1(Y \geq t_i)(t_{i+1} - t_i) = \sum_{i=1}^{m} f_i\lambda_i\,,$$

where

$$f_i = \Omega(t_i)\{Z(t_i) - \eta(t_i)\}Z(t_i)1(Y \geq t_i), \quad \lambda_i = t_{i+1} - t_i,$$

and $\{(t_1, t_2], \ldots, (t_m, \tau]\}$ forms a partition of the interval $(0, \tau]$. The set $\{f^m\}$ is the convex hull of $\mathcal{F} = \{f_i\}$, and thus a Donsker class by Theorem 2.10.3 in van der Vaart and Wellner (1996) since $\mathcal{F}$ is Donsker. Now we know that $f^m \to f$ both pointwise and in $L_2(P)$ by the boundedness of Y and η, then $\{f(\cdot)\}$ is Donsker by Theorem 2.10.2 in van der Vaart and Wellner (1996).

We now calculate the right side of equation (2.9). Direct calculation yields

$$\begin{aligned} n^{1/2}\dot{\Psi}_\eta(\theta_0, \eta_0)[\tilde{\eta}_n - \eta_0] = &-n^{1/2}P[\{\tilde{\eta}_n(Y) - \eta_0(Y)\}\Delta] \\ &+n^{1/2}P\bigg[\int \{\tilde{\eta}_n(t) - \eta_0(t)\}1(Y \geq t)\theta_0 Z(t)dt\bigg] \end{aligned} \tag{B.1}$$

by applying $E(\Omega|X) = 1$. Let $d^{(0)}(t) \equiv P\{W(t)1(Y \geq t)\} = P\{1(Y \geq t)\}$ and $d^{(1)}(t) \equiv P\{W(t)Z(t)1(Y \geq t)\} = P\{Z(t)1(Y \geq t)\}$, where $E(W|X) = 1$. The first term on the right side of (B.1) can be written as

$$\begin{aligned} &- n^{1/2}P[\{\tilde{\eta}_n(Y) - \eta_0(Y)\}\Delta] \\ &\quad = \; - \mathbb{G}_n\bigg[\int W(t)\{Z(t) - \eta_0(t)\}1(Y \geq t)d^{(0)}(t)^{-1}dP_{Y,\Delta}(t, 1)\bigg] + o_{p^*}(1) \\ &\quad = \; - \mathbb{G}_n\bigg[\int W(t)\{Z(t) - \eta_0(t)\}1(Y \geq t)\lambda_0(t)dt \\ &\qquad + \int W(t)\{Z(t) - \eta_0(t)\}1(Y \geq t)\theta_0\eta_0(t)dt\bigg] + o_{p^*}(1) \end{aligned}$$

since, from the joint distribution of $(Y, \Delta, \bar{Z}(Y))$, we have

$$\begin{aligned} \frac{dP_{Y,\Delta}(t, 1)}{dt} &= \int \{\lambda_0(t) + \theta_0 z(t)\}\{1 - F(t|\bar{z}(t))\}\{1 - G(t^-|\bar{z}(t))\}dH_t(\bar{z}(t)) \\ &= \lambda_0(t)P\{1(Y \geq t)\} + \theta_0 P\{Z(t)1(Y \geq t)\} \\ &= \lambda_0(t)d^{(0)}(t) + \theta_0 d^{(1)}(t)\,. \end{aligned}$$

From the proof of Proposition 1 we have

$$n^{1/2}\{\tilde{\eta}_n(t) - \eta_0(t)\} = d^{(0)}(t)^{-1}\mathbb{G}_n\Big[W(t)\{Z(t) - \eta_0(t)\}1(Y \geq t)\Big] + o_{p^*}(1)\,,$$

so the second term on the right side of (B.1) can be rewritten as

$$\begin{aligned}
&\int n^{1/2}\{\tilde{\eta}_n(t)-\eta_0(t)\}P\{1(Y\geq t)\theta_0 Z(t)\}\,dt\\
&\quad=\int d^{(0)}(t)^{-1}\mathbb{G}_n\Big[W(t)\{Z(t)-\eta_0(t)\}1(Y\geq t)\Big]\theta_0 d^{(1)}(t)\,dt+o_{p^*}(1)\\
&\quad=\mathbb{G}_n\bigg[\int W(t)\{Z(t)-\eta_0(t)\}1(Y\geq t)\theta_0\eta_0(t)\,dt\bigg]+o_{p^*}(1).
\end{aligned}$$

Thus from (2.9) we obtain

$$\begin{aligned}
n^{1/2}(\tilde{\theta}_n-\theta_0)=&\bigg[P\bigg\{\int\Omega(t)\{Z(t)-\eta_0(t)\}1(Y\geq t)Z(t)\,dt\bigg\}\bigg]^{-1}\\
&\times\mathbb{G}_n\bigg[\Omega(Y)\{Z(Y)-\eta_0(Y)\}\Delta\\
&-\int\{\Omega(t)\theta_0 Z(t)+W(t)\lambda_0(t)\}\{Z(t)-\eta_0(t)\}1(Y\geq t)dt\bigg]+o_{p^*}(1),
\end{aligned}$$

which is asymptotic normal by the Central Limit Theorem. This asymptotic representation reduces to that in Kulich and Lin (2000) when $\Omega_i = W_i$. Again, we do not require Ω_i and W_i to be predictable.

References

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

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* Johns Hopkins University Press, Baltimore.

Borgan, O., Langholz, B., Samuelsen, S. O., Goldstein, L., and Pogoda, J. (2000). Exposure stratified case-cohort designs. *Lifetime Data Anal.* **6**, 39-58.

Borgan, O., Goldstein, L., and Langholz, B. (1995). Methods for the analysis of sampled cohort data in the Cox proportional hazards model. *Ann. Statist.* **23**, 1749-1778.

Breslow, N. E. and Wellner, J. A. (2007). Weighted likelihood for semiparametric models and two-phase stratified samples, with applications to Cox regression. *Scand. J. Statist.* **34**, 86-102.

Chatterjee, N., Chen, Y.-H., and Breslow, N. E. (2003). A pseudoscore estimator for regression problems with two-phase sampling. *J. Amer. Statist. Assoc.* **98**, 158-168.

Chen, X., Linton, O. and Van Keilegom, I. (2003). Estimation of semiparametric models when the criterion function is not smooth. *Econometrica* **71**, 1591-1608.

Chen, K. and Lo, S.-H. (1999). Case-cohort and case-control analysis with Cox's model. *Biometrika* **86**, 755-764.

Chen, Y.-H. and Zucker D. M. (2009). Case-cohort analysis with semiparametric transformation models. *J. Ststist. Plann. Inference* **139**, 3706-3717.

Chung, K. L. and Williams, R. J. (1990). *Introduction to Stochastic Integration.* 2nd edition. Birkhäuser, Boston.

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

Ding, Y. and Nan, B. (2011). A sieve M-theorem for bundled parameters in semiparametric models, with application to the efficient estimation in a linear model for censored data. *Ann. Statist.* **39**, 3032-3061.

Goldstein, L. and Langholz, B. (1992). Asymptotic theory for nested case-control sampling in the Cox regression model. *Ann. Statist.* **20**, 1903-1928.

Hu, H. (1998). Large sample theory for pseudo-maximum likelihood estimates in semiparametric models. Ph.D. dissertation. Dept. Statistics, Univ. Washington.

Huang, J. and Wellner, J. A. (1995). Asymptotic normality of the NPMLE of linear functionals for interval censored data, case 1. *Statistica Neerlandica* **49**, 153-163.

Huang, J. and Wellner, J. A. (1997). Interval censored survival data: A review of recent progress. In *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis Lecture Notes in Statistics* **123**, 123-169. Springer, New York.

Kalbfleisch, J. D. and Lawless, J. F. (1988). Likelihood analysis of multi-state models for disease incidence and mortality. *Stat. Med.* **7**, 149-160.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd Ed. John Wiley, New York.

Kong, L., Cai, J. and Sen, P. K. (2004). Weighted estimating equations for semiparametric transformation models with censored data from a case-cohort design. *Biometrika* **91**, 305-319.

Kulich, M. and Lin, D. Y. (2000). Additive hazards regression for case-cohort studies. *Biometrika* **87**, 73-87.

Kulich, M. and Lin, D. Y. (2004). Improving the efficiency of relative-risk estimation in case-cohort studies. *J. Amer. Statist. Assoc.* **99**, 832-844.

Li, Z., Gilbert, P. and Nan, B. (2008). Weighted likelihood method for grouped survival data in case-cohort studies with application to HIV vaccine trials. *Biometrics* **64**, 1247-1255.

Li, Z. and Nan, B. (2011). Relative risk regression for current status data in case-cohort studies. *Canad. J. Statist.* **39**, 557-577.

Lin, D. Y. and Ying, Z. (1994) Semiparametric analysis for the additive risk model. *Biometrika* **81**, 61-71.

Lindsay, B. G. (1987). Composite likelihood methods. *In Statistical Inference from Stochastic Processes, Contemp. Math.* **80**, 221-239.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. 2nd edition. John Wiley, New Jersey.

Lu W. and Tsiatis, A. A. (2006). Semiparametric transformation models for the case-cohort study. *Biometrika* **93**, 207-214.

Nan, B. (2004). Efficient estimation for case-cohort studies. *Canad. J. Statist.* **32**, 403-419.

Nan, B., Emond, M., and Wellner, J. A. (2004). Information bounds for Cox regression models with missing data. *Ann. Statist.* **32**, 723-753.

Nan, B., Kalbfleisch, J. D. and Yu, M. (2009). Asymptotic theory for the semiparametric accelerated failure time model with missing data. *Ann. Statist.* **37**, 2351-2376.

Nan, B., Yu, M., and Kalbfleisch, J. D. (2006). Censored linear regression for case-cohort studies. *Biometrika* **93**, 747-762.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica* **62**, 1349-1382.

Pepe, M. S. and Fleming, T. R. (1991). A non-parametric method for dealing with mismeasured covariate data. *J. Amer. Statist. Assoc.* **86**, 108-113.

Præstgaard, J. and Wellner, J. A. (1993). Exchangeably weighted bootstraps of the general empirical process. *Ann. Probab.* **21**, 2053-2086.

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

Protter, P. E. (2004). *Stochastic Integration and Differential Equations*. 2nd edition. Springer-Verlag, Heidelberg.

Pugh, M., Robins, J., Lipsitz, S. and Harrington, D. (1992). Inference in the Cox proportional hazards model with missing covariates. Technical Report 758Z, Department of Biostatistics, Harvard School of Public Health.

Reilly, M. and Pepe, M. S. (1995). A mean score method for missing and auxilliary covariate data in regression models. *Biometrika* **82**, 299-314.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89**, 846-866.

Saegusa, T. and Wellner, J. A. (2012). Weighted likelihood estimation under two-phase sampling. Technical Report, University of Washington.

Self, S. G. and Prentice, R. L. (1988). Asymptotic distribution theory and efficiency results for case-cohort studies. *Ann. Statist.* **16**, 64-81.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag, New York.

van der Vaart, A. W. and Wellner, J. A. (2007). Empirical processes indexed by estimated functions. In *Asymptotics: Particles, Processes and Inverse Problems, IMS Lecture Notes Monogr. Ser.* **55**, 234-252. Inst. Math. Statist.

Varin, C., Reid, N. and Firth, D. (2011). An overview of composite likelihood methods. *Statist. Sinica* **21**, 5-42.

Wong, W. H. and Severini, T. A. (1991). On maximum likelihood estimation in infinite dimensional parameter space. *Ann. Statist.* **16**, 603-632.

Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109-2029, USA.

E-mail: bnan@umich.edu

Department of Statistics, University of Washington, Seattle, WA 98195-4322, USA.

E-mail: jaw@stat.washington.edu

(Received April 2012; accepted October 2012)