SOME NEW INSIGHTS ABOUT THE ACCELERATED FAILURE TIME MODEL

by

Ying Ding

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2010

Doctoral Committee:

Associate Professor Bin Nan, Chair
Professor John D. Kalbfleisch
Associate Professor Ji Zhu
Research Associate Professor Mousumi Banerjee

To Wei and my parents

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere appreciation to my dissertation advisor Dr. Bin Nan for his invaluable guidance for the completion of this thesis. He brought me to the area of survival analysis through his insightful ideas, broad knowledge and great patience, which I have enjoyed very much and learned a lot from. I benefit so much from his constructive suggestions and profound insight into statistics. Without him, this dissertation would not be in the current form.

My deep gratitude also goes to Dr. John D. Kalbfleisch, Dr. Mousumi Banerjee, and Dr. Ji Zhu for providing valuable comments on my thesis and being on the committee. I always feel inspired and encouraged after discussing the research ideas with them. Special thanks are due to Dr. Mousumi Banerjee who has provided me a great opportunity to work on an interesting survival tree project and offered me financial support. In addition, I would like to thank Dr. Debashis Ghosh for being my academic advisor in my first year of Ph.D. study and it has been a very pleasant experience working as a research assistant under his supervision.

I am very thankful to all knowledgable faculty, helpful staff members and smart fellow students in the Department of Biostatistics at the University of Michigan. I feel extremely lucky and honored to be a graduate student here.

This thesis is dedicated to my husband Wei Chen, my inspiring parents and Wei's parents. Their boundless love and consistent encouragement are the biggest part of every accomplishment that I have achieved. Wei is always extraordinary helpful in

programming. He made my three years' Ph.D. study in the University of Michigan the most brilliant time in my life. My parents shall be very pleased to see their daughter achieve her educational goals after over twenty years of nurturing.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

The semiparametric linear model is an important alternative to the Cox proportional hazards model for censored survival outcomes. In this dissertation, we provide some new insights for the parameter estimators and their asymptotic properties in the semiparametric linear model with censored data.

In Chapter 2, we have shown that in a linear regression model, where the outcome variable is subject to right censoring and the error distribution is unspecified, the intercept parameter is consistent and asymptotically normal when the support of some covariates with nonzero coefficients is unbounded. This holds even with finite follow-up times. In a practical setting, it makes the prediction of survival time possible under a linear regression model when the covariate range is wide. Without the commonly assumed regularity condition of bounded covariates, we have also shown that the slope estimators obtained by solving the Gehan-weighted rank based estimating equation are consistent and asymptotically normal, which provides a crucial condition for the asymptotic properties of the intercept estimator.

In Chapter 3, we have proposed a new approach to estimate the slope parameters in the semiparametric linear model by directly maximizing the log likelihood function in a sieve space, in which the log hazard function of the error term is approximated by B-splines. The maximization can be achieved through the gradient-based search algorithm over the sieve space. The resulting slope estimators have been shown to be consistent and asymptotically normal. In addition, the limiting covariance matrix of the proposed estimators reaches the semiparametric efficiency bound and can be estimated nicely by inverting either the information matrix based on the efficient score function of the regression parameters or the observed information matrix for all parameters including the “nuisance” parameters for estimating the log hazard function.

CHAPTER I

Introduction

Censored survival data often appear in biomedical research. The Cox proportional hazards model (Cox 1972) and the semiparametric transformed linear model are the two major approaches for analyzing such data. Instead of modeling the hazard function, the later model postulates a direct relationship between failure time and predictors. In this dissertation, we focus on the semiparametric linear model with unspecified error distribution and provide some new insights for the parameter estimates and corresponding asymptotic properties.

1.1 Estimation of the Intercept in the Accelerated Failure Time Model

The accelerated failure time (AFT) model assumes

$$h(T_i) = \alpha_0 + X_i'\beta_0 + \zeta_i, \quad i = 1, \cdots, n,$$

where T_i are the survival times that are subject to censoring, h is a pre-specified monotone function, e.g. the logarithm function, α_0 is the intercept, β_0 is the slope, and ζ_i are the independent random errors following an unknown distribution that is assumed to have zero mean and bounded variance.

There is a rich literature on the slope parameter estimation in the above regression model. We provide a brief review about existing slope estimating methods in Section

3.2.1. A good summary can be found in Chapter 7 of Kalbfleisch and Prentice (2002).

However, the estimation of intercept has not been thoroughly studied mostly because the follow-up time is usually finite in practice so the intercept, directly related to the mean survival time, is likely to be underestimated. Buckley and James (1979) first claimed that the intercept cannot be estimated consistently due to the existence of censoring. However, through simulation studies, Schneider and Weissfeld (1986) and Heller and Simonoff (1990) found that the intercept can sometimes be estimated very well using the Buckley-James method. Motivated by the existing literature of consistent estimation of the mean survival time, which is reviewed in Section 2.1, in this thesis we show that the aforementioned concern is not always necessary. In fact, the intercept can be consistently estimated and an "approximated" estimator is asymptotically normal when the support of some covariates with nonzero coefficients is unbounded. This result holds even with limited follow-up times, which is always the case in most human disease studies. It makes the prediction of survival time possible under a linear regression model when the covariate support is wide in the practice.

Without the commonly assumed regularity condition of bounded covariates, additional consideration on the slope estimation is required since its theoretical developments to date are mostly under the bounded covariates assumption. We next show that, without the restriction of bounded covariates, the slope estimators obtained by solving the rank-based estimating equations with Gehan weights are still consistent and asymptotically normal, which provides a crucial condition for the asymptotic properties of the intercept estimator.

The theoretical findings are further verified for finite samples by simulation studies. Simulations also show that, when both models are correctly specified, the semiparametric linear model yields reasonable mean square prediction errors and outperforms the Cox model for censored data, particularly for heavy censoring and short follow-up time. An illustrative example using the AFT model to predict the failure times is given in Section 2.5.

1.2 Sieve Maximum Likelihood Estimation of the Slope Parameter in the AFT Model

It is well known that the partial likelihood estimator in the Cox proportional hazards model is semiparametric efficient. Despite previous research efforts, developing an efficient estimator in the semiparametric linear model is not completely satisfactory. In this thesis, we propose a different approach to the existing semiparametric estimating equation methods that are known to be statistically inefficient. Specifically, we directly maximize the log likelihood function over a sieve space, in which the log hazard function is approximated by a linear span of a set of basis functions in that sieve space. Such bases can be B-splines, trigonometric polynomials, hermite polynomials or wavelets. We consider the B-splines basis (Schumaker 1981) in this proposed method because of its computational convenience. The numerical implementation can be achieved through the conventional gradient-based search algorithms such as the Newton-Raphson algorithm.

We show that the proposed estimators are consistent and asymptotically normal. Moreover, the limiting covariance matrix of the estimators reaches the semiparametric efficiency bound. The proof of the asymptotic normality and semiparametric efficiency of the proposed estimators is based on our extended general theorem on the asymptotic normality of semiparametric M-estimators, where the infinite dimensional nuisance parameter is a function of the parameters of interest. This is the case for the semiparametric linear regression model since the likelihood is built upon residuals, which is a function of the slope parameters.

We propose two ways to estimate the variance of the estimators. The first approach is to invert the information matrix based on the efficient score function of the slope parameters derived by Ritov and Wellner (1988). The second approach is to invert the observed information matrix of all parameters including the "nuisance" parameters in the sieve space for the log hazard function. Although the second method does not have a theoretical justification, simulation studies show that it is numerically robust and yields quite similar variance estimates as the first method. Simulation studies also demonstrate that the proposed method performs well in practical settings and yields more efficient estimates than existing estimating equation based methods. Illustrations with two real data examples are provided in Section 3.5.

CHAPTER II

Asymptotics of the Intercept Estimator in the Semiparametric Accelerated Failure Time Model

2.1 Introduction

As an important alternative to the Cox model (Cox 1972), the linear regression model for transformed censored survival data including the accelerated failure time model (Kalbfleisch and Prentice 2002) as a special case has been extensively studied in recent years, see e.g. Wei et al. (1990), Tsiatis (1990), Ritov (1990), Ying (1993), and Jin et al. (2003), among many others. This type of model appeals in many ways because it models the failure time directly and thus has a more intuitive interpretation. In situations where proportional hazards assumption is violated, this model may provide more accurate summarization of the data. Since it directly models the failure time, there might be chances that the linear model can be used to predict the failure time in a straightforward way.

The study of such a linear model has primarily focused on the slope parameter estimation. Commonly used estimating methods include: the Buckley-James method (Buckley and James 1979) that imputes the censored failure time by its estimated conditional expectation given the corresponding censoring time and covariates, the weighted least squares method of Stute (1993, 1996) with weights obtained from the Kaplan-Meier estimator for the transformed survival time, and the rank-based method (Prentice 1978; Tsiatis 1990; Ying 1993) that is derived by using linear rank tests for the right censored data. Ritov (1990) showed that the class of weighted rank-based estimating functions of Tsiatis (1990) is asymptotically equivalent to the class of Buckley-James estimating functions on transformed residuals.

The estimation of intercept when the error distribution is unspecified, however, has not been thoroughly studied mostly because the follow-up time is usually finite in practice so the intercept, directly related to the mean survival time, is generally believed to be underestimated. Obviously, a good prediction of survival time requires a good estimator for the intercept parameter which is the expectation of the error term in the semiparametric linear model. The inconsistency of the intercept estimator was first claimed by Buckley and James (1979). In some of their simulations, however, Schneider and Weissfeld (1986) and Heller and Simonoff (1990) found that the intercept can sometimes be estimated quite well using the Buckley-James method. Based on the work of Susarla and Van Ryzin (1980) and Susarla et al. (1984), Wang et al. (2008) conjectured that the intercept can be consistently estimated when the supports of some covariates are not restricted to finite intervals. In this chapter, we confirm such a conjecture by formally establishing the consistency result for the intercept estimator, as well as the asymptotic normality for an "approximated" estimator. This result makes the prediction of survival time possible under a linear regression model when covariate support is wide in a practical setting.

Without the presence of covariates, using an integration by parts argument with a truncation technique, Susarla and Van Ryzin (1980) showed that when the support of censoring time distribution contains the support of failure time distribution together with appropriate assumptions for the tail probability, the mean failure time estimation based on a Kaplan-Meier type estimator is consistent almost surely under random censoring. Using the reverse martingale approach, Stute and Wang (1993) established more general strong consistency results including the mean failure time estimation without using the truncation argument. When covariates are present and a linear model is considered for the transformed failure time, the intercept estimation is equivalent to the mean failure time estimation on the residual scale if true values of the slope parameters are given. In reality, however, the slope parameters need to be estimated, which dramatically complicates the study of asymptotic properties of the intercept estimation. For the consistency of intercept estimation when slopes are estimated, to the best of our knowledge, we are only aware of Lai and Ying (1991) who assumed bounded covariates, bounded support of the failure time distribution and wider support of the censoring time distribution. Their latter assumption, however, is often violated in practice due to the nature of limited follow-up time in, for example, most of the human disease studies. Instead of assuming wider support of the censoring time distribution, we consider the setting that the supports of some covariates with nonzero coefficients are not restricted to finite intervals, which however requires additional consideration on the slope estimation because its theoretical developments to date are primarily under the assumption of bounded covariates. The unbounded covariate support is a technical condition, and corresponds to the practical situation where the ranges of the explanatory covariates are wide.

The rest of the chapter is organized as follows. In Section 2.2 we present a general strong consistency property of the intercept estimator under the assumption of unbounded covariates, followed by the asymptotic normality result for an "approximated" intercept estimator where the modification is applied to simplify the technical derivation. In Section 2.3 we present the both in probability and almost sure consistency properties as well as the asymptotic normality result for the Gehan-weighted rank based slope estimators without assuming bounded covariates. In Section 2.4 we conduct simulation studies by varying the covariate support and the censoring rate under different error distributions with different sample sizes. We also compare the failure time prediction performance of the semiparametric model to that of the Cox model under the standard extreme value error distribution for which both models fit the data correctly. In Section 2.5 we provide an application to a major medical study, the Mayo primary biliary cirrhosis (PBC) study, for illustration. We provide some concluding remarks in Section 2.6. Proofs of the technical results heavily depend on the empirical process theory and are deferred to the Appendix.

2.2 Intercept Estimation

2.2.1 The Model and Notation

Consider the linear regression model:

$$T_i = \alpha_0 + X_i^{'}\beta_0 + \zeta_i, \quad i = 1, \ldots, n, \tag{2.1}$$

where ζ_i, $i = 1, \ldots, n$, are independent and identically distributed (i.i.d.) with zero mean. The response variable T_i for the ith subject is the failure time transformed by a known monotone function, e.g. the logarithm transformation that yields the so-called accelerated failure time model (Kalbfleisch and Prentice 2002, Chap. 7). When T_i is subject to right censoring, we only observe (Y_i, Δ_i, X_i), where $Y_i = \min(T_i, C_i)$, C_i is the censoring time transformed by the same function that yields T_i, and $\Delta_i = 1(T_i \leq C_i)$. Here we assume that (X_i, C_i), $i = 1, \ldots, n$, are i.i.d. and independent of ζ_i.

Throughout the sequel we consider one-dimensional β_0 for notational simplicity and assume that its parameter space $\mathcal{B}$ is compact. For any $\beta \in \mathcal{B}$ we denote

$$e_{\beta,i} = T_i - \beta X_i, \quad e_{0,i} = T_i - \beta_0 X_i = \alpha_0 + \zeta_i,$$

and

$$\epsilon_{\beta,i} = Y_i - \beta X_i, \quad \epsilon_{0,i} = Y_i - \beta_0 X_i.$$

Here, $e_{\beta,i}$ are the transformed failure time in the residual scale with β_0 being replaced by β, $\epsilon_{\beta,i}$ is the corresponding observed time in the residual scale for a fixed β, and $e_{0,i}$ is the error term that has absorbed the intercept in model (2.1). We shall use F and G to denote the distribution functions of $e_{0,i}$ and C_i, and f and g to denote their density functions, respectively. Now we adopt the empirical process notation of van der Vaart and Wellner (1996). In particular, for a function f of a random variable U that follows distribution P,

$$\begin{aligned} Pf &= \int f(u)\, dP(u), \\ \mathbb{P}_n f &= n^{-1} \sum_{i=1}^{n} f(U_i), \\ \mathbb{G}_n f &= n^{1/2}(\mathbb{P}_n - P)f, \end{aligned}$$

and refer all the details to the reference. Set $\epsilon_\beta = Y - \beta X$ and $\epsilon_0 = Y - \beta_0 X$, and define

$$H_n^{(0)}(\beta, s) = \mathbb{P}_n\{1(\epsilon_\beta \le s, \Delta = 1)\}, \quad h^{(0)}(\beta, s) = P\{1(\epsilon_\beta \le s, \Delta = 1)\}; \tag{2.2}$$

$$H_n^{(1)}(\beta, s) = \mathbb{P}_n\{1(\epsilon_\beta \ge s)\}, \quad h^{(1)}(\beta, s) = P\{1(\epsilon_\beta \ge s)\}. \tag{2.3}$$

Since $\alpha_0 = Ee_{0,i} = \int_{-\infty}^{\infty} t\, dF(t)$, if the slope β_0 is known, then a natural estimator of α_0 is given by

$$\hat{\alpha}_n = \int_{-\infty}^{\infty} t\, d\hat{F}_n(t), \tag{2.4}$$

where $\hat{F}_n(t)$ is the Kaplan-Meier estimator of the distribution function $F(t)$ of $e_0 = T - \beta_0 X$. In a regression setting, however, β_0 is unknown and hence has to be estimated. Let $\hat{\beta}_n$ be an estimator of β_0, a direct extension of (2.4) yields

$$\hat{\alpha}_{n,\hat{\beta}_n} = \int_{-\infty}^{\infty} t\, d\hat{F}_{n,\hat{\beta}_n}(t), \tag{2.5}$$

where $\hat{F}_{n,\beta}(t)$ is the Kaplan-Meier estimator of the distribution function $F_\beta(t)$ of $e_\beta = T - \beta X$ and is given by

$$\hat{F}_{n,\beta}(t) = 1 - \prod_{i:\epsilon_{\beta,i} \le t} \left\{ 1 - \frac{\Delta_i / n}{H_n^{(1)}(\beta, \epsilon_{\beta,i})} \right\}. \tag{2.6}$$

Note that the above estimator does not automatically provide a consistent estimator of $F_\beta(t)$ because $T - \beta X$ and $C - \beta X$ are not independent except when $\beta = \beta_0$. We will follow the method used by Lai and Ying (1991) to argue that $\hat{F}_{n,\hat{\beta}_n}(t)$ does converge to $F(t)$ when $\hat{\beta}_n$ converges to β_0 with a certain polynomial rate.

When there is no covariates (equivalently $\beta_0 = 0$) or β_0 is given, Susarla and Van Ryzin (1980) and Stute and Wang (1993) studied the asymptotic properties of the mean survival time estimator (2.4). They provided the following key sufficient condition

$$\{t : t \in \text{ the support of } T - \beta_0 X\} \subseteq \{t : t \in \text{ the support of } C - \beta_0 X\} \tag{2.7}$$

for the consistency of (2.4). Now we replace β_0 by its estimator $\hat{\beta}_n$ and want to show the consistency of (2.5). The proof of Stute and Wang (1993) for the consistence of the mean survival time estimation makes use of the martingale theory that cannot be directly adopted here due to the dependence between $T - \beta X$ and $C - \beta X$ when $\beta \neq \beta_0$. We shall use the empirical process theory as well as the properties of stochastic integrals with censored data in Lai and Ying (1988) to show the desirable result.

2.2.2 Consistency

First we introduce the following regularity conditions:

Condition 1. The covariates X_i are i.i.d. random variables with finite second moment.

Condition 2. The error e_0's density f and its derivative $\dot{f}$ are bounded and

$$\int_{-\infty}^{\infty} \big(\dot{f}(t)/f(t)\big)^2 f(t)\ dt < \infty.$$

Condition 3. The conditional density of C given X is continuous and uniformly bounded for all possible values of X. That is,

$$\sup_{x\in\mathcal{X},\ t\in\mathcal{C}} g_{C|X}(t \mid X = x) < \infty,$$

where $\mathcal{X}$ and $\mathcal{C}$ denote the support of X and C, respectively.

Condition 4. The error e_0 has a finite second moment, i.e., $Ee_0^2 < \infty$.

Condition 1 is different to the common assumption of bounded covariates in Tsiatis (1990), Lai and Ying (1991), Ying (1993), and many others. Here we do not assume bounded covariates. Instead, we only assume finite second moment. Hence, even with a short follow-up time, the support of the censoring time in the residual scale can be extended to infinity provided that the support of X is the real line and $\beta_0 \neq 0$, which yields that the supports of e_0 and $C - \beta_0 X$ are equivalent and thus the sufficient condition (2.7) is satisfied. Condition 2 is exactly the same as Condition 2 in Ying (1993). Condition 3 implies Condition 3 in Ying (1993) and also Condition (3.5) in Lai and Ying (1991) when X_i are bounded. Condition 4 implies Condition 4 in Ying (1993) with $\theta_0 = 2$.

We then have the consistency results given in the following Theorems 2.2.1 and 2.2.2. We omit the constants in front of the rate expressions to simplify the notation.

Theorem 2.2.1. Suppose Conditions 1-3 hold, and define

$$F(\beta, t) = 1 - \exp\bigg\{ - \int_{u\le t} \frac{dh^{(0)}(\beta, u)}{h^{(1)}(\beta, u)} \bigg\}, \tag{2.8}$$

with $h^{(0)}(\beta,u)$ and $h^{(1)}(\beta,u)$ being defined in (2.2) and (2.3), respectively. Then for every $\varepsilon > 0$, with probability 1 we have

$$\sup\bigl\{|\hat{F}_{n,\beta}(t) - F(\beta,t)| : \beta \in \mathcal{B}, H_n^{(1)}(\beta,t) \geq n^{-\varepsilon}\bigr\} = o(n^{-\frac{1}{2}+3\varepsilon}), \tag{2.9}$$

$$\sup\bigl\{|F(\beta,t) - F(t)| : |\beta-\beta_0| \leq n^{-3\varepsilon}, h^{(1)}(\beta,t) \geq n^{-\varepsilon}\bigr\} = O(n^{-\varepsilon}), \tag{2.10}$$

where $\hat{F}_{n,\beta}(t)$ is given in (2.6). In addition, for every $0 < \varepsilon \leq \frac{1}{8}$, with probability 1 we have

$$\sup\bigl\{|\hat{F}_{n,\beta}(t) - F(t)| : |\beta-\beta_0| \leq n^{-3\varepsilon}, H_n^{(1)}(\beta,t) \geq n^{-\varepsilon}\bigr\} = O(n^{-\varepsilon}). \tag{2.11}$$

Introduced by Lai and Ying (1991), $F(\beta,t)$ defined in (2.8) is an important intermediate quantity. On one hand, it is the limit of the Kaplan-Meier estimator $\hat{F}_{n,\beta}(t)$ for a fixed β; on the other hand, it equals to $F(t)$, the true distribution function of the error e_0, when β is replaced by the true slope β_0 in (2.8).

Theorem 2.2.2. Suppose Conditions 1-4 hold, and in addition, assume $\beta_0 \neq 0$ and that the support of X, $\mathcal{X}$ is the whole real line, i.e., $f_X(x) > 0$ for all $-\infty < x < \infty$. Then for every $0 < \varepsilon \leq \frac{1}{8}$, with probability 1 we have

$$\sup\left\{\left|\int_{-\infty}^{\infty} t\, d\hat{F}_{n,\beta}(t) - \alpha_0\right| : |\beta-\beta_0| \leq n^{-3\varepsilon}, H_n^{(1)}(\beta,t) \geq n^{-\varepsilon}\right\} = o(1). \tag{2.12}$$

Theorem 2.2.2 implies that $\hat{\alpha}_{n,\hat{\beta}_n} = \int_{-\infty}^{\infty} t\ d\hat{F}_{n,\hat{\beta}_n}(t)$ is a consistent estimator of the intercept α_0 when $\hat{\beta}_n$ is a consistent estimator of the slope β_0 with polynomial convergence rate. Define

$$T_n = \sup\bigl\{t : H_n^{(1)}(\beta,t) \geq n^{-\varepsilon}, |\beta-\beta_0| \leq n^{-3\varepsilon}\bigr\}, \tag{2.13}$$

so the restriction on t inside the supremum is equivalent to set the Kaplan-Meier estimator $\hat{F}_{n,\beta}(t)$ to 1 for $t > T_n$, and thus $\int_{-\infty}^{\infty} t\ d\hat{F}_{n,\beta}(t) = \int_{-\infty}^{T_n} t\ d\hat{F}_{n,\beta}(t)$ within the set $\{t : H_n^{(1)}(\beta,t) \geq n^{-\varepsilon}, |\beta-\beta_0| \leq n^{-3\varepsilon}\}$. Then the unbounded covariate support assumption plays an important role to guarantee that $T_n \to \infty$ as $n \to \infty$, and hence $\int_{-\infty}^{T_n} t\ d\hat{F}_{n,\beta}(t) \to \int_{-\infty}^{\infty} t\ dF(t)$ almost surely when $|\beta - \beta_0| \leq n^{-3\varepsilon}$. Susarla et al. (1984) showed that the above $\hat{\alpha}_{n,\hat{\beta}_n}$ is identical to the Buckley-James estimator of α_0 for a fixed $\hat{\beta}_n$.

2.2.3 Asymptotic Normality

In the situation of no covariates or equivalently $\beta = \beta_0$, to the best of our knowledge, we are only aware of Susarla and Van Ryzin (1980), who proved the asymptotic normality result for the mean survival time estimator. A truncation technique was used in their proof and some stringent conditions on the tail probability were introduced, which are often very difficult to justify. Even for the Kaplan-Meier estimator itself without covariates, there has also been much effort to show its asymptotic normality for the past two decades. See e.g. Wellner (2007). When covariates are present, besides the complexity from the estimated slope parameter $\hat{\beta}_n$, another challenge comes from the integrand in the intercept estimator in (2.5), which integrates from $-\infty$ up to T_n, with $T_n \to \infty$ as $n \to \infty$. In this section, we try to avoid assuming the hard-to-justify technical condition for the tail probability of the error term and to simply the derivation, therefore, instead of showing the asymptotic normality for the original intercept estimator in (2.5), we consider the following "approximated" intercept estimator

$$\hat{\alpha}^*_{n,\hat{\beta}_n} = \int_{S^*}^{T^*} t\ d\hat{F}_{n,\hat{\beta}_n}, \tag{2.14}$$

where S^* and T^* are any fixed time points with $-\infty < S^* < T^* < \infty$. $\hat{\beta}_n$ is an estimate of β which is consistent and asymptotically normal. The asymptotic normality result for the estimator in (2.14) is stated in the following theorem.

Theorem 2.2.3. Suppose Conditions 1-4 hold, and in addition, assume $\beta_0 \neq 0$ and that the support of X is the whole real line, i.e., $f_X(x) > 0$ for all $-\infty < x < \infty$. Let $\hat{\alpha}^*_{n,\hat{\beta}_n}$ be the "approximated" intercept estimator defined in (2.14) with $\hat{\beta}_n$ being a consistent and asymptotically normal estimate of β. Let $\alpha^*_0 = \int_{S^*}^{T^*} t\ dF(t)$, then $n^{1/2}(\hat{\alpha}^*_{n,\hat{\beta}_n} - \alpha^*_0)$ is asymptotically normal with the following asymptotic representation

$$
\begin{aligned}
(2.15)\, n^{1/2}(\hat{\alpha}^*_{n,\hat{\beta}_n} - \alpha^*_0)& \\
= \ \mathbb{G}_n \Big\{ & T^* \big[\bar{F}(T^*)(m_1(\beta_0, \epsilon_0; T^*) + m_2(\beta_0, \epsilon_0; T^*, \Delta)) + \dot{F}_\beta(\beta_0, T^*) m_3(\beta_0, \epsilon_0; X, \Delta)\big] \\
& - S^* \big[\bar{F}(S^*)(m_1(\beta_0, \epsilon_0; S^*) + m_2(\beta_0, \epsilon_0; S^*, \Delta)) + \dot{F}_\beta(\beta_0, S^*) m_3(\beta_0, \epsilon_0; X, \Delta)\big] \\
& - \int_{S^*}^{T^*} \big[\bar{F}(t)(m_1(\beta_0, \epsilon_0; t) + m_2(\beta_0, \epsilon_0; t, \Delta)) + \dot{F}_\beta(\beta_0, t) m_3(\beta_0, \epsilon_0; X, \Delta)\big]\ dt \Big\} \\
& + o_p(1).
\end{aligned}
$$

The functions m_1 and m_2 in the above representation are defined as

$$
m_1(\beta, s; t) = -P\left\{ \frac{\Delta 1(s \geq \epsilon_\beta) 1(t \geq \epsilon_\beta)}{h^{(1)}(\beta, \epsilon_\beta)^2} \right\}, \tag{2.16}
$$

$$
m_2(\beta, s; t, \Delta) = \frac{\Delta 1(t \geq s)}{h^{(1)}(\beta, s)}. \tag{2.17}
$$

The function $m_3(\beta_0, \epsilon_0; X, \Delta)$ is the influence function such that

$$
n^{1/2}(\hat{\beta}_n - \beta_0) = \mathbb{G}_n\{m_3(\beta_0, \epsilon_0; X, \Delta)\} + o_p(1),
$$

where the explicit form of $m_3(\beta_0, \epsilon_0; X, \Delta)$ is given in (2.21). $\dot{F}_\beta(\beta, t)$ is the derivative of $F(\beta, t)$ with respect to β.

This theorem does not provide the asymptotic normality result for the original intercept estimator. However, since S^* and T^* can be arbitrarily large, in practice one is often able to find two values to bound the observed residual time.

2.3 Slope Estimation with Unbounded Covariates

It is easily seen from Theorem 2.2.2 that we can obtain the consistency of $\hat{\alpha}_{n,\hat{\beta}_n}$ in probability and almost surely by providing a consistent estimator $\hat{\beta}_n$ with certain polynomial convergence rate in probability and almost surely, respectively. We consider both in this section using the estimator obtained by the Gehan-weighted rank based estimating method.

Define

$$(2.18) \qquad H_n^{(2)}(\beta, s) = \mathbb{P}_n\{1(\epsilon_\beta \geq s)X\} \quad \text{and} \quad h^{(2)}(\beta, s) = P\{1(\epsilon_\beta \geq s)X\},$$

Then the general rank-based estimating function of Tsiatis (1990) is given by

$$(2.19) \qquad \mathbb{P}_n\left\{\omega_n(\beta, \epsilon_\beta)\left[X - \frac{H_n^{(2)}(\beta, \epsilon_\beta)}{H_n^{(1)}(\beta, \epsilon_\beta)}\right]\Delta\right\},$$

where $\omega_n(\beta, s)$ is a weight function and $H_n^{(1)}(\beta, s) = \mathbb{P}_n\{1(\epsilon_\beta \geq s)\}$ is defined in (2.3). We consider the Gehan weight function $\omega_n(\beta, s) = H_n^{(1)}(\beta, s)$, which yields the following estimating function

$$(2.20) \qquad \Psi_n(\beta, H_n^{(1)}, H_n^{(2)}) = \mathbb{P}_n\big\{[H_n^{(1)}(\beta, \epsilon_\beta)X - H_n^{(2)}(\beta, \epsilon_\beta)]\Delta\big\}.$$

2.3.1 Convergence in Probability and Asymptotic Normality

The only reason of assuming bounded covariates and/or truncated residual time in the current literature is to bound the denominator $H_n^{(1)}(\beta, \epsilon_\beta)$ in (2.19) away from zero. Such an issue disappears in (2.20) and hence none of such assumptions is needed when the Gehan weight function is used. Fygenson and Ritov (1994) showed that the estimating function $\Psi_n(\beta, H_n^{(1)}, H_n^{(2)})$ in (2.20) is monotone in β.

Proposition 2.2.4. Suppose Conditions 1-3 hold. Assume $\beta_0 \in \mathcal{B}$ is the unique root of $\Psi(\beta, h^{(1)}, h^{(2)}) = P\big\{[h^{(1)}(\beta, \epsilon_\beta)X - h^{(2)}(\beta, \epsilon_\beta)]\Delta\big\}$. Then,

(1) The approximate root $\hat{\beta}_n$ satisfying $\Psi_n(\hat{\beta}_n, H_n^{(1)}(\hat{\beta}_n, \cdot), H_n^{(2)}(\hat{\beta}_n, \cdot)) = o_p(1)$ is a consistent estimator of β_0.

(2) Suppose $\Psi(\beta, h^{(1)}, h^{(2)})$ is differentiable with bounded continuous derivative $\dot{\Psi}_\beta(\beta, h^{(1)}(\beta, \cdot), h^{(2)}(\beta, \cdot))$ in a neighborhood of β_0, and $\dot{\Psi}_\beta(\beta_0, h^{(1)}(\beta_0, \cdot), h^{(2)}(\beta_0, \cdot))$ is nonsingular. Then for an approximate root $\hat{\beta}_n$ satisfying

$$\Psi_n(\hat{\beta}_n, H_n^{(1)}(\hat{\beta}_n, \cdot), H_n^{(2)}(\hat{\beta}_n, \cdot)) = o_p(n^{-1/2}),$$

we have $|\hat{\beta}_n - \beta_0| = O_p(n^{-1/2})$.

(3) Suppose the same assumptions given in (2) hold, then $n^{1/2}(\hat{\beta}_n - \beta_0)$ is asymptotically normal with the following asymptotic representation

$$n^{1/2}(\hat{\beta}_n - \beta_0) = \mathbb{G}_n\{m_3(\beta_0, \epsilon_0; \Delta, X)\} + o_p(1),$$

where

$$\begin{aligned} (2.21) \qquad & m_3(\beta_0, \epsilon_0; \Delta, X) \\ = \quad & \left\{-\dot{\Psi}_\beta(\beta_0, h^{(1)}(\beta_0, \cdot), h^{(2)}(\beta_0, \cdot))\right\}^{-1} \Big\{ [h^{(1)}(\beta_0, \cdot)X - h^{(2)}(\beta_0, \cdot)]\Delta \\ & - \int [1(\epsilon_0 \geq t)X] \; dP_{\epsilon_0, \Delta}(t, 1) + \int [1(\epsilon_0 \geq t)]x \; dP_{\epsilon_0, \Delta, X}(t, 1, x) \Big\}. \end{aligned}$$

The estimator $\hat{\beta}_n$ is in fact in a neighborhood of the zero-crossing point due to the discrete nature of $\Psi_n(\beta, H_n^{(1)}, H_n^{(2)})$, see e.g. Kalbfleisch and Prentice (2002). Proposition 2.2.4 implies that $|\hat{\beta}_n - \beta_0| = O_p(n^{-3\varepsilon})$ for any $0 < \varepsilon \leq \frac{1}{8}$ with probability approaching to 1. Hence by Theorem 2.2.2, $\hat{\alpha}_{n,\hat{\beta}_n}$ converges to α_0 in probability.

2.3.2 Almost Sure Convergence with Polynomial Rate

Following by the Theorem 5 in Ying (1993), the almost sure consistency of the slope estimator with a polynomial rate can be also achieved under the unbounded covariate support assumption, which is given in the following proposition.

Proposition 2.2.5. Suppose Conditions 1-4 hold, and in addition assume that the tail probability of the covariate X satisfies

$$P(|X| > t) \le Mt^{\theta}\exp(-\eta t^{\gamma}) \tag{2.22}$$

for some constants $M > 0$, $-\infty < \theta < \infty$, $\eta > 0$, and $\gamma > 0$. Then the estimator $\hat{\beta}_n$ satisfying $\Psi_n(\hat{\beta}_n, H_n^{(1)}(\hat{\beta}_n, \cdot), H_n^{(2)}(\hat{\beta}_n, \cdot)) = o(n^{-1/2})$ almost surely converges with probability 1 to β_0 with a polynomial rate, that is, $|\hat{\beta}_n - \beta_0| = o(n^{-1/2+\varepsilon})$ almost surely for every $\varepsilon > 0$.

Comparing to Proposition 2.2.4, an exponential tail probability bound for the covariate is assumed to guarantee the almost sure consistency of $\hat{\beta}_n$ with a polynomial rate. Such assumption yields the relaxed Condition 1 in Ying (1993), page 83, i.e., $\max_{i\le n}|X_i| = o(n^{\varepsilon})$ almost surely for every $\varepsilon > 0$. This is because when (2.22) holds, for every $t > 0$ we have

$$\begin{aligned} P(\max_{i\le n}|X_i| > t) &= 1 - P(\max_{i\le n}|X_i| \le t) \\ &= 1 - [1 - P(|X| > t)]^n \le 1 - [1 - Mt^{\theta}\exp(-\eta t^{\gamma})]^n \\ &\le 1 - [1 - nMt^{\theta}\exp(-\eta t^{\gamma})] = nMt^{\theta}\exp(-\eta t^{\gamma}), \end{aligned}$$

where the last inequality holds due to the fact that $(1-s)^n \ge 1 - ns$ for $0 \le s \le 1$. Therefore, for every fixed $t > 0$ and $\varepsilon > 0$,

$$\begin{aligned} \sum_{n=1}^{\infty} P(n^{-\varepsilon}\max_{i\le n}|X_i| > t) &= \sum_{n=1}^{\infty} P(\max_{i\le n}|X_i| > n^{\varepsilon}t) \\ &\le \sum_{n=1}^{\infty} nM(n^{\varepsilon}t)^{\theta}\exp\{-\eta(n^{\varepsilon}t)^{\gamma}\} < \infty. \end{aligned}$$

It then follows by the Borel-Cantelli lemma that $P(\lim_{n\to\infty} n^{-\varepsilon}\max_{i\le n}|X_i| = 0) = 1$, i.e., $\max_{i\le n}|X_i| = o(n^{\varepsilon})$ almost surely. As mentioned in Section 2.2.2, Conditions 2-4 imply Conditions 2-4 in Ying (1993), respectively. Furthermore, Ying (1993) pointed out that the Gehan weights satisfy their Condition 5 and (4.7). Hence the conclusion in Proposition 2.2.5 follows directly from their (4.8) in Theorem 5 of Ying (1993). The detailed argument is thus omitted. The exponential tail probability condition holds for many commonly used distributions such as normal, weibull, extreme value distributions and etc.

2.4 Simulations

We conduct extensive simulations to investigate the finite sample performance of the intercept estimator and the slope estimator under different scenarios. Failure times are generated from the following model

$$T = 2 + X + \zeta,$$

where five different error distributions are considered, which are (a) $\zeta \sim N(0, 0.5^2)$; (b) $\zeta \sim \mathit{Gumbel}(-0.5\gamma, 0.5)$ that has mean zero, where γ is the Euler constant; (c) $\zeta \sim \mathit{Laplace}(0, 0.5)$; (d) $\zeta \sim \mathit{Logistic}(0, 0.5)$; and (e) $\zeta \sim t(0, \mathit{df} = 30)$. In each scenario, three different settings of covariate X are investigated, which are (1) $X \sim N(0, 1)$; (2) $X \sim U(-2, 2)$; and (3) $X \sim U(-1, 1)$. The censoring distribution is $C \sim U(0, 4) \wedge \tau$, here τ is the follow-up time. We choose $\tau = 1$ and $\tau = 3$ to yield censoring rate ranges $(75\%, 92\%)$ and $(51\%, 53\%)$, respectively. For each setting, we simulate 1000 runs with four different sample sizes: 50, 200, 500 and 2000. Since $\tau = 1$ yields very high censoring rate that causes numerical instability, we drop $\tau = 1$ for sample size 50. The simulation results are summarized in Table 2.1.

The first covariate setting corresponds to the unbounded covariate support. It is clearly seen that the bias of the intercept estimator is minimal even with a short follow-up time $\tau = 1$ for all error distributions. The bias is also very small in the second covariate setting, where the support of X is bounded, but wide. The bias becomes noticeable when the support of X gets narrow in the third setting with the short follow-up time $\tau = 1$. With the longer follow-up time $\tau = 3$, the bias of the intercept estimator is negligible even for a small sample size ($n = 50$) under all error distributions. The bias for the slope estimator is minimal across all simulation settings.

For the short follow-up setting ($\tau = 1$), Figure 2.1 displays the Kaplan-Meier curves of the estimated residual survival time $T_i - \hat{\beta}_n X_i$ under five error distributions with sample size $n = 2000$. The left panel corresponds to the unbounded covariate scenario with $X \sim N(0,1)$, the middle panel corresponds to the scenario with $X \sim U(-2,2)$, and the right panel corresponds to the scenario with $X \sim U(-1,1)$. We notice that none of the Kaplan-Meier curves in the right panel goes to zero in the right tail when the support of X is narrow. This implies that the condition $\{t : t \text{ in the support of } T - \hat{\beta}_n X\} \subseteq \{t : t \text{ in the support of } C - \hat{\beta}_n X\}$ is violated in this situation, and hence the intercept estimators are biased. With unbounded X, all five Kaplan-Meier curves go to zero in the right tail, which indicates that a good estimator of the intercept can be obtained. The similar pattern is also observed for the case with bounded but wide support X, i.e., $X \sim U(-2,2)$. This type of plot is suggested to examine the quality of the intercept estimation in the real data analysis. Based on additional simulations (results are not shown here), we suggest that if the right tail of the Kaplan-Meier curve goes below 0.1-0.15, a good estimate of the intercept can be obtained.

Moreover, for $\tau = 3$, we plot the empirical variances of $\hat{\alpha}_{n,\hat{\beta}_n}$ and $\hat{\beta}_n$ versus the reciprocal of the sample sizes respectively, and display the graphs in Figure 2.2. It is clear that there is a linear relationship between the empirical variances of each estimator and the reciprocal of the sample size. This provides numerical evidence of the root-n convergence rate for both $\hat{\alpha}_{n,\hat{\beta}_n}$ and $\hat{\beta}_n$.

We also study the survival time prediction accuracy of the semiparametric linear model via simulations and compare it to the Cox model. In order to have a fair comparison, we generate data from the following model:

$$h(T) = X + e_0, \tag{2.23}$$

where e_0 follows the standard extreme value distribution with $F(t) = 1 - e^{-e^t}$, and $h(\cdot)$ is some monotone transformation. In such a setting, we have $\beta_0 = 1$ and $\alpha_0 = Ee_0 = -\gamma$, where γ is the Euler constant. Note that this is different to the previous simulation setting with $\zeta \sim Gumbel(-0.5\gamma, 0.5)$ where the mean is shifted to zero and the scale parameter is 0.5. It is well known that both the semiparametric linear regression model and the Cox model correctly fit the data generated from (2.23). We choose the censoring distribution as $h(C) \sim U(-3, 3) \wedge \tau$, where τ is a fixed follow-up time taking different values to generate different censoring rates. As in the first simulation study, covariate X is generated from three distributions, namely, $N(0, 1)$, $U(-2, 2)$ and $U(-1, 1)$. Two transformations are considered: the identity (with a constant added to shift all the survival times to positive values) and the logarithm transformations. For each simulation setting, two independent data sets of equal size are generated from the same model, namely the training set and the test set. Both the semiparametric linear model and the Cox model are fitted using the training set, and survival times are predicted for the test set using the fitted models. For the linear regression model, the predicted survival time is calculated by $\hat{T}_i^{LR} = h^{-1}(\hat{\alpha}_n + \hat{\beta}_n^{LR} X_i^*)$, where X_i^* is the observed covariate for the ith subject in the test set, $\hat{\beta}_n^{LR}$ is solved by the Gehan-weighted rank based estimating equation, and $\hat{\alpha}_n$ is estimated from (2.5). For the Cox model, the predicted survival time is calculated by $\hat{T}_i^{Cox} = \int t \; d\big(1 - \exp\{-\hat{\Lambda}_{0,n}(t) e^{\hat{\beta}_n^{Cox} X_i^*}\}\big)$, where $\hat{\Lambda}_{0,n}(t)$ is the Breslow estimator of the baseline cumulative hazard function $\Lambda_0(t)$ and $\hat{\beta}_n^{Cox}$ is the partial likelihood estimator. We use the following measure to determine the prediction accuracy:

$$MSE_p = \frac{1}{n}\sum_{i=1}^{n}(T_i^* - \hat{T}_i)^2, \tag{2.24}$$

where $\hat{T}_i$ is either $\hat{T}_i^{LR}$ or $\hat{T}_i^{Cox}$ depending on which model is used and T_i^* is the true survival time for the ith subject in the test set. Two sample sizes are considered: $n = 200$ and $n = 2000$, and 1000 runs are conducted for each simulation setting. The results are summarized in Table 2.2. For each scenario, in addition to the empirical mean and standard deviation of MSE_p, given in the parenthesis, we also calculate the relative prediction accuracy to the uncensored case, that is, the ratio of the empirical mean MSE_p under uncensored case to that under each corresponding censored case. The MSE_p obtained from ordinary least squares (OLS) is also listed for each uncensored scenario.

Under both transformations and all three covariate distributions, the semiparametric linear model yields larger relative prediction accuracy than the Cox model for all censored cases with four different censoring rates. Even with short follow-up times, the relative prediction accuracy is close to 1 under the semiparametric linear model, especially with large sample size, but much smaller than 1 under the Cox model for both sample sizes. This is not surprising because for the Cox model, the baseline hazard function after the last observation time in the training set is not estimable. For a study with the follow-up time τ, without any parametric assumption for the baseline hazard function, the convention is to set $\hat{\Lambda}_{0,n}(t) = \infty$ for $t > \tau$. This will introduce bias when predicting the survival time for new observations and obviously, the bias becomes more severe when follow-up time is shorter.

Moreover, under the identity transformation, the absolute MSE_p value from the semiparametric linear model is smaller compared to that from the Cox model, especially when the follow-up time is short. In the scenario when X has an unbounded or a wide support, i.e., $X \sim N(0,1)$ or $X \sim U(-2,2)$, the linear model yields larger relative prediction accuracy and smaller absolute MSE_p value comparing with the scenario when $X \sim U(-1,1)$. One possible explanation for this finding is that the intercept estimate is more accurate under the situation when the support of X is unbounded or wide. It also can be seen that when there is no censoring, both models yield the same MSE_p value as that obtained by the ordinary least squares. Under the logarithm transformation, the absolute MSE_p value from the linear model is also smaller than that from the Cox model when censoring exists, except for the longest follow-up case under $X \sim U(-1,1)$. While there is no censoring, the Cox model yields smaller MSE_p compared to that obtained by both the semiparametric linear model and the ordinary least squares. This is because the linear model predicts the survival time as $e^{E(\log T)}$, which underestimates ET and hence causes bias.

In Figure 2.3, we plot the predicted survival time versus the true survival time for both semiparametric linear model and Cox model under the identity transformation with $X \sim N(0,1)$ for the follow-up time $\tau = -2$ and 0 ($n = 2000$). It is clearly seen that the semiparametric linear model provides a nice prediction of the survival time even with a short follow-up time. However the prediction from the Cox model under two scenarios are both poor and we observe more severe bias when follow-up time is shorter.

2.5 A Real Data Example

We consider the well-known Mayo primary biliary cirrhosis (PBC) study as an illustrative example (Fleming and Harrington 1991, app. D.1). The data contain information about the survival time and prognostic factors for 418 patients. Jin et al. (2003) and Jin et al. (2006) fitted the accelerated failure time model with five covariates, namely age, log(albumin), log(bilirubin), edema, and log(protime). They used the rank-based and least squares estimators and reported only the slope estimators for the five covariates. We fit the same model with the slope estimators obtained by the rank based estimating equation with Gehan weights and the intercept estimator obtained by (2.5). The estimated coefficients for the five prognostic factors are -0.025, 1.498, -0.554, -0.904, and -2.822 with estimated standard errors of 0.005, 0.479, 0.052, 0.234 and 0.923. Our estimates are similar to those reported in Jin et al. (2003). The intercept estimator is 8.692. The Kaplan-Meier curve of the residual survival time under the estimated slope parameters goes to zero in the right tail (shown in Figure 2.4a), which indicates no evidence that the intercept estimator obtained by (2.5) is biased.

We then perform the leave-one-out cross-validation to check the prediction performance of the model. If the fitted model is adequate, the predicted survival time is expected to be close to the observed time for those patients who failed. On the other hand, for patients who were censored, the predicted survival time is expected to be greater than the observed time. Figure 2.4b shows the predicted survival time against the observed time in the logarithm scale. The circles correspond to the patients who failed and the triangles correspond to the patients who were censored. The figure suggests that the accelerated failure time model provides reasonably good prediction of the survival time for this dataset, except for a few subjects who might be outliers. For example, subject 87 (circled in Figure 2.4b) was a 37 year old woman with quite good prognostic status: no edema, good albumin (4.4), low bilirubin (1.1) and moderate protime (10.7). Yet she survived for no longer than roughly half a year. Subject 293, on the other hand, was a 57 year old woman with poor prognostic status. In spite of low albumin (2.98), high bilirubin (8.5) and protime (12.3), and edema resistent to diuretics, she remains alive after more than 3.5 years. This same subject was also detected as an outlier in the residual plot for the covariate edema from a Cox model for the same data (Fleming and Harrington 1991, p. 184).

2.6 Concluding Remarks

Other important features such as the asymptotic distribution of the original intercept estimator (2.5) and the procedure to estimate its variance remain unknown and are worth further investigation.

In a practical situation with a finite follow-up time τ, instead of estimating the unconditional expectation of the squared error loss $[T - E(T|X)]^2$ to assess the model prediction performance, it seems more reasonable to consider the conditional expectation by conditioning on that the survival time is no greater than τ, i.e.,

$$E\{[T - E(T|X)]^2 \mid T \leq \tau\}. \tag{2.25}$$

This is because the model is believed to be correct only for the individuals whose survival time is no greater than τ. Then as an analogue to the MSE_p in (2.24), a natural finite sample estimation of (2.25) is

$$MSE_{p,\tau} = n_\tau^{-1} \sum_{i=1}^{n} \big[(T_i^* - \hat{T}_i)^2 \cdot 1(T_i^* \leq \tau)\big],$$

where $n_\tau = \sum_{i=1}^{n} 1(T_i^* \leq \tau)$, $\hat{T}_i$ is the predicted survival time and T_i^* is the true survival time for individual i. Simulations (results are not shown here) reveal that when the follow-up time τ is short, the Cox model cannot provide a reasonable $MSE_{p,\tau}$ value, while the linear model estimates the $MSE_{p,\tau}$ reasonably well.

In the real data analysis, model checking is a very important issue. One possible procedure is to follow the method developed for the Cox model by plotting the cumulative sums of the martingale-based residuals to assess how unusual the observed residual patterns would be, see e.g. Lin et al. (1993) and Lin et al. (1996).

Bias and variance trade-off plays an important role in assessing prediction errors, which requires knowing the asymptotic joint distribution of both the intercept and slope parameter estimators. We do not consider it here. We also want to point out that any prediction beyond the last observed failure time needs to be interpreted cautiously because it lacks empirical verification without obtaining new data with longer observed survival times.

2.7 Appendix: Proofs of the Technical Results

In this section, we provide the proofs of Theorems 2.2.1-2.2.3 and Proposition 2.3.4. We first provide several lemmas that will be used for the proofs.

2.7.1 Technical Lemmas

Lemma A.1. For every $\varepsilon > 0$, with probability one we have

$$\sup_{\beta\in\mathcal{B},-\infty<s<\infty} n^{1/2}|H_n^{(k)}(\beta,s) - h^{(k)}(\beta,s)| = o(n^{\varepsilon}),$$

where $H_n^{(k)}(\beta,s)$ and $h^{(k)}(\beta,s)$, $k = 0,1$, are defined in (2.2) and (2.3) respectively.

Proof: We shall use the empirical process theory to prove this result. Since the class of indicator functions of half spaces is a VC-class, see e.g. Exercise 9 on page 151 and Exercise 14 on page 152 in van der Vaart and Wellner (1996), and thus a Donsker class. Then the sets of functions $\mathcal{F}_0 = \{1(\epsilon_\beta \leq s, \Delta = 1)\}$=$\{\Delta 1(\epsilon_\beta \leq s)\}$ and $\mathcal{F}_1 = \{1(\epsilon_\beta \geq s)\}$ are both Donsker classes. Let $\bar{\mathcal{F}}_k$ be the closure of $\mathcal{F}_k$, $k = 0,1$, respectively. Then $H_n^{(k)}(\beta,s)$ and $h^{(k)}(\beta,s)$ are in the convex hull of $\bar{\mathcal{F}}_k$, $k = 0,1$, and thus belong to Donsker classes. See e.g. Theorems 2.10.2 and 2.10.3 in van der Vaart and Wellner (1996). Hence by their Theorem 2.6.7 and Theorem 2.14.9, it follows that for every $t > 0$,

$$P\biggl(\sup_{\beta\in\mathcal{B},-\infty<s<\infty} n^{1/2}|H_n^{(k)}(\beta,s) - h^{(k)}(\beta,s)| > t\biggr) \le Mt^V e^{-2t^2},$$

where $M > 0$ is a constant and $V = 2V(\mathcal{F}) - 2$ with $V(\mathcal{F})$ being the index of the VC-class $\mathcal{F}$, which is 4 in this case for one-dimensional β_0, hence $V = 6$. When $\beta_0 \in \mathbb{R}^d$ for a fixed d, the index of the VC-class is $V(\mathcal{F}) = d + 3$ and the following argument still holds. Then for any $\varepsilon > 0$, let

$$A_{n,\varepsilon} = \sup_{\beta\in\mathcal{B},-\infty<s<\infty} n^{1/2-\varepsilon}|H_n^{(k)}(\beta,s) - h^{(k)}(\beta,s)|.$$

Since $t^V \le e^{1.5t^2}$ for large enough $t > 0$ and a fixed $V > 0$, then

$$\sum_{n=1}^{\infty} P(|A_{n,\varepsilon} - 0| > t) \le M\sum_{n=1}^{\infty}\exp\{-0.5(n^{\varepsilon}t)^2\} < \infty.$$

It then follows by the Borel-Cantelli lemma that $P\bigl(\lim_{n\to\infty} A_{n,\varepsilon} = 0\bigr) = 1$. Hence we obtain the desired result.

Lemma A.2. Assume Conditions 1-3 hold, then for every $\varepsilon \ge 0$ we have

$$\sup_{|\beta-\beta'|+|s-s'|\le n^{-\varepsilon}} |h^{(k)}(\beta,s) - h^{(k)}(\beta',s')| = O(n^{-\varepsilon}),$$

where $h^{(k)}(\beta,s)$, $k = 0, 1$ and 2, are defined in (2.2), (2.3) and (2.18) respectively.

Proof: Since $e_0 = T - \beta_0 X$ is independent of (X, C), the joint density function of (T, C, X) can then be decomposed as

$$f_{T,C,X}(t,c,x) = f_{e_0,C,X}(t - \beta_0 x, c, x) = f(t - \beta_0 x) f_{C,X}(c,x)$$

where f is the density of e_0. So

$$f(t - \beta_0 x) = f_{T|C,X}(t|C = c, X = x) = f_{T|X}(t|X = x).$$

Then the joint density function of (Y, Δ, X) follows

$$f_{Y,\Delta,X}(y,\delta,x) = f(y-\beta_0 x)^{\delta}\bar{F}(y-\beta_0 x)^{1-\delta} g_{C|X}(y|X=x)^{1-\delta}\bar{G}_{C|X}(y|X=x)^{\delta} f_X(x),$$

where $\bar{F}(\cdot) = 1 - F(\cdot)$ and $\bar{G}_{C|X}(\cdot|X=x) = 1 - G_{C|X}(\cdot|X=x)$.

For $h^{(0)}(\beta,s)$, the joint sub-density function of $(Y, \Delta = 1, X)$ can be written as $f_{Y,\Delta,X}(y,1,x) = f(y-\beta_0 x)\bar{G}_{C|X}(y|X=x)f_X(x)$. So

$$\begin{aligned} h^{(0)}(\beta,s) &= P\{1(\epsilon_\beta \le s, \Delta = 1)\} \\ &= \int_{\mathcal{X}}\left\{\int_{-\infty}^{s} f(u+(\beta-\beta_0)x)\bar{G}_{C|X}(u+\beta x|X=x)\,du\right\} f_X(x)\,dx. \end{aligned}$$

Then for any $\beta, \beta' \in \mathcal{X}$ and $-\infty < s < \infty$, by the mean value theorem, there exists a value $\tilde{\beta}$ between β and β' such that

$$\begin{aligned} & |h^{(0)}(\beta,s) - h^{(0)}(\beta',s)| \\ = & \left|\int_{\mathcal{X}}\left\{\int_{-\infty}^{s} [f(u+(\beta-\beta_0)x)\bar{G}_{C|X}(u+\beta x|X=x) \right.\right. \\ & \left.\left. - f(u+(\beta'-\beta_0)x)\bar{G}_{C|X}(u+\beta' x|X=x)]\,du\right\} f_X(x)\,dx\right| \\ = & \left|\int_{\mathcal{X}}\left\{\int_{-\infty}^{s} \big[\dot{f}(u+(\tilde{\beta}-\beta_0)x)\bar{G}_{C|X}(u+\tilde{\beta}x|X=x) \right.\right. \\ & \left.\left. - f(u+(\tilde{\beta}-\beta_0)x)g_{C|X}(u+\tilde{\beta}x|X=x)\big](\beta-\beta')x\,du\right\} f_X(x)\,dx\right| \\ \le & |\beta-\beta'|\int_{\mathcal{X}}\left\{\int_{-\infty}^{s} \big|\dot{f}(u+(\tilde{\beta}-\beta_0)x)\bar{G}_{C|X}(u+\tilde{\beta}x|X=x) \right. \\ & \left. - f(u+(\tilde{\beta}-\beta_0)x)g_{C|X}(u+\tilde{\beta}x|X=x)\big|\,du\right\}|x| f_X(x)\,dx \\ \le & C_1|\beta-\beta'|\int_{\mathcal{X}}\left\{\int_{-\infty}^{s} |\dot{f}(u+(\tilde{\beta}-\beta_0)x)| + f(u+(\tilde{\beta}-\beta_0)x)\,du\right\}|x| f_X(x)\,dx \\ \le & C_1|\beta-\beta'|\int_{\mathcal{X}}\left\{\int_{-\infty}^{\infty} |\dot{f}(u)| + f(u)\,du\right\}|x| f_X(x)\,dx \\ \le & C_1 C_2|\beta-\beta'|\int_{\mathcal{X}} |x| f_X(x)\,dx, \end{aligned}$$

where the second inequality holds for some $C_1 \ge 1$ such that $g_{C|X}(\cdot|X=x) \le C_1$ uniformly, which is guaranteed by Condition 3; the third inequality holds since both

$|\dot{f}(u)|$ and $f(u)$ are nonnegative and thus for any s, $\tilde{\beta}$ and x, the following inequality holds:

$$\begin{aligned} & \int_{-\infty}^{s} |\dot{f}(u+(\tilde{\beta}-\beta_0)x)| + f(u+(\tilde{\beta}-\beta_0)x)\ du \\ = & \int_{-\infty}^{s+(\tilde{\beta}-\beta_0)x} |\dot{f}(u)| + f(u)\ du \le \int_{-\infty}^{\infty} |\dot{f}(u)| + f(u)\ du; \end{aligned}$$

and the last inequality holds since under Condition 2, it follows by the Cauchy-Schwartz inequality that

$$\begin{aligned} \left\{\int_{-\infty}^{\infty} |\dot{f}(u)|\ du\right\}^2 &\le \int_{-\infty}^{\infty} \left(\frac{|\dot{f}(u)|}{\sqrt{f(u)}}\right)^2 du \cdot \int_{-\infty}^{\infty} \left(\sqrt{f(u)}\right)^2 du \\ &= \left\{\int_{-\infty}^{\infty} \left(\frac{\dot{f}(u)}{f(u)}\right)^2 f(u)\ du\right\} \cdot 1 < \infty, \end{aligned}$$

so we can find a constant $C_2 > 1$ such that

$$\int_{-\infty}^{\infty} |\dot{f}(u)| + f(u)\ du = \int_{-\infty}^{\infty} |\dot{f}(u)|\ du + 1 \le C_2.$$

Therefore, by Condition 1 that X has a finite second moment and thus a finite first moment, it follows that

$$|h^{(0)}(\beta, s) - h^{(0)}(\beta', s)| = O(|\beta - \beta'|)$$

for all $\beta, \beta' \in \mathcal{B}$ and $-\infty < s < \infty$. Moreover, for any $\beta \in \mathcal{B}$ and $-\infty < s, s' < \infty$, we have

$$\begin{aligned} & |h^{(0)}(\beta, s) - h^{(0)}(\beta, s')| \\ = & \left|\int_{\mathcal{X}} \left\{\int_{s}^{s'} f(u+(\beta-\beta_0)x)\bar{G}_{C|X}(u+\beta x|X=x)\ du\right\} f_X(x)\ dx\right| \\ \le & \int_{\mathcal{X}} \left|\int_{s}^{s'} f(u+(\beta-\beta_0)x)\ du\right| f_X(x)\ dx \\ \le & C_3|s-s'| \int_{\mathcal{X}} f_X(x)\ dx = O(|s-s'|), \end{aligned}$$

where $C_3 > 0$ is a constant such that $f(\cdot) \leq C_3$, which is guaranteed by Condition 2. Hence, for any $\beta, \beta' \in \mathcal{X}$ and $-\infty < s, s' < \infty$, it follows that

$$\begin{aligned} & |h^{(0)}(\beta, s) - h^{(0)}(\beta', s')| \\ \leq \; & |h^{(0)}(\beta, s) - h^{(0)}(\beta', s)| + |h^{(0)}(\beta', s) - h^{(0)}(\beta', s')| \\ = \; & O(|\beta - \beta'|) + O(|s - s'|), \end{aligned}$$

and therefore for any $\varepsilon > 0$, we have

$$\sup_{|\beta-\beta'|+|s-s'|\leq n^{-\varepsilon}} |h^{(0)}(\beta, s) - h^{(0)}(\beta', s')| = O(n^{-\varepsilon}).$$

For $h^{(1)}(\beta, s)$, it is easy to obtain that

$$P\{1(\epsilon_\beta \geq s)|X = x\} = \bar{F}(s + (\beta - \beta_0)x)\bar{G}_{C|X}(s + \beta x|X = x). \tag{2.26}$$

Then for any $\beta, \beta' \in \mathcal{X}$ and $-\infty < s < \infty$, by the mean value theorem, there exists a value $\tilde{\beta}$ between β and β' such that

$$\begin{aligned} & |h^{(1)}(\beta, s) - h^{(1)}(\beta', s)| = |P\{1(\epsilon_\beta \geq s)\} - P\{1(\epsilon_{\beta'} \geq s)\}| \\ & = \Bigg| \int_{\mathcal{X}} \big\{ \bar{F}(s + (\beta - \beta_0)x)\bar{G}_{C|X}(s + \beta x|X = x) \\ & \quad - \bar{F}(s + (\beta' - \beta_0)x)\bar{G}_{C|X}(s + \beta' x|X = x) \big\} f_X(x)\, dx \Bigg| \\ & = \Bigg| \int_{\mathcal{X}} \big\{ -f(s + (\tilde{\beta} - \beta_0)x)\bar{G}_{C|X}(s + \tilde{\beta} x|X = x) \\ & \quad - \bar{F}(s + (\tilde{\beta} - \beta_0)x) g_{C|X}(s + \tilde{\beta} x|X = x) \big\} (\beta - \beta') x f_X(x)\, dx \Bigg| \\ & \leq |\beta - \beta'| \int_{\mathcal{X}} \big\{ f(s + (\tilde{\beta} - \beta_0)x) + g_{C|X}(s + \tilde{\beta} x|X = x) \big\} |x| f_X(x)\, dx \\ & \leq (C_1 + C_3)|\beta - \beta'| \int_{\mathcal{X}} |x| f_X(x)\, dx = O(|\beta - \beta'|), \end{aligned}$$

where C_1 and C_3 are two constants such that $g_{C|X}(\cdot|X = x) \leq C_1$ and $f(\cdot) \leq C_3$ (defined before), and the last equality holds because $E|X| < \infty$. Moreover, for any

$\beta \in \mathcal{B}$ and $-\infty < s, s' < \infty$, by the mean value theorem, there exists a value $\tilde{s}$ between s and s' such that

$$\begin{aligned}
&|h^{(1)}(\beta, s) - h^{(1)}(\beta, s')| = |P\{1(\epsilon_\beta \geq s)\} - P\{1(\epsilon_\beta \geq s')\}| \\
&= \bigg| \int_{\mathcal{X}} \big\{ -f(\tilde{s} + (\beta - \beta_0)x) \bar{G}_{C|X}(\tilde{s} + \beta x | X = x) \\
&\quad - \bar{F}(\tilde{s} + (\beta - \beta_0)x) g_{C|X}(\tilde{s} + \beta x | X = x) \big\} (s - s') f_X(x)\, dx \bigg| \\
&\leq |s - s'| \int_{\mathcal{X}} \big\{ f(\tilde{s} + (\beta - \beta_0)x) + g_{C|X}(\tilde{s} + \beta x | X = x) \big\} f_X(x)\, dx \\
&\leq (C_1 + C_3)|s - s'| \int_{\mathcal{X}} f_X(x)\, dx = O(|s - s'|).
\end{aligned}$$

Hence, for any $\varepsilon > 0$, we have

$$\sup_{|\beta - \beta'| + |s - s'| \leq n^{-\varepsilon}} |h^{(1)}(\beta, s) - h^{(1)}(\beta', s')| = O(n^{-\varepsilon}).$$

Finally for $h^{(2)}(\beta, s)$, by using the similar argument for $h^{(1)}(\beta, s)$, we can easily obtain that

$$\begin{aligned}
&|h^{(2)}(\beta, s) - h^{(2)}(\beta', s)| = |P\{1(\epsilon_\beta \geq s)X\} - P\{1(\epsilon_{\beta'} \geq s)X\}| \\
&\quad \leq (C_1 + C_3)|\beta - \beta'| \int_{\mathcal{X}} x^2 f_X(x)\, dx = O(|\beta - \beta'|),
\end{aligned}$$

and

$$\begin{aligned}
&|h^{(2)}(\beta, s) - h^{(2)}(\beta, s')| = |P\{1(\epsilon_\beta \geq s)X\} - P\{1(\epsilon_\beta \geq s')X\}| \\
&\quad \leq (C_1 + C_3)|s - s'| \int_{\mathcal{X}} |x| f_X(x)\, dx = O(|s - s'|).
\end{aligned}$$

Therefore, for any $\varepsilon > 0$, we have

$$\sup_{|\beta - \beta'| + |s - s'| \leq n^{-\varepsilon}} |h^{(2)}(\beta, s) - h^{(2)}(\beta', s')| = O(n^{-\varepsilon}).$$

Thus, we have proved Lemma A.2.

Lemma A.3. Let $U_n(\beta, s)$ be random variables for which there exist non-random Borel functions $u_n(\beta, s)$ such that for every $\varepsilon > 0$, (A1) $\sup_{\beta\in\mathcal{B},-\infty<s<\infty} |U_n(\beta,s) - u_n(\beta,s)| = o(n^{-1/2+\varepsilon})$ almost surely.

(A2) $U_n(\beta,s)$ has a bounded variation in s uniformly on $\mathcal{B}$, that is,

$$\sup_{\beta\in\mathcal{B}} \int_{s=-\infty}^{\infty} |dU_n(\beta,s)| = O(1) \text{ almost surely.}$$

(A3) u_n satisfies

$$\sup_{\beta\in\mathcal{B},-\infty<s<\infty} |u_n(\beta,s)| = O(1).$$

Then under Conditions 1-3, for every $0 < \varepsilon \le 1/2$, with probability 1 we have

$$\sup_{\beta\in\mathcal{B},-\infty<y<\infty} \left| \int_{s=-\infty}^{y} U_n(\beta,s)\ dH_n^{(0)}(\beta,s) - \int_{s=-\infty}^{y} u_n(\beta,s)\ dh^{(0)}(\beta,s) \right| = o(n^{-1/2+\varepsilon}).$$

Proof: By the triangle inequality and integration by parts, we have

$$\begin{aligned} &\left| \int_{s=-\infty}^{y} U_n(\beta,s)\ dH_n^{(0)}(\beta,s) - \int_{s=-\infty}^{y} u_n(\beta,s)\ dh^{(0)}(\beta,s) \right| \\ &\quad \le \int_{s=-\infty}^{y} |U_n(\beta,s) - u_n(\beta,s)|\ dh^{(0)}(\beta,s) + \big|U_n(\beta,y)\big(H_n^{(0)}(\beta,y) - h^{(0)}(\beta,y)\big)\big| \\ &\quad\quad + \int_{s=-\infty}^{y} |H_n^{(0)}(\beta,s) - h^{(0)}(\beta,s)|\ |dU_n(\beta,s)|. \end{aligned}$$

Then it is easy to see that each term on the right hand side of the above inequality is $o(n^{-1/2+\varepsilon})$ almost surely under (A1)-(A3) and Lemma A.1.

2.7.2 Proof of Theorem 2.2.1

By the first order Taylor expansion of function $\log(1-x)$, for large n we have

$$\begin{aligned} \hat{F}_{n,\beta}(t) &= 1 - \exp\left\{ \sum_{i:\epsilon_{\beta,i}\le t} \log\left(1 - \frac{\Delta_i/n}{H_n^{(1)}(\beta,\epsilon_{\beta,i})}\right) \right\} \\ &= 1 - \exp\left\{ -\int_{u\le t} \frac{dH_n^{(0)}(\beta,u)}{H_n^{(1)}(\beta,u)} - \sum_{i:\epsilon_{\beta,i}\le t} O\big(\{nH_n^{(1)}(\beta,\epsilon_{\beta,i})\}^{-2}\big) \right\}. \end{aligned}$$

Then by the mean value theorem and the fact that $e^x \le 1$ for any $x \le 0$, it follows that

$$\begin{aligned}
& |\hat{F}_{n,\beta}(t) - F(\beta,t)| \\
= & \bigg|\exp\Big\{-\int_{-\infty}^{t} \frac{dh^{(0)}(\beta,u)}{h^{(1)}(\beta,u)}\Big\} \\
& - \exp\Big\{-\int_{-\infty}^{t} \frac{dH_n^{(0)}(\beta,u)}{H_n^{(1)}(\beta,u)} - n^{-2}\sum_{i:\epsilon_{\beta,i}\le t} O\big(H_n^{(1)}(\beta,\epsilon_{\beta,i})^{-2}\big)\Big\}\bigg| \\
\le & \bigg|\int_{-\infty}^{t} \frac{dH_n^{(0)}(\beta,u)}{H_n^{(1)}(\beta,u)} - \int_{-\infty}^{t} \frac{dh^{(0)}(\beta,u)}{h^{(1)}(\beta,u)} + n^{-2}\sum_{i:\epsilon_{\beta,i}\le t} O\big(H_n^{(1)}(\beta,\epsilon_{\beta,i})^{-2}\big)\bigg|.
\end{aligned}$$

Under the condition $H_n^{(1)}(\beta,t) \ge n^{-\varepsilon}$, we have

$$n^{-2}\sum_{i:\epsilon_{\beta,i}\le t} O\big(H_n^{(1)}(\beta,\epsilon_{\beta,i})^{-2}\big) \le n^{-2}\cdot O(n^{2\varepsilon})\cdot n = O(n^{-1+2\varepsilon}) = o(n^{-\frac{1}{2}+3\varepsilon}).$$

So in order to show (2.9), we only need to show

(2.27)
$$\sup\Big\{\Big|\int_{-\infty}^{t} \frac{dH_n^{(0)}(\beta,u)}{H_n^{(1)}(\beta,u)} - \int_{-\infty}^{t} \frac{dh^{(0)}(\beta,u)}{h^{(1)}(\beta,u)}\Big| : \beta\in\mathcal{B}, H_n^{(1)}(\beta,t)\ge n^{-\varepsilon}\Big\} = o(n^{-\frac{1}{2}+3\varepsilon})$$

almost surely. Now we define $\tilde{T}_n = \sup\{t : \beta\in\mathcal{B}, H_n^{(1)}(\beta,t)\ge n^{-\varepsilon}\}$, and let

$$\tilde{H}_n^{(1)}(\beta,t) = \begin{cases} H_n^{(1)}(\beta,t), & \text{if } t\le\tilde{T}_n, \\ H_n^{(1)}(\beta,\tilde{T}_n), & \text{if } t>\tilde{T}_n. \end{cases}$$

Then $\tilde{H}_n^{(1)}(\beta,t)\ge n^{-\varepsilon}$ for all $\beta\in\mathcal{B}$ and $-\infty<t<\infty$. Define $\tilde{h}^{(1)}(\beta,t)$ similarly as $\tilde{H}_n^{(1)}(\beta,t)$ and apply Lemma A.3 to $U_n(\beta,u) = n^{-2\varepsilon}/\tilde{H}_n^{(1)}(\beta,u)$ and $u_n(\beta,u) = n^{-2\varepsilon}/\tilde{h}^{(1)}(\beta,u)$, we obtain (2.27) and thus (2.9) holds.

Next we show (2.10). Notice that $F(t) = F(\beta_0,t)$, then under the restriction

$\{|\beta - \beta_0| \le n^{-3\varepsilon}, h^{(1)}(\beta, t) \ge n^{-\varepsilon}\}$, we have

$$\begin{aligned}
&|F(\beta,t) - F(t)| = |F(\beta,t) - F(\beta_0,t)| \\
&\quad = \left|\exp\left\{-\int_{u\le t} \frac{dh^{(0)}(\beta,u)}{h^{(1)}(\beta,u)}\right\} - \exp\left\{-\int_{u\le t} \frac{dh^{(0)}(\beta_0,u)}{h^{(1)}(\beta_0,u)}\right\}\right| \\
&\quad \le \left|\int_{u\le t} \frac{dh^{(0)}(\beta,u)}{h^{(1)}(\beta,u)} - \int_{u\le t} \frac{dh^{(0)}(\beta_0,u)}{h^{(1)}(\beta_0,u)}\right| \\
&\quad \le \left|\int_{u\le t} \frac{d\{h^{(0)}(\beta,u) - h^{(0)}(\beta_0,u)\}}{h^{(1)}(\beta,u)}\right| \\
&\qquad + \left|\int_{u\le t} \left(\frac{h^{(1)}(\beta_0,u) - h^{(1)}(\beta,u)}{h^{(1)}(\beta,u)h^{(1)}(\beta_0,u)}\right) dh^{(0)}(\beta_0,u)\right| \\
&\quad \le \{n^{\varepsilon} + n^{2\varepsilon} h^{(0)}(\beta_0,t)\} \cdot \sup\{|h^{(1)}(\beta_0,t) - h^{(1)}(\beta,t)|\} \\
&\quad = O(n^{\varepsilon-3\varepsilon}) + O(n^{2\varepsilon-3\varepsilon}) = O(n^{-\varepsilon}),
\end{aligned}$$

where the third inequality holds since for any $u \le t$, $\frac{1}{h^{(1)}(\beta,u)} \le \frac{1}{h^{(1)}(\beta,t)} \le n^{\varepsilon}$, and the second to last equality holds because $h^{(0)}(\beta_0,t) \le 1$ and $|h^{(1)}(\beta,t) - h^{(1)}(\beta_0,t)| = O(|\beta - \beta_0|)$ by Lemma A.2. Thus (2.10) holds. Finally, (2.11) can be easily obtained by applying the triangle inequality to (2.9) and (2.10) provided that $-\frac{1}{2} + 3\varepsilon \le -\varepsilon$, i.e., $0 < \varepsilon \le \frac{1}{8}$.

2.7.3 Proof of Theorem 2.2.2

Notice that

$$\alpha_0 = \int_{-\infty}^{\infty} t\, dF(t) = \int_0^{\infty} \{1 - F(t)\}\, dt - \int_{-\infty}^0 F(t)\, dt.$$

We thus have

$$\begin{aligned}
&\int_{-\infty}^{\infty} t\, d\hat{F}_{n,\beta}(t) - \alpha_0 = \int_{-\infty}^{\infty} t\, d\hat{F}_{n,\beta}(t) - \int_{-\infty}^{\infty} t\, dF(t) \\
&\quad = \left\{\int_0^{\infty} \{1 - \hat{F}_{n,\beta}(t)\}\, dt - \int_0^{\infty} \{1 - F(t)\}\, dt\right\} \qquad (2.28) \\
&\qquad - \left\{\int_{-\infty}^0 \hat{F}_{n,\beta}(t)\, dt - \int_{-\infty}^0 F(t)\, dt\right\}.
\end{aligned}$$

Since $f_X(x) > 0$ for all $-\infty < x < \infty$, from the proof of Lemma A.2, we then have

$$h^{(1)}(\beta, t) = \int_{-\infty}^{\infty} \bar{F}(t + (\beta - \beta_0)x)\bar{G}_{C|X}(t + \beta x|X = x) f_X(x)\, dx.$$

With $\beta_0 \neq 0$, when β satisfies $|\beta - \beta_0| \leq n^{-3\varepsilon}$, we have $\beta \neq 0$ for sufficiently large n. Then for any fixed $\beta \neq 0$ and $t \in (-\infty, \infty)$, one can always find a range of x such that $\bar{G}_{C|X}(t + \beta x|X = x) > 0$ and $\bar{F}(t + (\beta - \beta_0)x) > 0$ (since $\bar{F}(t) > 0$ for all $t < \infty$, even with $\beta = \beta_0$, we still have $\bar{F}(t + (\beta - \beta_0)x) > 0$). Therefore, we have $h^{(1)}(\beta, t) > 0$ for all $t \in (-\infty, \infty)$. Moreover, since $H_n^{(1)}(\beta, t) \to h^{(1)}(\beta, t)$ almost surely as $n \to \infty$, then with n sufficiently large, we have $H_n^{(1)}(\beta, t) > 0$ almost surely for any fixed $\beta \neq 0$ and $t \in (-\infty, \infty)$. Hence $T_n \to \infty$ as $n \to \infty$, where $T_n = \sup\left\{t : H_n^{(1)}(\beta, t) \geq n^{-\varepsilon}, |\beta - \beta_0| \leq n^{-3\varepsilon}\right\}$, as defined in (2.13).

Then at $\beta = \beta_0$, by the independence of e_0 and $C - \beta_0 X$ and the Markov's inequality, it follows that

$$\begin{aligned} h^{(1)}(\beta_0, T_n) &= P\{1(e_0 \geq T_n)\} \cdot P\{1(C - \beta_0 X \geq T_n)\} \\ &\leq P\{1(e_0 \geq T_n)\} \leq \frac{Ee_0^2}{T_n^2}. \end{aligned}$$

Since $H_n^{(1)}(\beta_0, T_n) \geq n^{-\varepsilon}$ implies $h^{(1)}(\beta_0, T_n) \geq n^{-\varepsilon}$, i.e., $1/h^{(1)}(\beta_0, T_n) \leq n^{\varepsilon}$, together with Condition 4 that $Ee_0^2 < \infty$, we have $T_n^2 \leq Ee_0^2/h^{(1)}(\beta_0, T_n) = O(n^{\varepsilon})$, i.e., $T_n \leq O(n^{\varepsilon/2})$. This implies that $T_n \to \infty$ in a rate no faster than $n^{\varepsilon/2}$.

Since the Kaplan-Meier estimator $\hat{F}_{n,\beta}(t)$ is set to 1 for $t > T_n$, equation (2.28) becomes

$$\begin{aligned} &\int_{-\infty}^{\infty} t\, d\hat{F}_{n,\beta}(t) - \alpha_0 \\ &= \left\{\int_0^{T_n} \{1 - \hat{F}_{n,\beta}(t)\}\, dt - \int_0^{\infty} \{1 - F(t)\}\, dt\right\} \\ &\quad - \left\{\int_{-\infty}^{0} \hat{F}_{n,\beta}(t)\, dt - \int_{-\infty}^{0} F(t)\, dt\right\} \\ &= \int_0^{T_n} \{F(t) - \hat{F}_{n,\beta}(t)\}\, dt - \int_{T_n}^{\infty} \{1 - F(t)\}\, dt - \int_{-\infty}^{0} \{\hat{F}_{n,\beta}(t) - F(t)\}\, dt. \end{aligned}$$

Then by Theorem 2.2.1, we have

$$\sup\left\{\int_0^{T_n}|F(t)-\hat{F}_{n,\beta}(t)|\;dt : |\beta-\beta_0|\le n^{-3\varepsilon}\right\}\le T_n\cdot O(n^{-\varepsilon})\le O(n^{-\frac{\varepsilon}{2}})$$

almost surely. For the second term, by the Markov's inequality,

$$\int_{T_n}^{\infty}\{1-F(t)\}\;dt\le\int_{T_n}^{\infty}P\{1(|e_0|\ge t)\}\;dt\le\int_{T_n}^{\infty}\frac{Ee_0^2}{t^2}\;dt\le\frac{Ee_0^2}{T_n}=o(1),$$

For the third term, we have

$$\begin{aligned}&\int_{-\infty}^{0}\{\hat{F}_{n,\beta}(t)-F(t)\}\;dt\\&\quad=\int_{-T_n}^{0}\{\hat{F}_{n,\beta}(t)-F(t)\}\;dt+\int_{-\infty}^{-T_n}\{\hat{F}_{n,\beta}(t)-F(t)\}\;dt\end{aligned}$$

where for the first part, similarly we have

$$\sup\left\{\int_{-T_n}^{0}|F(t)-\hat{F}_{n,\beta}(t)|\;dt : |\beta-\beta_0|\le n^{-3\varepsilon}\right\}\le T_n\cdot O(n^{-\varepsilon})\le O(n^{-\frac{\varepsilon}{2}})$$

almost surely. For the second part, it follows that

$$\begin{aligned}\int_{-\infty}^{-T_n}|F(t)-\hat{F}_{n,\beta}(t)|\;dt&\le\int_{-\infty}^{-T_n}F(t)\;dt+\int_{-\infty}^{-T_n}\hat{F}_{n,\beta}(t)\;dt\\&=\int_{T_n}^{\infty}F(-t)\;dt+\int_{-\infty}^{-T_n}\hat{F}_{n,\beta}(t)\;dt\le\frac{Ee_0^2}{T_n}+o(1)=o(1),\end{aligned}$$

where the last inequality holds because of the Markov's inequality

$$F(-t)=P\{1(e_0\le -t)\}\le P\{1(|e_0|\ge t)\}\le\frac{Ee_0^2}{t^2},$$

and the fact $\int_{-\infty}^{-T_n}\hat{F}_{n,\beta}(t)\;dt\to 0$ as $n\to\infty$. Therefore,

$$\sup\left\{\left|\int_{-\infty}^{\infty}t\,d\hat{F}_{n,\beta}(t)-\alpha_0\right| : |\beta-\beta_0|\le n^{-3\varepsilon}, H_n^{(1)}(\beta,t)\ge n^{-\varepsilon}\right\}=o(1).$$

We now have proved Theorem 2.2.2.

2.7.4 Proof of Theorem 2.2.3

Throughout the proof, we consider $\beta \in \mathcal{B}_n^K = \{\beta : |\beta - \beta_0| \leq Kn^{-1/2}\}$, where $K > 0$ is a constant. By the rule of integration by parts for the stochastic integral, which is warranted by the fact that $\hat{F}_{n,\beta}(t)$ is a right continuous non-decreasing submartingale, we notice that

$$\begin{aligned} n^{1/2}(\hat{\alpha}^*_{n,\beta} - \alpha^*_0) &= \int_{S^*}^{T^*} t\, d\big\{n^{1/2}(\hat{F}_{n,\beta}(t) - F(t))\big\} \\ &= T^* n^{1/2}(\hat{F}_{n,\beta}(T^*) - F(T^*)) - S^* n^{1/2}(\hat{F}_{n,\beta}(S^*) - F(S^*)) \\ &\quad - \int_{S^*}^{T^*} n^{1/2}(\hat{F}_{n,\beta}(t) - F(t))\, dt, \end{aligned} \tag{2.29}$$

where

$$n^{1/2}(\hat{F}_{n,\beta}(t) - F(t)) = n^{1/2}(\hat{F}_{n,\beta}(t) - F(\beta,t)) + n^{1/2}(F(\beta,t) - F(t))$$

with $F(\beta,t)$ being defined in (2.8) and $t \in [S^*, T^*]$.

Then by mean value theorem, term $n^{1/2}(\hat{F}_{n,\beta}(t) - F(\beta,t))$ can be rewritten as

$$\begin{aligned} & n^{1/2}(\hat{F}_{n,\beta}(t) - F(\beta,t)) \\ = \;& n^{1/2}\Bigg[\exp\Bigg\{-\int_{-\infty}^{t} \frac{dh^{(0)}(\beta,u)}{h^{(1)}(\beta,u)}\Bigg\} \\ & - \;\exp\Bigg\{-\int_{-\infty}^{t} \frac{dH_n^{(0)}(\beta,u)}{H_n^{(1)}(\beta,u)} - n^{-2}\sum_{i:\epsilon_{\beta,i}\leq t} O(H_n^{(1)}(\beta,\epsilon_{\beta,i})^{-2})\Bigg\}\Bigg] \\ = \;& z_n(\beta,t)\cdot n^{1/2}\Bigg\{\int_{-\infty}^{t} \frac{dH_n^{(0)}(\beta,u)}{H_n^{(1)}(\beta,u)} - \int_{-\infty}^{t} \frac{dh^{(0)}(\beta,u)}{h^{(1)}(\beta,u)} \\ & + n^{-2}\sum_{i:\epsilon_{\beta,i}\leq t} O(H_n^{(1)}(b,\epsilon_{\beta,i})^{-2})\Bigg\}, \end{aligned}$$

where

$$\begin{aligned} z_n(\beta,t) \;=\; & \exp\Bigg\{-\lambda\int_{-\infty}^{t} \frac{dh^{(0)}(\beta,u)}{h^{(1)}(\beta,u)} \\ & - (1-\lambda)\Bigg[\int_{-\infty}^{t} \frac{dH_n^{(0)}(\beta,u)}{H_n^{(1)}(\beta,u)} + n^{-2}\sum_{i:\epsilon_{\beta,i}\leq t} O(H_n^{(1)}(\beta,\epsilon_{\beta,i})^{-2})\Bigg]\Bigg\} \end{aligned} \tag{2.30}$$

is an intermediate value with $\lambda \in (0,1)$. Notice that $0 < z_n(\beta, t) < 1$ for any β, t and n. For any fixed β and n, $H_n^{(1)}(\beta, u)$ is decreasing as u increasing, so for $\epsilon_{\beta,i} \leq t \leq T^*$, we have $H_n^{(1)}(\beta, \epsilon_{\beta,i}) \geq H_n^{(1)}(\beta, T^*)$. Since $H_n^{(1)}(\beta, T^*) > 0$ almost surely, which is shown in the proof of Theorem 2.2.2, we have $O(H_n^{(1)}(\beta, T^*)^{-2}) = O_p(1)$ and therefore,

$$n^{1/2} \cdot n^{-2} \sum_{i:\epsilon_{\beta,i} \leq t} O(H_n^{(1)}(\beta, \epsilon_{\beta,i})^{-2}) \leq n^{-1/2} \cdot O(H_n^{(1)}(\beta, T^*)^{-2}) = O_p(n^{-1/2}) = o_p(1).$$

Then term $n^{1/2}(\hat{F}_{n,\beta}(t) - F(\beta, t))$ can be further rewritten as

$$\begin{aligned} & n^{1/2}(\hat{F}_{n,\beta}(t) - F(\beta, t)) \\ = \ & z_n(\beta, t) \cdot n^{1/2} \left\{ \int_{-\infty}^{t} \frac{dH_n^{(0)}(\beta, u)}{H_n^{(1)}(\beta, u)} - \int_{-\infty}^{t} \frac{dh^{(0)}(\beta, u)}{h^{(1)}(\beta, u)} \right\} + o_p(1). \end{aligned}$$

By subtracting and adding the term $\int_{-\infty}^{t} \frac{dH_n^{(0)}(\beta,u)}{h^{(1)}(\beta,u)}$, we get

$$\begin{aligned} & n^{1/2} \left\{ \int_{-\infty}^{t} \frac{dH_n^{(0)}(\beta, u)}{H_n^{(1)}(\beta, u)} - \int_{-\infty}^{t} \frac{dh^{(0)}(\beta, u)}{h^{(1)}(\beta, u)} \right\} \\ = \ & n^{1/2} \left\{ \int_{-\infty}^{t} \frac{dH_n^{(0)}(\beta, u)}{H_n^{(1)}(\beta, u)} - \int_{-\infty}^{t} \frac{dH_n^{(0)}(\beta, u)}{h^{(1)}(\beta, u)} \right\} \qquad (2.31) \\ & + n^{1/2} \left\{ \int_{-\infty}^{t} \frac{dH_n^{(0)}(\beta, u)}{h^{(1)}(\beta, u)} - \int_{-\infty}^{t} \frac{dh^{(0)}(\beta, u)}{h^{(1)}(\beta, u)} \right\}. \qquad (2.32) \end{aligned}$$

Term (2.31) can be rewritten as

$$\begin{aligned} (2.31) &= \int_{-\infty}^{t} \frac{-n^{1/2}(H_n^{(1)}(\beta, u) - h^{(1)}(\beta, u))}{H_n^{(1)}(\beta, u) h^{(1)}(\beta, u)} \, dH_n^{(0)}(\beta, u) \\ &= \int_{-\infty}^{t} \frac{-\mathbb{G}_n\{1(\epsilon_\beta \geq u)\}}{H_n^{(1)}(\beta, u) h^{(1)}(\beta, u)} \, dH_n^{(0)}(\beta, u) \\ &= \frac{1}{n} \sum_{i:\epsilon_{\beta,i} \leq t} \frac{-\mathbb{G}_n\{1(\epsilon_\beta \geq \epsilon_{\beta,i})\}\Delta_i}{H_n^{(1)}(\beta, \epsilon_{\beta,i}) h^{(1)}(\beta, \epsilon_{\beta,i})} \\ &= \frac{1}{n} \sum_{i=1}^{n} \frac{-\mathbb{G}_n\{1(\epsilon_\beta \geq \epsilon_{\beta,i})\}\Delta_i 1(\epsilon_{\beta,i} \leq t)}{H_n^{(1)}(\beta, \epsilon_{\beta,i}) h^{(1)}(\beta, \epsilon_{\beta,i})} \\ &= -\mathbb{G}_n \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{\Delta_i 1(\epsilon_\beta \geq \epsilon_{\beta,i}) 1(\epsilon_{\beta,i} \leq t)}{H_n^{(1)}(\beta, \epsilon_{\beta,i}) h^{(1)}(\beta, \epsilon_{\beta,i})} \right\}, \end{aligned}$$

where t is any fixed time point between S^* and T^*. Now for a given t, define

$$A_1(\beta, s; t) = \frac{\Delta 1(\epsilon_\beta \geq s) 1(s \leq t)}{H_n^{(1)}(\beta, s) h^{(1)}(\beta, s)} \quad \text{and} \quad A_2(\beta, s; t) = \frac{\Delta 1(\epsilon_\beta \geq s) 1(s \leq t)}{h^{(1)}(\beta, s)^2}.$$

As the same argument in the proof of Lemma A.1, the class of indicator functions of a half space is a VC-class, and thus a Donsker class. So $\{\Delta 1(\epsilon_\beta \geq s) 1(s \leq t)\}$ belongs to a Donsker class. Also as shown in Lemma A.1, $H_n^{(1)}(\beta, s)$ and $h^{(1)}(\beta, s)$ are both Donsker classes. Moveover, for $s \leq t \leq T^*$, we have already shown that $H_n^{(1)}(\beta, s)$ is bounded away from zero in probability and $h^{(1)}(\beta, s)$ is bounded away from zero. Hence $\{A_1(\beta, s; t)\}$ and $\{A_2(\beta, s; t)\}$ are both Donsker classes. Then the convex combinations

$$\left\{ \sum_{i=1}^n \frac{1}{n} A_1(\beta, \epsilon_{\beta,i}; t) \right\} = \left\{ \frac{1}{n} \sum_{i=1}^n \frac{\Delta_i 1(\epsilon_\beta \geq \epsilon_{\beta,i}) 1(\epsilon_{\beta,i} \leq t)}{H_n^{(1)}(\beta, \epsilon_{\beta,i}) h^{(1)}(\beta, \epsilon_{\beta,i})} \right\}$$

and

$$\left\{ \sum_{i=1}^n \frac{1}{n} A_2(\beta, \epsilon_{\beta,i}; t) \right\} = \left\{ \frac{1}{n} \sum_{i=1}^n \frac{\Delta_i 1(\epsilon_\beta \geq \epsilon_{\beta,i}) 1(\epsilon_{\beta,i} \leq t)}{h^{(1)}(\beta, \epsilon_{\beta,i})^2} \right\}$$

belong to the convex hull of $\{A_1(\beta, s)\}$ and $\{A_2(\beta, s)\}$ respectively, and thus are also Donsker classes. We then show that

$$\frac{1}{n} \sum_{i=1}^n \frac{\Delta_i 1(\epsilon_\beta \geq \epsilon_{\beta,i}) 1(\epsilon_{\beta,i} \leq t)}{h^{(1)}(\beta, \epsilon_{\beta,i})} \left[\frac{1}{H_n^{(1)}(\beta, \epsilon_{\beta,i})} - \frac{1}{h^{(1)}(\beta, \epsilon_{\beta,i})} \right]$$

converges to zero in quadratic mean through the following argument.

$$
\begin{aligned}
& \int \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{\Delta_i 1(y - \beta x \geq \epsilon_{\beta,i}) 1(\epsilon_{\beta,i} \leq t)}{h^{(1)}(\beta, \epsilon_{\beta,i})} \left[\frac{1}{H_n^{(1)}(\beta, \epsilon_{\beta,i})} - \frac{1}{h^{(1)}(\beta, \epsilon_{\beta,i})} \right] \right\}^2 dP_{Y,X} \\
\leq \; & \int \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^{(1)}(\beta, \epsilon_{\beta,i})} \left| \frac{1}{H_n^{(1)}(\beta, \epsilon_{\beta,i})} - \frac{1}{h^{(1)}(\beta, \epsilon_{\beta,i})} \right| \right\}^2 dP_{Y,X} \\
\leq \; & \left\{ \frac{1}{h^{(1)}(\beta, T^*)} \sup_{\beta \in \mathcal{B}_n^K, S^* \leq s \leq T^*} \left| \frac{1}{H_n^{(1)}(\beta, s)} - \frac{1}{h^{(1)}(\beta, s)} \right| \right\}^2 \int 1 \; dP_{Y,X} \\
= \; & \frac{1}{h^{(1)}(\beta, T^*)^2} \left\{ \sup_{\beta \in \mathcal{B}_n^K, S^* \leq s \leq T^*} \left| \frac{1}{H_n^{(1)}(\beta, s)} - \frac{1}{h^{(1)}(\beta, s)} \right| \right\}^2 \\
\leq \; & \frac{1}{h^{(1)}(\beta, T^*)^4 H_n^{(1)}(\beta, T^*)^2} \left\{ \sup_{\beta \in \mathcal{B}_n^K, S^* \leq s \leq T^*} |H_n^{(1)}(\beta, s) - h^{(1)}(\beta, s)| \right\}^2 \\
= \; & \left\{ \frac{1}{h^{(1)}(\beta, T^*)^6} + o_p(1) \right\} \cdot o_p(1) = o_p(1),
\end{aligned}
$$

where the second to last equality holds since

$$
|H_n^{(1)}(\beta, s) - h^{(1)}(\beta, s)| = |(\mathbb{P}_n - P) 1(\epsilon_\beta \geq s)|,
$$

with $\{1(\epsilon_\beta \geq s)\}$ being a Donsker class, and thus a Glivenko-Cantelli class, therefore,

$$
\sup_{\beta \in \mathcal{B}_n^K, S^* \leq s \leq T^*} |H_n^{(1)}(\beta, s) - h^{(1)}(\beta, s)| = o_p(1),
$$

and in addition, $1/H_n^{(1)}(\beta, T^*) = 1/h^{(1)}(\beta, T^*) + o_p(1)$ since $h^{(1)}(\beta, T^*)$ is away from 0 by the argument in the proof of Theorem 2.2.2. Hence by the relationship between the Donsker and the equicontinuity condition by Corollary 2.3.12 of van der Vaart and Wellner (1996), we have

$$
\mathbb{G}_n \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{\Delta_i 1(\epsilon_\beta \geq \epsilon_{\beta,i}) 1(\epsilon_{\beta,i} \leq t)}{H_n^{(1)}(\beta, \epsilon_{\beta,i}) h^{(1)}(\beta, \epsilon_{\beta,i})} \right\} = \mathbb{G}_n \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{\Delta_i 1(\epsilon_\beta \geq \epsilon_{\beta,i}) 1(\epsilon_{\beta,i} \leq t)}{h^{(1)}(\beta, \epsilon_{\beta,i})^2} \right\} + o_p(1),
$$

and thus

$$
(2.31) = -\mathbb{G}_n \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{\Delta_i 1(\epsilon_\beta \geq \epsilon_{\beta,i}) 1(\epsilon_{\beta,i} \leq t)}{h^{(1)}(\beta, \epsilon_{\beta,i})^2} \right\} + o_p(1).
$$

Now for a given t, define

$$
m_{1n}(\beta, s; t) = -\frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{\Delta_i 1(s \geq \epsilon_{\beta,i}) 1(t \geq \epsilon_{\beta,i})}{h^{(1)}(\beta, \epsilon_{\beta,i})^2} \right\} = -\mathbb{P}_n \left\{ \frac{\Delta 1(s \geq \epsilon_\beta) 1(t \geq \epsilon_\beta)}{h^{(1)}(\beta, \epsilon_\beta)^2} \right\},
$$

and we further show that

$$\mathbb{G}_n\{m_{1n}(\beta,\epsilon_\beta;t)\} = \mathbb{G}_n\{m_1(\beta,\epsilon_\beta;t)\} + o_p(1),$$

where $m_1(\beta,s;t)$ is defined in (2.16). First, using the same argument as before, $\{\Delta 1(s \geq \epsilon_\beta)1(t \geq \epsilon_\beta)/h^{(1)}(\beta,\epsilon_\beta)^2\}$ belongs to a Donsker class. Then $m_{1n}(\beta,s;t)$ and $m_1(\beta,s;t)$ are in the convex hull of $\{\Delta 1(s \geq \epsilon_\beta)1(t \geq \epsilon_\beta)/h^{(1)}(\beta,\epsilon_\beta)^2\}$ and thus also belong to Donsker classes. Moreover,

$$\begin{aligned}
& \int \{m_{1n}(\beta,\epsilon_\beta;t) - m_1(\beta,\epsilon_\beta;t)\}^2 \, dP_{Y,X} \\
= & \int \left\{(\mathbb{P}_n - P)\left[\frac{\Delta 1(y - \beta x \geq \epsilon_\beta)1(t \geq \epsilon_\beta)}{h^{(1)}(\beta,\epsilon_\beta)^2}\right]\right\}^2 dP_{Y,X} \\
\leq & \sup_{\beta\in\mathcal{B}_n^K, -\infty<s<\infty} \left\{(\mathbb{P}_n - P)\left[\frac{\Delta 1(s \geq \epsilon_\beta)1(t \geq \epsilon_\beta)}{h^{(1)}(\beta,\epsilon_\beta)^2}\right]\right\}^2 \int 1 \, dP_{Y,X} \\
= & \left\{\sup_{\beta\in\mathcal{B}_n^K, -\infty<s<\infty} \left|(\mathbb{P}_n - P)\left[\frac{\Delta 1(s \geq \epsilon_\beta)1(t \geq \epsilon_\beta)}{h^{(1)}(\beta,\epsilon_\beta)^2}\right]\right|\right\}^2 \\
= & o_p(1),
\end{aligned}$$

where the last equality holds because $\{\Delta 1(s \geq \epsilon_\beta)1(t \geq \epsilon_\beta)/h^{(1)}(\beta,\epsilon_\beta)^2\}$ is a Donsker class, and thus a Glivenko-Cantelli class, it follows that

$$\sup_{\beta\in\mathcal{B}_n^K, -\infty<s<\infty} \left|(\mathbb{P}_n - P)\left[\frac{\Delta 1(s \geq \epsilon_\beta)1(t \geq \epsilon_\beta)}{h^{(1)}(\beta,\epsilon_\beta)^2}\right]\right| \to 0$$

in probability. Again by corollary 2.3.12 in van der Vaart and Wellner (1996), we have (2.31) $= \mathbb{G}_n\{m_1(\beta,\epsilon_\beta;t)\} + o_p(1)$. Next we will show that

$$\mathbb{G}_n\{m_1(\beta,\epsilon_\beta;t)\} = \mathbb{G}_n\{m_1(\beta_0,\epsilon_0;t)\} + o_p(1).$$

Since we have already shown that $\{m_1(\beta,s;t)\}$ belongs to a Donsker class, we only have to show that $m_1(\beta,\epsilon_\beta;t) - m_1(\beta_0,\epsilon_0;t)$ converges to zero in quadratic mean.

First, it follows that

$$\begin{aligned}
& \int \{m_1(\beta, \epsilon_\beta; t) - m_1(\beta_0, \epsilon_0; t)\}^2 \, dP_{Y,X} \\
= \; & \int \left\{ -P \left[\frac{\Delta 1(y - \beta x \geq \epsilon_\beta) 1(t \geq \epsilon_\beta)}{h^{(1)}(\beta, \epsilon_\beta)^2} - \frac{\Delta 1(y - \beta_0 x \geq \epsilon_0) 1(t \geq \epsilon_0)}{h^{(1)}(\beta_0, \epsilon_0)^2} \right] \right\}^2 dP_{Y,X} \\
\leq \; & \frac{1}{h^{(1)}(\beta, T^*)^2 h^{(1)}(\beta_0, T^*)^2} \int PI^2 \, dP_{Y,X} \\
= \; & \left[\frac{1}{h^{(1)}(\beta_0, T^*)^4} + o(1) \right] \int PI^2 \, dP_{Y,X},
\end{aligned}$$

where

$$I = 1(y - \beta x \geq \epsilon_\beta) 1(t \geq \epsilon_\beta) h^{(1)}(\beta_0, \epsilon_0)^2 - 1(y - \beta_0 x \geq \epsilon_0) 1(t \geq \epsilon_0) h^{(1)}(\beta, \epsilon_\beta)^2.$$

Since both the indicator and $h^{(1)}$ functions are between 0 and 1, so $-1 \leq I \leq 1$ and it follows that

$$\begin{aligned}
I^2 \; \leq \; & \left| 1(y - \beta x \geq \epsilon_\beta) 1(t \geq \epsilon_\beta) h^{(1)}(\beta_0, \epsilon_0)^2 - 1(y - \beta_0 x \geq \epsilon_0) 1(t \geq \epsilon_0) h^{(1)}(\beta, \epsilon_\beta)^2 \right| \\
\leq \; & 1(y - \beta x \geq \epsilon_\beta) 1(t \geq \epsilon_\beta) |h^{(1)}(\beta_0, \epsilon_0)^2 - h^{(1)}(\beta, \epsilon_\beta)^2| \\
& \quad + h^{(1)}(\beta, \epsilon_\beta)^2 |1(y - \beta x \geq \epsilon_\beta) 1(t \geq \epsilon_\beta) - 1(y - \beta_0 x \geq \epsilon_0) 1(t \geq \epsilon_0)| \\
\leq \; & 2 |h^{(1)}(\beta_0, \epsilon_0) - h^{(1)}(\beta, \epsilon_\beta)| \\
& \quad + |1\{\epsilon_\beta \leq \min(y - \beta x, t)\} - 1\{\epsilon_0 \leq \min(y - \beta_0 x, t)\}| \\
\leq \; & 2 \big\{ |h^{(1)}(\beta_0, \epsilon_0) - h^{(1)}(\beta_0, \epsilon_\beta)| + |h^{(1)}(\beta_0, \epsilon_\beta) - h^{(1)}(\beta, \epsilon_\beta)| \big\} \\
& \quad + |1\{\epsilon_\beta \leq \min(y - \beta x, t)\} - 1\{\epsilon_\beta \leq \min(y - \beta_0 x, t)\}| \\
& \quad + |1\{\epsilon_\beta \leq \min(y - \beta_0 x, t)\} - 1\{\epsilon_0 \leq \min(y - \beta_0 x, t)\}| \\
= \; & 2\{I_1 + I_2\} + I_3 + I_4.
\end{aligned}$$

For I_1, by Lemma A.2 we have $I_1 = O(|\epsilon_0 - \epsilon_\beta|) = O(|(\beta - \beta_0) X|)$, Then together with Condition 1, it follows that $PI_1 = O(|\beta - \beta_0|) = O(n^{-1/2})$. For I_2, it also follows by lemma A.2 that

$$PI_2 \leq \sup_{\beta \in \mathcal{B}_n^K, -\infty < s < \infty} |h^{(1)}(\beta_0, s) - h^{(1)}(\beta, s)| = O(|\beta - \beta_0|) = O(n^{-1/2}).$$

Then for I_3, since the joint density for (Y, X) follows

$$f_{Y,X}(y,x) = \big\{f(y-\beta_0 x)\bar{G}_{C|X}(y|X=x) + g_{C|X}(y|X=x)\bar{F}(y-\beta_0 x)\big\}f_X(x),$$

it is easily seen that for any s and s',

$$\begin{aligned}
& P|1(\epsilon_\beta \le s) - 1(\epsilon_\beta \le s')| \\
= \;& P|1(Y-\beta X \le s) - 1(Y - \beta X \le s')| \\
= \;& \int_{-\infty}^{\infty}\left\{\int_{\beta x+\min(s,s)}^{\beta x+\max(s,s')} \left[f(y-\beta_0 x)\bar{G}_{C|X}(y|x) + g_{C|X}(y|x)\bar{F}(y-\beta_0 x)\right]\,dy\right\}f_X(x)\,dx \\
\le \;& \int_{-\infty}^{\infty}\left\{\int_{\beta x+\min(s,s)}^{\beta x+\max(s,s')} \left[f(y-\beta_0 x) + g_{C|X}(y|x)\right]\,dy\right\}f_X(x)\,dx \\
\le \;& (C_1+C_3)|s'-s|\int_{-\infty}^{\infty} f_X(x)\,dx = (C_1+C_3)|s'-s|,
\end{aligned}$$

where C_1 and C_3 are two constants introduced in the proof of Lemma A.2 such that $f(\cdot) \le C_3$ and $g_{C|X}(\cdot|X=x) \le C_1$, guaranteed by Conditions 2 and 3. So $PI_3 \le (C_1+C_3)|(y-\beta x)-(y-\beta_0 x)| = (C_1+C_3)|(\beta-\beta_0)x|$. Finally for I_4, similarly it follows that for any s,

$$\begin{aligned}
& P|1(\epsilon_\beta \le s) - 1(\epsilon_0 \le s)| \\
= \;& \int_{-\infty}^{\infty}\left\{\int_{\min(\beta x,\beta_0 x)+s}^{\max(\beta x,\beta_0 x)+s} [f(y-\beta_0 x)\bar{G}_{C|X}(y|x) + g_{C|X}(y|x)\bar{F}(y-\beta_0 x)]\,dy\right\}f_X(x)\,dx \\
\le \;& (C_1+C_3)|\beta-\beta_0|\int_{-\infty}^{\infty}|x|f_X(x)\,dx \\
= \;& O(|\beta-\beta_0|) = O(n^{-1/2}).
\end{aligned}$$

Hence $PI_4 = P|1\{\epsilon_\beta \le \min(y-\beta_0 x, t)\} - 1\{\epsilon_0 \le \min(y-\beta_0 x,t)\}| = O(n^{-1/2})$. Thus,

$$\begin{aligned}
\int PI^2\,dP_{Y,X} &\le \int \{O(n^{-1/2}) + (C_1+C_3)|(\beta-\beta_0)x|\}\,dP_{Y,X} \\
&= O(n^{-1/2}) + O(|\beta-\beta_0|)\int |x|\,dP_{Y,X} = O(n^{-1/2}).
\end{aligned}$$

Therefore,

$$\int \{m_1(\beta,\epsilon_\beta;t) - m_1(\beta_0,\epsilon_0;t)\}^2\,dP_{Y,X} = O(n^{-1/2}) = o(1).$$

By corollary 2.3.12 of van der Vaart and Wellner (1996), we have

$$\mathbb{G}_n\{m_1(\beta, \epsilon_\beta; t)\} = \mathbb{G}_n\{m_1(\beta_0, \epsilon_0; t)\} + o_p(1).$$

Thus, term (2.31) $= \mathbb{G}_n\{m_1(\beta_0, \epsilon_0; t)\} + o_p(1)$.

Next we look at term (2.32). For any fixed time point $t \in [S^*, T^*]$, it follows that

$$\begin{aligned}
(2.32) &= \int_{-\infty}^{t} \frac{dn^{1/2}(H_n^{(0)}(\beta, u) - h^{(0)}(\beta, u))}{h^{(1)}(\beta, u)} \\
&= \int_{-\infty}^{t} \frac{d\mathbb{G}_n\{\Delta 1(\epsilon_\beta \le u)\}}{h^{(1)}(\beta, u)} \\
&= \mathbb{G}_n\left\{\int_{-\infty}^{t} \frac{d\{\Delta 1(\epsilon_\beta \le u)\}}{h^{(1)}(\beta, u)}\right\} \\
&= \mathbb{G}_n\left\{\frac{\Delta 1(\epsilon_\beta \le t)}{h^{(1)}(\beta, \epsilon_\beta)}\right\} = \mathbb{G}_n\{m_2(\beta, \epsilon_\beta; t, \Delta)\},
\end{aligned}$$

where $m_2(\beta, s; t, \Delta)$ is defined in (2.17). Then we will show that

$$\mathbb{G}_n\{m_2(\beta, \epsilon_\beta; t, \Delta)\} = \mathbb{G}_n\{m_2(\beta_0, \epsilon_0; t, \Delta)\} + o_p(1). \tag{2.33}$$

By the same argument as before, $\{\Delta 1(\epsilon_\beta \le t)/h^{(1)}(\beta, \epsilon_\beta)\}$ is a Donsker class. Furthermore,

$$\begin{aligned}
&\int \left\{\frac{\Delta 1(\epsilon_\beta \le t)}{h^{(1)}(\beta, \epsilon_\beta)} - \frac{\Delta 1(\epsilon_0 \le t)}{h^{(1)}(\beta_0, \epsilon_0)}\right\}^2 dP_{Y,\Delta,X} \\
\le\ & \frac{1}{h^{(1)}(\beta, T^*)^2 h^{(1)}(\beta_0, T^*)^2} \int \{1(\epsilon_\beta \le t)h^{(1)}(\beta_0, \epsilon_0) - 1(\epsilon_0 \le t)h^{(1)}(\beta, \epsilon_\beta)\}^2\, dP_{Y,X} \\
\le\ & \left[\frac{1}{h^{(1)}(\beta_0, T^*)^4} + o(1)\right] \int |1(\epsilon_\beta \le t)h^{(1)}(\beta_0, \epsilon_0) - 1(\epsilon_0 \le t)h^{(1)}(\beta, \epsilon_\beta)|\, dP_{Y,X} \\
=\ & \left[\frac{1}{h^{(1)}(\beta, T^*)^4} + o(1)\right] O(n^{-1/2}) = o(1),
\end{aligned}$$

where the first equality holds since by the similar argument as for $\int PI^2\, dP_{Y,X}$, we have

$$\int |1(\epsilon_\beta \le t)h^{(1)}(\beta_0, \epsilon_0) - 1(\epsilon_0 \le t)h^{(1)}(\beta, \epsilon_\beta)|\, dP_{Y,X} = O(n^{-1/2}).$$

Again, by corollary 2.3.12 of van der Vaart and Wellner (1996), equation (2.33) holds. Hence we have

(2.34)

$$\mathbb{G}_n\{m_1(\beta,\epsilon_\beta;t)+m_2(\beta,\epsilon_\beta;t,\Delta)\}=\mathbb{G}_n\{m_1(\beta_0,\epsilon_0;t)+m_2(\beta_0,\epsilon_0;t,\Delta)\}+o_p(1),$$

and thus

$$\begin{aligned}
& n^{1/2}(\hat{F}_{n,\beta}(t)-F(\beta,t)) \\
= \;& z_n(\beta,t)\cdot\mathbb{G}_n\{m_1(\beta,\epsilon_\beta;t)+m_2(\beta,\epsilon_\beta;t,\Delta)\}+o_p(1) \\
= \;& (1-F(\beta,t))\cdot\mathbb{G}_n\{m_1(\beta,\epsilon_\beta;t)+m_2(\beta,\epsilon_\beta;t,\Delta)\}+o_p(1) \\
= \;& \bar{F}(t)\cdot\mathbb{G}_n\{m_1(\beta_0,\epsilon_0;t)+m_2(\beta_0,\epsilon_0;t,\Delta)\}+o_p(1),
\end{aligned}$$

where the second equality holds because $\hat{F}_{n,\beta}(t)\to F(\beta,t)$ as $n\to\infty$, and thus

$$z_n(\beta,t)\to\exp\bigg\{-\int_{-\infty}^{t}\frac{dh^{(0)}(\beta,u)}{h^{(1)}(\beta,u)}\bigg\}=1-F(\beta,t),$$

with $z_n(\beta,t)$ being defined in (2.30), and the last equality holds because of equation (2.34) and the fact that

$$1-F(\beta,t)=1-F(\beta_0,t)+o(1)=\bar{F}(t)+o(1)$$

for $\beta\in\mathcal{B}_n^K$. Finally for the term $n^{1/2}(F(\beta,t)-F(t))$, first order Taylor expansion gives

$$\begin{aligned}
n^{1/2}(F(\beta,t)-F(t)) &= n^{1/2}(F(\beta,t)-F(\beta_0,t)) \\
&= n^{1/2}(\beta-\beta_0)\{\dot{F}_\beta(\beta_0,t)+o_p(1)\},
\end{aligned}$$

where $\dot{F}_\beta(\beta,t)$ is the derivative of $F(\beta,t)$ with respect to β. Proposition 2.3.4 shows that $n^{1/2}(\beta-\beta_0)$ converges to a normal random variable when $\beta\in\mathcal{B}_n^K$, i.e.,

$$n^{1/2}(\beta-\beta_0)=\mathbb{G}_n\{m_3(\beta_0,\epsilon_0;X,\Delta)\}+o_p(1),$$

where the representation of $m_3(\beta_0, \epsilon_0; X, \Delta)$ is given in Proposition 2.3.4. Hence

$$n^{1/2}(F(\beta, t) - F(t)) = \dot{F}_\beta(\beta_0, t) \cdot \mathbb{G}_n\{m_3(\beta_0, \epsilon_0; X, \Delta)\} + o_p(1),$$

and therefore

$$\begin{aligned} & n^{1/2}(\hat{F}_{n,\beta}(t) - F(t)) \\ = \;& \mathbb{G}_n\big\{\bar{F}(t)(m_1(\beta_0, \epsilon_0; t) + m_2(\beta_0, \epsilon_0; t, \Delta)) + \dot{F}_\beta(\beta_0, t) m_3(\beta_0, \epsilon_0; X, \Delta)\big\} + o_p(1). \end{aligned}$$

Thus by (2.29) and the above representation, when $\hat{\beta}_n$ is a consistent and asymptotically normal estimate of β_0, $n^{1/2}(\hat{\alpha}^*_{n,\hat{\beta}_n} - \alpha^*_0)$ is asymptotically normal with the following asymptotic representation:

$$\begin{aligned} & n^{1/2}(\hat{\alpha}^*_{n,\hat{\beta}_n} - \alpha^*_0) \\ = \;& \mathbb{G}_n\bigg\{T^*\big[\bar{F}(T^*)(m_1(\beta_0, \epsilon_0; T^*) + m_2(\beta_0, \epsilon_0; T^*, \Delta)) + \dot{F}_\beta(\beta_0, T^*) m_3(\beta_0, \epsilon_0; X, \Delta)\big] \\ & - S^*\big[\bar{F}(S^*)(m_1(\beta_0, \epsilon_0; S^*) + m_2(\beta_0, \epsilon_0; S^*, \Delta)) + \dot{F}_\beta(\beta_0, S^*) m_3(\beta_0, \epsilon_0; X, \Delta)\big] \\ & - \int_{S^*}^{T^*} \big[\bar{F}(t)(m_1(\beta_0, \epsilon_0; t) + m_2(\beta_0, \epsilon_0; t, \Delta)) + \dot{F}_\beta(\beta_0, t) m_3(\beta_0, \epsilon_0; X, \Delta)\big]\, dt\bigg\} \\ & + o_p(1). \end{aligned}$$

2.7.5 Proof of Proposition 2.3.4

We will show the consistency of $\hat{\beta}_n$ first, then prove the root-n convergence rate and finally prove the asymptotic normality.

(1) For any $\beta \in \mathcal{B}$, first we show that with probability approaching to one,

$$\|\Psi_n(\beta, H_n^{(1)}, H_n^{(2)}) - \Psi(\beta, h^{(1)}, h^{(2)})\| \to 0, \tag{2.35}$$

where $\|\cdot\|$ denotes the supremum norm. By the triangle inequality, we have

$$\begin{aligned} & \|\Psi_n(\beta, H_n^{(1)}, H_n^{(2)}) - \Psi(\beta, h^{(1)}, h^{(2)})\| \\ & \le \|(\mathbb{P}_n - P)\{(H_n^{(1)} X - H_n^{(2)})\Delta\}\| + \|P\{(H_n^{(1)} - h^{(1)}) X \Delta\}\| \\ & \quad + \|P\{(H_n^{(2)} - h^{(2)})\Delta\}\| \end{aligned}$$

By the same argument in the proof of Lemma A.1, the class of functions $\{1(\epsilon_\beta \geq t)\}$ is a VC-class and thus a Donsker class, and the set of functions $\{1(\epsilon_\beta \geq t)X\}$ is also a Donsker class when X is a random variable with a finite second moment. Since Donsker classes are Glivenko-Cantelli classes, it follows that,

$$\|H_n^{(1)}(\beta,t) - h^{(1)}(\beta,t)\| \;\; = \;\; \|(\mathbb{P}_n - P)1(\epsilon_\beta \geq t)\| \to 0$$

in probability and

$$\|H_n^{(2)}(\beta,t) - h^{(2)}(\beta,t)\| \;\; = \;\; \|(\mathbb{P}_n - P)\{1(\epsilon_\beta \geq t)X\}\| \to 0$$

in probability. Since $P|X\Delta| \leq P|X| < \infty$ and $P|\Delta| \leq 1$, we then have

$$\|P\{(H_n^{(1)} - h^{(1)})X\Delta\}\| \leq \|H_n^{(1)} - h^{(1)}\| P|X\Delta| \to 0$$

and

$$\|P\{(H_n^{(2)} - h^{(2)})\Delta\}\| \leq \|H_n^{(2)} - h^{(2)}\| P|\Delta| \to 0$$

in probability. Since $H_n^{(k)}(\beta,t)$ and $h^{(k)}(\beta,t)$, $k = 1, 2$, are in the convex hull of $\{1(\epsilon_\beta \geq t)\}$ and $\{1(\epsilon_\beta \geq t)X\}$, respectively, by Theorems 2.10.2 and 2.10.3 in van der Vaart and Wellner (1996), $H_n^{(k)}(\beta,t)$ and $h^{(k)}(\beta,t)$ are Donsker classes and thus Glivenko-Cantelli classes. Since both X and Δ have finite second moments, $\{(H_n^{(1)}X - H_n^{(2)})\Delta\}$ is also a Donsker class and thus a Glivenko-Cantelli class. So we have

$$\|(\mathbb{P}_n - P)\{(H_n^{(1)}X - H_n^{(2)})\Delta\}\| \to 0$$

in probability. Therefore, (2.35) holds. Together with the assumption that $\hat{\beta}_n$ satisfying the equation

$$\Psi_n(\hat{\beta}_n, H_n^{(1)}(\hat{\beta}_n, \cdot), H_n^{(2)}(\hat{\beta}_n, \cdot)) = o_p(1),$$

we have the following inequalities:

$$\begin{aligned}
&|\Psi(\hat{\beta}_n, h^{(1)}(\hat{\beta}_n,\cdot), h^{(2)}(\hat{\beta}_n,\cdot))| \\
&\quad \le |\Psi_n(\hat{\beta}_n, H_n^{(1)}(\hat{\beta}_n,\cdot), H_n^{(2)}(\hat{\beta}_n,\cdot))| \\
&\qquad + |\Psi(\hat{\beta}_n, h^{(1)}(\hat{\beta}_n,\cdot), h^{(2)}(\hat{\beta}_n,\cdot)) - \Psi_n(\hat{\beta}_n, H_n^{(1)}(\hat{\beta}_n,\cdot), H_n^{(2)}(\hat{\beta}_n,\cdot))| \\
&\quad \le o_p(1) + \ \|\Psi_n(\beta, H_n^{(1)}, H_n^{(2)}) - \Psi(\beta, h^{(1)}, h^{(2)})\| = o_p(1). \qquad (2.36)
\end{aligned}$$

Since β_0 is the unique solution to $\Psi(\beta, h^{(1)}(\beta,\cdot), h^{(2)}(\beta,\cdot)) = 0$, then for any fixed $\varepsilon > 0$, there exists a $\delta > 0$ such that

$$\{\hat{\beta}_n : |\hat{\beta}_n - \beta_0| > \varepsilon\} \subseteq \{\hat{\beta}_n : |\Psi(\hat{\beta}_n, h^{(1)}(\hat{\beta}_n,\cdot), h^{(2)}(\hat{\beta}_n,\cdot)) - 0| > \delta\}.$$

From (2.36), for any $\delta > 0$, we have $P\{|\Psi(\hat{\beta}_n, h^{(1)}(\hat{\beta}_n,\cdot), h^{(2)}(\hat{\beta}_n,\cdot)) - 0| > \delta\} \to 0$ as $n \to 0$. Hence $P\{|\hat{\beta}_n - \beta_0| > \varepsilon\} \to 0$ as $n \to \infty$, i.e., $|\hat{\beta}_n - \beta_0| = o_p(1)$.

(2) Now let $\mathcal{B}_0 \subset \mathcal{B}$ be a neighborhood of β_0 and $\|\cdot\|_0$ be the supremum norm in $\mathcal{B}_0$. For any $\beta \in \mathcal{B}_0$, we have

$$\begin{aligned}
(2.37)\ &\left\| n^{1/2}\{\Psi_n(\beta, H_n^{(1)}(\beta,t), H_n^{(2)}(\beta,t)) - \Psi(\beta, h^{(1)}(\beta,t), h^{(2)}(\beta,t))\}\right\|_0 \\
= \ &\Big\| n^{1/2}(\mathbb{P}_n - P)\{[H_n^{(1)}(\beta,t)X - H_n^{(2)}(\beta,t)]\Delta\} \\
&\quad + n^{1/2}P\{[H_n^{(1)}(\beta,t) - h^{(1)}(\beta,t)]X\Delta\} + n^{1/2}P\{[H_n^{(2)}(\beta,t) - h^{(2)}(\beta,t)]\Delta\}\Big\|_0 \\
\le \ &\left\|\mathbb{G}_n\{[H_n^{(1)}(\beta,t)X - H_n^{(2)}(\beta,t)]\Delta\}\right\|_0 \\
&\quad + \left\|\mathbb{G}_n\{1(\epsilon_\beta \ge t)\}\right\|_0 P|X\Delta| + \|\mathbb{G}_n\{1(\epsilon_\beta \ge t)X\}\|_0 P|\Delta|.
\end{aligned}$$

Since $\{1(\epsilon_\beta \ge t)\}$, $\{1(\epsilon_\beta \ge t)X\}$ and $\{[H_n^{(1)}(\beta,t)X - H_n^{(2)}(\beta,t)]\Delta\}$ are all Donsker classes, and $P|X\Delta| < \infty$, $P|\Delta| \le 1$, we have (2.37) $= O_p(1)$. Together with the fact that $\Psi(\beta_0, h^{(1)}(\beta_0,\cdot), h^{(2)}(\beta_0,\cdot)) = 0$ and $n^{1/2}\Psi_n(\hat{\beta}_n, H_n^{(1)}(\hat{\beta}_n,\cdot), H_n^{(2)}(\hat{\beta}_n,\cdot)) = o_p(1)$, we have

$$\begin{aligned} O_p(1) &= -n^{1/2}\{\Psi_n(\hat{\beta}_n, H_n^{(1)}(\hat{\beta}_n, \cdot), H_n^{(2)}(\hat{\beta}_n, \cdot)) - \Psi(\hat{\beta}_n, h^{(1)}(\hat{\beta}_n, \cdot), h^{(2)}(\hat{\beta}_n, \cdot))\} \\ &= o_p(1) + n^{1/2}\Psi(\hat{\beta}_n, h^{(1)}(\hat{\beta}_n, \cdot), h^{(2)}(\hat{\beta}_n, \cdot)) - n^{1/2}\Psi(\beta_0, h^{(1)}(\beta_0, \cdot), h^{(2)}(\beta_0, \cdot)) \\ &= o_p(1) + n^{1/2}(\hat{\beta}_n - \beta_0)\dot{\Psi}_\beta(\tilde{\beta}, h^{(1)}(\tilde{\beta}, \cdot), h^{(2)}(\tilde{\beta}, \cdot)), \end{aligned}$$

where $\tilde{\beta}$ is a value between β_0 and $\hat{\beta}_n$, and thus $|\tilde{\beta} - \beta_0| = o_p(1)$. By the continuity of $\dot{\Psi}_\beta(\beta, h^{(1)}(\beta, \cdot), h^{(2)}(\beta, \cdot))$ we have

$$O_p(1) = o_p(1) + n^{1/2}(\hat{\beta}_n - \beta_0)\{\dot{\Psi}_\beta(\beta_0, h^{(1)}(\beta_0, \cdot), h^{(2)}(\beta_0, \cdot)) + o_p(1)\},$$

and by the assumption that $\dot{\Psi}_\beta(\beta_0, h^{(1)}(\beta_0, \cdot), h^{(2)}(\beta_0, \cdot))$ is nonsingular, it follows that

$$n^{1/2}(\hat{\beta}_n - \beta_0) = \dot{\Psi}_\beta^{-1}(\beta_0, h^{(1)}(\beta_0, \cdot), h^{(2)}(\beta_0, \cdot))O_p(1) + o_p(1) = O_p(1).$$

Therefore, $|\hat{\beta}_n - \beta_0| = O_p(n^{-1/2})$.

(3) Now we prove the asymptotic normality of $\hat{\beta}_n$. First, we show that $h^{(1)}(\beta, t)$ and $h^{(2)}(\beta, t)$ are differentiable in β with uniformly bounded and continuous derivatives. The conditional density of $Y|X$ is

$$f_{Y|X}(y|X = x) = f(y - \beta_0 X)\bar{G}_{C|X}(y|X = x) + g_{C|X}(y|X = x)\bar{F}(y - \beta_0 x),$$

which is uniformly bounded and continuous by Conditions 2 and 3. Then

$$h^{(1)}(\beta, t) = P\{1(Y - \beta X \geq t)\} = \int_{-\infty}^{\infty} \bar{F}_{Y|X}(t + \beta x|X = x)f_X(x)\ dx,$$

and

$$\frac{\partial h^{(1)}(\beta, t)}{\partial \beta} = \int_{-\infty}^{\infty} -f_{Y|X}(t + \beta x|X = x)xf_X(x)\ dx.$$

Since $f_{Y|X}(\cdot|X = x)$ is uniformly bounded and X has a finite second moment, then

$$\int_{-\infty}^{\infty} f_{Y|X}(t + \beta x|X = x)|x|f_X(x)\ dx < \infty.$$

Hence $\frac{\partial h^{(1)}(\beta,t)}{\partial\beta}$ is bounded and continuous. Similarly, $h^{(2)}(\beta,t)$ also has an uniformly bounded and continuous derivative in β. Then by the dominated convergence theorem, $\Psi(\beta, h^{(1)}(\beta,\cdot), h^{(2)}(\beta,\cdot))$ is differentiable with respect to β and the derivative is continuous and bounded.

Since we have already shown that $|\hat{\beta}_n - \beta_0| = O_p(n^{-1/2})$, now we consider a root-n neighborhood of β_0, i.e., $\mathcal{B}_n^K = \{\beta \in \mathcal{B} : |\beta - \beta_0| \le Kn^{-1/2}\}$, where $K > 0$ is a constant. For any $\beta \in \mathcal{B}_n^K$, it follows that

$$\begin{aligned}
& n^{\frac{1}{2}}\big\{\Psi_n(\beta, H_n^{(1)}(\beta,\epsilon_\beta), H_n^{(2)}(\beta,\epsilon_\beta)) - \Psi_n(\beta_0, H_n^{(1)}(\beta_0,\epsilon_0), H_n^{(2)}(\beta_0,\epsilon_0))\big\} \\
= \;& -n^{\frac{1}{2}}\mathbb{P}_n\big\{[H_n^{(2)}(\beta,\epsilon_\beta) - H_n^{(2)}(\beta_0,\epsilon_0)]\Delta\big\} + n^{\frac{1}{2}}\mathbb{P}_n\big\{[H_n^{(1)}(\beta,\epsilon_\beta) - H_n^{(1)}(\beta_0,\epsilon_0)]X\Delta\big\} \\
= \;& -\mathbb{G}_n\big\{[H_n^{(2)}(\beta,\epsilon_\beta) - H_n^{(2)}(\beta_0,\epsilon_0)]\Delta\big\} - n^{\frac{1}{2}}P\big\{[H_n^{(2)}(\beta,\epsilon_\beta) - H_n^{(2)}(\beta_0,\epsilon_0)]\Delta\big\} \\
& + \mathbb{G}_n\{[H_n^{(1)}(\beta,\epsilon_\beta) - H_n^{(1)}(\beta_0,\epsilon_0)]X\Delta\} + n^{\frac{1}{2}}P\{[H_n^{(1)}(\beta,\epsilon_\beta) - H_n^{(1)}(\beta_0,\epsilon_0)]X\Delta\} \\
= \;& A_1 + A_2 + A_3 + A_4.
\end{aligned}$$

Term A_1 converges to zero in probability because $\{H_n^{(2)}(\beta,\cdot)\Delta\}$ belongs to a Donsker class and $\{[H_n^{(2)}(\beta,\epsilon_\beta) - H_n^{(2)}(\beta_0,\epsilon_0)]\Delta\}$ converges to zero in quadratic mean, which is warranted through the following argument.

$$\begin{aligned}
& P\big\{[H_n^{(2)}(\beta,\epsilon_\beta) - H_n^{(2)}(\beta_0,\epsilon_0)]^2\Delta^2\big\} \\
\le \;& P\big\{|H_n^{(2)}(\beta,\epsilon_\beta) - H_n^{(2)}(\beta_0,\epsilon_0)|\big\} \\
\le \;& P\big\{|H_n^{(2)}(\beta,\epsilon_\beta) - h^{(2)}(\beta,\epsilon_\beta)| + |h^{(2)}(\beta,\epsilon_\beta) - h^{(2)}(\beta_0,\epsilon_0)| \\
& + |h^{(2)}(\beta_0,\epsilon_0) - H_n^{(2)}(\beta_0,\epsilon_0)|\big\} \\
= \;& o(1) + P|h^{(2)}(\beta,\epsilon_\beta) - h^{(2)}(\beta_0,\epsilon_0)| + o(1) \\
= \;& o(1) + O(n^{-1/2}) + o(1) = o(1),
\end{aligned}$$

where the first inequality holds since both $H_n^{(2)}(\cdot,\cdot)$ and Δ are between 0 and 1; the first equality holds since $\{1(\epsilon_\beta \ge t)X\}$ is a Glivenko-Cantelli class and thus

$H_n^{(2)}(\beta,t)-h^{(2)}(\beta,t)=(\mathbb{P}_n-P)\{1(\epsilon_\beta\geq t)X\}=o(1)$ for any $\beta\in\mathcal{B}$ and $-\infty<t<\infty$; and the second equality holds since $|\epsilon_\beta-\epsilon_0|=|(\beta-\beta_0)X|$, it follows by Lemma A.2 that for any $\beta\in\mathcal{B}_n^K$, $|h^{(2)}(\beta,\epsilon_\beta)-h^{(2)}(\beta_0,\epsilon_0)|=O(|\beta-\beta_0|+|(\beta-\beta_0)X|)=O(n^{-1/2})+O(n^{-1/2})|X|$, thus $P|h^{(2)}(\beta,\epsilon_\beta)-h^{(2)}(\beta_0,\epsilon_0)|=O(n^{-1/2})$ by Condition 1. Then let $t'=t-(\beta-\beta_0)x$, term A_2 can be rewritten as

$$\begin{aligned}
A_2 &= -n^{1/2}P\{[H_n^{(2)}(\beta,\epsilon_\beta)-h^{(2)}(\beta,\epsilon_\beta)]\Delta\}+n^{1/2}P\{[H_n^{(2)}(\beta_0,\epsilon_0)-h^{(2)}(\beta_0,\epsilon_0)]\Delta\}\\
&\quad -n^{1/2}P\{[h^{(2)}(\beta,\epsilon_\beta)-h^{(2)}(\beta_0,\epsilon_0)]\Delta\}\\
&= -n^{1/2}\int\{[H_n^{(2)}(\beta,t')-h^{(2)}(\beta,t')]\Delta\}\;dP_{\epsilon_0,\Delta,X}(t,\delta,x)\\
&\quad +n^{1/2}\int\{[H_n^{(2)}(\beta,t)-h^{(2)}(\beta,t)]\Delta\}\;dP_{\epsilon_0,\Delta,X}(t,\delta,x)\\
&\quad -n^{1/2}P\{[h^{(2)}(\beta,\epsilon_\beta)-h^{(2)}(\beta_0,\epsilon_0)]\Delta\}\\
&= -\int\mathbb{G}_n\{1(\epsilon_\beta\geq t')X\}\;dP_{\epsilon_0,\Delta,X}(t,1,x)+\int\mathbb{G}_n\{1(\epsilon_0\geq t)X\}\;dP_{\epsilon_0,\Delta,X}(t,1,x)\\
&\quad -n^{1/2}(\beta-\beta_0)P\{\dot{h}_\beta^{(2)}(\tilde{\beta},\epsilon_{\tilde{\beta}})\Delta\}\\
&= -\int\mathbb{G}_n\{[1(\epsilon_\beta\geq t')-1(\epsilon_0\geq t)]X\}\;dP_{\epsilon_0,\Delta,X}(t,1,x)\\
&\quad -n^{1/2}(\beta-\beta_0)P\{\dot{h}_\beta^{(2)}(\tilde{\beta},\epsilon_{\tilde{\beta}})\Delta\},
\end{aligned}$$

where $\tilde{\beta}$ is an intermediate value between β and β_0 and $\dot{h}_\beta^{(2)}$ is the derivative of $h^{(2)}$ with respect to β. Since $\{1(\epsilon_\beta\geq t)X\}$ is a Donsker class and by using the similar argument for I_3 and I_4 in the proof of Theorem 2.2.3, it is easy to show that $\{[1(\epsilon_\beta\geq t')-1(\epsilon_0\geq t)]X\}$ converges to zero in quadratic mean. Therefore, term $\mathbb{G}_n\{[1(\epsilon_\beta\geq t')-1(\epsilon_0\geq t)]X\}=o_p(1)$. Moreover, since $\dot{h}_\beta^{(2)}(\beta,\epsilon_\beta)$ is continuous in β, and $|\tilde{\beta}-\beta_0|=O(n^{-1/2})$, we have

$$-n^{1/2}(\beta-\beta_0)P\{\dot{h}_\beta^{(2)}(\tilde{\beta},\epsilon_{\tilde{\beta}})\Delta\}=-n^{1/2}(\beta-\beta_0)P\{\dot{h}_\beta^{(2)}(\beta_0,\epsilon_0)\Delta+o_p(1)\}.$$

Therefore, $A_2=-n^{1/2}(\beta-\beta_0)P\{\dot{h}_\beta^{(2)}(\beta_0,\epsilon_0)\Delta\}+o_p(1)$. Finally, by the similar argument for A_1, it can be shown that term A_3 converges to zero in probability. And by using the same argument for A_2, term A_4 can be re-written as

$$A_4 = n^{1/2}(\beta - \beta_0)P\{\dot{h}_\beta^{(1)}(\beta_0, \epsilon_0)X\Delta\} + o_p(1).$$

Therefore,

$$\begin{aligned}
&n^{\frac{1}{2}}\big\{\Psi_n(\beta, H_n^{(1)}(\beta, \epsilon_\beta), H_n^{(2)}(\beta, \epsilon_\beta)) - \Psi_n(\beta_0, H_n^{(1)}(\beta_0, \epsilon_0), H_n^{(2)}(\beta_0, \epsilon_0))\big\} \\
&\quad = -n^{1/2}(\beta - \beta_0)P\{\dot{h}_\beta^{(2)}(\beta_0, \epsilon_0)\Delta\} + n^{1/2}(\beta - \beta_0)P\{\dot{h}_\beta^{(1)}(\beta_0, \epsilon_0)X\Delta\} + o_p(1) \\
&\quad = n^{1/2}(\beta - \beta_0)\dot{\Psi}_\beta(\beta_0, h^{(1)}(\beta_0, \cdot), h^{(2)}(\beta_0, \cdot)) + o_p(1).
\end{aligned}$$

Replacing β by $\hat{\beta}_n$ and using the assumption that $n^{1/2}\Psi_n(\hat{\beta}_n, H_n^{(1)}(\hat{\beta}_n, \cdot), H_n^{(2)}(\hat{\beta}_n, \cdot)) = o_p(1)$ will yield the asymptotic linearity of Ψ_n, which further yields:

$$\begin{aligned}
n^{1/2}(\hat{\beta}_n - \beta_0) \;&=\; \big\{-\dot{\Psi}_\beta(\beta_0, h^{(1)}(\beta_0, \cdot), h^{(2)}(\beta_0, \cdot))\big\}^{-1} \\
&\qquad \cdot n^{1/2}\Psi_n(\beta_0, H_n^{(1)}(\beta_0, \cdot), H_n^{(2)}(\beta_0, \cdot)) + o_p(1).
\end{aligned}$$

Direct calculation with the fact that $P\{[h^{(1)}(\beta_0, \cdot)X - h^{(2)}(\beta_0, \cdot)]\Delta\} = 0$ gives

$$\begin{aligned}
&n^{1/2}\Psi_n(\beta_0, H_n^{(1)}(\beta_0, \cdot), H_n^{(2)}(\beta_0, \cdot)) = n^{1/2}\mathbb{P}_n\big\{[H_n^{(1)}(\beta_0, \cdot)X - H_n^{(2)}(\beta_0, \cdot)]\Delta\big\} \\
&\quad = \mathbb{G}_n\big\{[H_n^{(1)}(\beta_0, \cdot) - h^{(1)}(\beta_0, \cdot)]X\Delta\big\} - \mathbb{G}_n\{[H_n^{(2)}(\beta_0, \cdot) - h^{(2)}(\beta_0, \cdot)]\Delta\big\} \\
&\qquad + n^{1/2}P\big\{[H_n^{(1)}(\beta_0, \cdot) - h^{(1)}(\beta_0, \cdot)]X\Delta\big\} - n^{1/2}P\big\{[H_n^{(2)}(\beta_0, \cdot) - h^{(2)}(\beta_0, \cdot)]\Delta\big\} \\
&\qquad + \mathbb{G}_n\big\{[h^{(1)}(\beta_0, \cdot)X - h^{(2)}(\beta_0, \cdot)]\Delta\big\} \\
&\quad = A_5 + A_6 + A_7 + A_8 + \mathbb{G}_n\big\{[h^{(1)}(\beta_0, \cdot)X - h^{(2)}(\beta_0, \cdot)]\Delta\big\}.
\end{aligned}$$

For term A_5, since both $\{H_n^{(1)}(\beta_0, \cdot)X\Delta\}$ and $\{h^{(1)}(\beta_0, \cdot)X\Delta\}$ belong to Donsker classes and $\{[H_n^{(1)}(\beta_0, \epsilon_0) - h^{(1)}(\beta_0, \epsilon_0)]X\Delta\}$ converges to zero in quadratic mean through the following argument:

$$\begin{aligned}
&P\big\{[H_n^{(1)}(\beta_0, \epsilon_0) - h^{(1)}(\beta_0, \epsilon_0)]^2X^2\Delta^2\big\} \le \big\|[H_n^{(1)}(\beta_0, t) - h^{(1)}(\beta_0, t)]^2\big\|P\{X^2\Delta^2\} \\
&\quad \le \big\|(\mathbb{P}_n - P)[1(\epsilon_0 \ge t)]\big\|P\{X^2\} = o(1),
\end{aligned}$$

where the second inequality holds since both $H_n^{(1)}$ and $h^{(1)}$ are between 0 and 1, it follows that

$$[H_n^{(1)}(\beta_0,t)-h^{(1)}(\beta_0,t)]^2 \le |H_n^{(1)}(\beta_0,t)-h^{(1)}(\beta_0,t)| \le \|(\mathbb{P}_n-P)[1(\epsilon_0\ge t)]\|,$$

and the last equality holds because X has a finite second moment and $\{1(\epsilon_0 \ge t)\}$ is a Donsker class and thus a Glivenko-Cantelli class. Hence by corollary 2.3.12 of van der Vaart and Wellner (1996), term $A_5 = o_p(1)$. By the same argument for term A_5, we also have $A_6 = o_p(1)$. Finally, for terms A_7 and A_8, they can be re-written as

$$\begin{aligned} A_7 &= \int \mathbb{G}_n\{1(\epsilon_0 \ge t)\}x \; dP_{\epsilon_0,\Delta,X}(t,1,x), \\ A_8 &= -\int \mathbb{G}_n\{1(\epsilon_0\ge t)X\} \; dP_{\epsilon_0,\Delta}(t,1). \end{aligned}$$

Therefore,

$$\begin{aligned} & n^{1/2}\Psi_n(\beta_0, H_n^{(1)}(\beta_0,\cdot), H_n^{(2)}(\beta_0,\cdot)) \\ &\quad = \int \mathbb{G}_n\{1(\epsilon_0\ge t)\}x \; dP_{\epsilon_0,\Delta,X}(t,1,x) - \int \mathbb{G}_n\{1(\epsilon_0 \ge t)X\} \; dP_{\epsilon_0,\Delta}(t,1) \\ &\qquad + \mathbb{G}_n\big\{[h^{(1)}(\beta_0,\cdot)X - h^{(2)}(\beta_0,\cdot)]\Delta\big\} + o_p(1). \end{aligned}$$

Thus, we have proved the asymptotic normality of $\hat\beta_n$ with the following representation form

$$\begin{aligned} & n^{1/2}(\hat\beta_n - \beta_0) \\ &\quad = \big\{-\dot\Psi_\beta(\beta_0, h^{(1)}(\beta_0,\cdot), h^{(2)}(\beta_0,\cdot))\big\}^{-1}\mathbb{G}_n\Big\{[h^{(1)}(\beta_0,\cdot)X - h^{(2)}(\beta_0,\cdot)]\Delta \\ &\qquad - \int [1(\epsilon_0\ge t)X] \; dP_{\epsilon_0,\Delta}(t,1) + \int [1(\epsilon_0 \ge t)]x \; dP_{\epsilon_0,\Delta,X}(t,1,x)\Big\} \\ &\qquad + o_p(1). \end{aligned}$$

Table 2.1: Summary of the simulation statistics. The slope estimator is obtained by solving the Gehan-weighted rank based estimating equation and the intercept estimator is obtained by (2.5). The true parameters are $\alpha = 2$ and $\beta = 1$. The empirical mean (standard deviation) for each of the two parameters is provided. (a) $\zeta \sim N(0, 0.5^2)$; (b) $\zeta \sim \mathit{Gumbel}(-0.5\gamma, 0.5)$; (c) $\zeta \sim \mathit{Laplace}(0, 0.5)$; (d) $\zeta \sim \mathit{Logistic}(0, 0.5)$; and (e) $\zeta \sim T(0, \mathit{df} = 30)$. †: $\tau = 1$ and ‡: $\tau = 3$.

Err.	Cen.	n = 50		n = 200		n = 500		n = 2000	
dist	rate	α	β	α	β	α	β	α	β
$X \sim N(0,1)$:									
(a)	.84†	—	—	1.99 (.18)	1.00 (.12)	2.00 (.11)	1.00 (.08)	2.00 (.06)	1.00 (.04)
	.52‡	2.00 (.11)	1.01(.11)	2.00 (.05)	1.00 (.05)	2.00 (.03)	1.00 (.03)	2.00 (.02)	1.00 (.02)
(b)	.82†	—	—	1.95 (.16)	1.00 (.10)	1.97 (.10)	1.00 (.07)	1.99 (.05)	1.00 (.03)
	.52‡	1.99 (.14)	1.00 (.12)	1.99 (.07)	1.00 (.05)	2.00 (.04)	1.00 (.03)	2.00 (.02)	1.00 (.02)
(c)	.82†	—	—	1.98 (.24)	1.00 (.18)	1.98 (.15)	1.01 (.12)	1.99 (.08)	1.00 (.06)
	.52‡	1.99 (.14)	1.01 (.14)	2.00 (.07)	1.00 (.07)	2.00 (.04)	1.00 (.04)	2.00 (.02)	1.00 (.02)
(d)	.80†	—	—	1.96 (.23)	1.01 (.16)	1.96 (.14)	1.00 (.10)	1.98 (.07)	1.00 (.05)
	.52‡	1.98 (.18)	1.01 (.19)	2.00 (.09)	1.01 (.09)	2.00 (.06)	1.00 (.06)	2.00 (.03)	1.00 (.03)
(e)	.78†	—	—	1.94 (.22)	1.00 (.16)	1.96 (.14)	1.01 (.10)	1.97 (.07)	1.00 (.05)
	.52‡	1.99 (.21)	1.02 (.21)	1.99 (.10)	1.00 (.10)	1.99 (.06)	1.00 (.06)	2.00 (.03)	1.00 (.03)
$X \sim U(-2,2)$:									
(a)	.78†	—	—	2.00 (.19)	1.01 (.14)	2.00 (.11)	1.00 (.08)	2.00 (.06)	1.00 (.04)
	.52‡	2.00 (.12)	1.00 (.11)	2.00 (.06)	1.00 (.05)	2.00 (.04)	1.00 (.03)	2.00 (.02)	1.00 (.02)
(b)	.77†	—	—	1.95 (.17)	1.00 (.12)	1.96 (.10)	1.00 (.07)	1.96 (.05)	1.00 (.03)
	.52‡	1.99 (.14)	1.01 (.10)	2.00 (.07)	1.00 (.05)	2.00 (.04)	1.00 (.03)	2.00 (.02)	1.00 (.01)
(c)	.77†	—	—	1.99 (.25)	1.02 (.18)	1.98 (.15)	1.01 (.11)	1.97 (.07)	1.00 (.06)
	.53‡	2.00 (.16)	1.01 (.14)	2.00 (.08)	1.00 (.07)	2.00 (.05)	1.00 (.04)	2.00 (.02)	1.00 (.02)
(d)	.76†	—	—	1.95 (.23)	1.02 (.17)	1.94 (.14)	1.01 (.10)	1.94 (.07)	1.00 (.05)
	.52‡	2.00 (.19)	1.01 (.16)	1.99 (.09)	1.00 (.08)	2.00 (.06)	1.00 (.05)	2.00 (.03)	1.00 (.03)
(e)	.75†	—	—	1.91 (.21)	1.02 (.16)	1.91 (.13)	1.01 (.10)	1.91 (.06)	1.00 (.05)
	.52‡	2.00 (.21)	1.00 (.19)	2.00 (.11)	1.00 (.09)	2.00 (.07)	1.00 (.06)	2.00 (.03)	1.00 (.03)
$X \sim U(-1,1)$:									
(a)	.92†	—	—	1.85 (.32)	1.08 (.37)	1.81 (.16)	1.02 (.19)	1.80 (.08)	1.00 (.09)
	.52‡	2.00 (.10)	1.01 (.18)	2.00 (.05)	1.00 (.08)	2.00 (.03)	1.00 (.05)	2.00 (.02)	1.00 (.03)
(b)	.90†	—	—	1.77 (.20)	1.03 (.25)	1.76 (.13)	1.01 (.15)	1.76 (.06)	1.01 (.07)
	.51‡	1.98 (.12)	1.00 (.18)	1.99 (.06)	1.00 (.08)	1.99 (.04)	1.00 (.05)	1.99 (.02)	1.00 (.03)
(c)	.89†	—	—	1.79 (.33)	1.06 (.40)	1.77 (.19)	1.03 (.23)	1.75 (.09)	1.00 (.12)
	.52‡	1.98 (.13)	1.00 (.22)	1.99 (.06)	1.00 (.10)	1.99 (.04)	1.00 (.06)	1.99 (.02)	1.00 (.03)
(d)	.85†	—	—	1.67 (.23)	1.03 (.31)	1.66 (.14)	1.02 (.19)	1.66 (.07)	1.00 (.09)
	.52‡	1.96 (.17)	1.01 (.30)	1.98 (.08)	1.00 (.13)	1.99 (.05)	1.00 (.09)	1.99 (.03)	1.00 (.04)
(e)	.82†	—	—	1.59 (.19)	1.02 (.27)	1.59 (.12)	1.01 (.16)	1.59 (.06)	1.00 (.08)
	.52‡	1.97 (.18)	1.01 (.33)	1.98 (.10)	1.00 (.16)	1.98 (.06)	1.00 (.10)	1.99 (.03)	1.00 (.05)

Table 2.2: Comparison of prediction accuracy between the semiparametric linear model and the Cox model. Relative prediction accuracy to the uncensored case, i.e., the ratio of the empirical mean MSE_p under the uncensored case to that under each corresponding censored case, is listed. The empirical mean ± standard deviation of MSE_p under each scenario is given in the parenthesis. The MSE_p obtained from ordinary least squares (OLS) is also listed for each uncensored scenario.

		Cen.	Sample Size			
X	τ	rate	n = 200		n = 2000	
			Linear	Cox	Linear	Cox
Identity Transformation:						
$N(0,1)$	-2	.86	0.86 (1.95 ± 0.51)	0.33 (5.08 ± 0.41)	0.98 (1.68 ± 0.08)	0.33 (4.98 ± 0.12)
	-1	.72	0.97 (1.72 ± 0.29)	0.58 (2.90 ± 0.29)	0.99 (1.66 ± 0.08)	0.58 (2.86 ± 0.09)
	0	.55	0.99 (1.69 ± 0.25)	0.84 (2.00 ± 0.24)	1.00 (1.65 ± 0.08)	0.84 (1.96 ± 0.08)
	1	.44	1.00 (1.67 ± 0.24)	0.97 (1.72 ± 0.23)	1.00 (1.64 ± 0.08)	0.97 (1.70 ± 0.08)
	–	.00	1.67 ± 0.25	1.67 ± 0.25	1.65 ± 0.08	1.65 ± 0.08
	OLS		1.67 ± 0.25		1.65 ± 0.08	
$U(-2,2)$	-2	.82	0.85 (1.93 ± 0.48)	0.31 (5.38 ± 0.44)	0.96 (1.71 ± 0.09)	0.31 (5.28 ± 0.12)
	-1	.67	0.96 (1.71 ± 0.28)	0.53 (3.12 ± 0.32)	1.00 (1.65 ± 0.08)	0.54 (3.08 ± 0.09)
	0	.54	0.99 (1.67 ± 0.26)	0.80 (2.07 ± 0.27)	1.00 (1.65 ± 0.08)	0.80 (2.05 ± 0.08)
	1	.46	0.99 (1.66 ± 0.25)	0.96 (1.72 ± 0.26)	1.00 (1.65 ± 0.08)	0.96 (1.71 ± 0.08)
	–	.00	1.65 ± 0.25	1.65 ± 0.25	1.65 ± 0.08	1.65 ± 0.08
	OLS		1.65 ± 0.25		1.65 ± 0.08	
$U(-1,1)$	-2	.86	0.68 (2.41 ± 0.58)	0.37 (4.51 ± 0.36)	0.74 (2.24 ± 0.18)	0.38 (4.38 ± 0.09)
	-1	.72	0.94 (1.75 ± 0.28)	0.67 (2.47 ± 0.28)	0.97 (1.70 ± 0.08)	0.67 (2.45 ± 0.08)
	0	.55	0.99 (1.66 ± 0.26)	0.93 (1.78 ± 0.26)	1.00 (1.65 ± 0.08)	0.93 (1.77 ± 0.08)
	1	.44	1.00 (1.65 ± 0.25)	1.00 (1.65 ± 0.25)	1.00 (1.65 ± 0.08)	1.00 (1.65 ± 0.08)
	–	.00	1.65 ± 0.25	1.65 ± 0.25	1.65 ± 0.08	1.65 ± 0.08
	OLS		1.65 ± 0.25		1.65 ± 0.08	
Logarithm Transformation:						
$N(0,1)$	-2	.83	0.74 (12.09 ± 12.76)	0.56 (14.50 ± 12.52)	0.96 (9.02 ± 3.99)	0.51 (14.36 ± 4.79)
	-1	.69	0.88 (10.25 ± 10.55)	0.59 (13.91 ± 12.44)	0.98 (8.82 ± 3.69)	0.54 (13.76 ± 4.77)
	0	.55	0.98 (9.16 ± 8.88)	0.64 (12.75 ± 12.24)	0.98 (8.82 ± 3.68)	0.59 (12.58 ± 4.71)
	1	.45	1.00 (8.99 ± 8.73)	0.74 (11.05 ± 11.79)	0.98 (8.82 ± 3.66)	0.68 (10.81 ± 4.57)
	–	.00	9.00 ± 8.60	8.19 ± 8.16	8.67 ± 3.25	7.39 ± 2.73
	OLS		8.99 ± 8.51		8.68 ± 3.26	
$U(-2,2)$	-2	.82	0.83 (9.64 ± 5.12)	0.53 (12.99 ± 4.65)	0.92 (8.85 ± 1.32)	0.52 (13.19 ± 1.41)
	-1	.67	0.95 (8.44 ± 4.11)	0.56 (12.33 ± 4.56)	0.99 (8.21 ± 1.15)	0.55 (12.52 ± 1.39)
	0	.54	0.99 (8.07 ± 3.58)	0.62 (11.00 ± 4.36)	1.00 (8.15 ± 1.11)	0.61 (11.12 ± 1.32)
	1	.46	1.00 (8.04 ± 3.58)	0.76 (9.00 ± 3.95)	1.00 (8.15 ± 1.09)	0.76 (9.00 ± 1.19)
	–	.00	8.03 ± 3.59	6.86 ± 2.88	8.14 ± 1.08	6.83 ± 0.87
	OLS		8.03 ± 3.59		8.14 ± 1.08	
$U(-1,1)$	-2	.86	0.77 (2.81 ± 1.75)	0.55 (3.33 ± 0.87)	0.78 (2.79 ± 0.27)	0.54 (3.35 ± 0.26)
	-1	.72	0.91 (2.35 ± 0.76)	0.62 (2.96 ± 0.83)	0.92 (2.36 ± 0.23)	0.61 (2.98 ± 0.25)
	0	.55	0.99 (2.18 ± 0.71)	0.76 (2.40 ± 0.75)	0.99 (2.18 ± 0.21)	0.76 (2.40 ± 0.22)
	1	.44	1.00 (2.15 ± 0.70)	0.94 (1.94 ± 0.65)	1.00 (2.17 ± 0.21)	0.94 (1.93 ± 0.19)
	–	.00	2.15 ± 0.69	1.83 ± 0.55	2.17 ± 0.20	1.82 ± 0.16
	OLS		2.15 ± 0.69		2.17 ± 0.20	

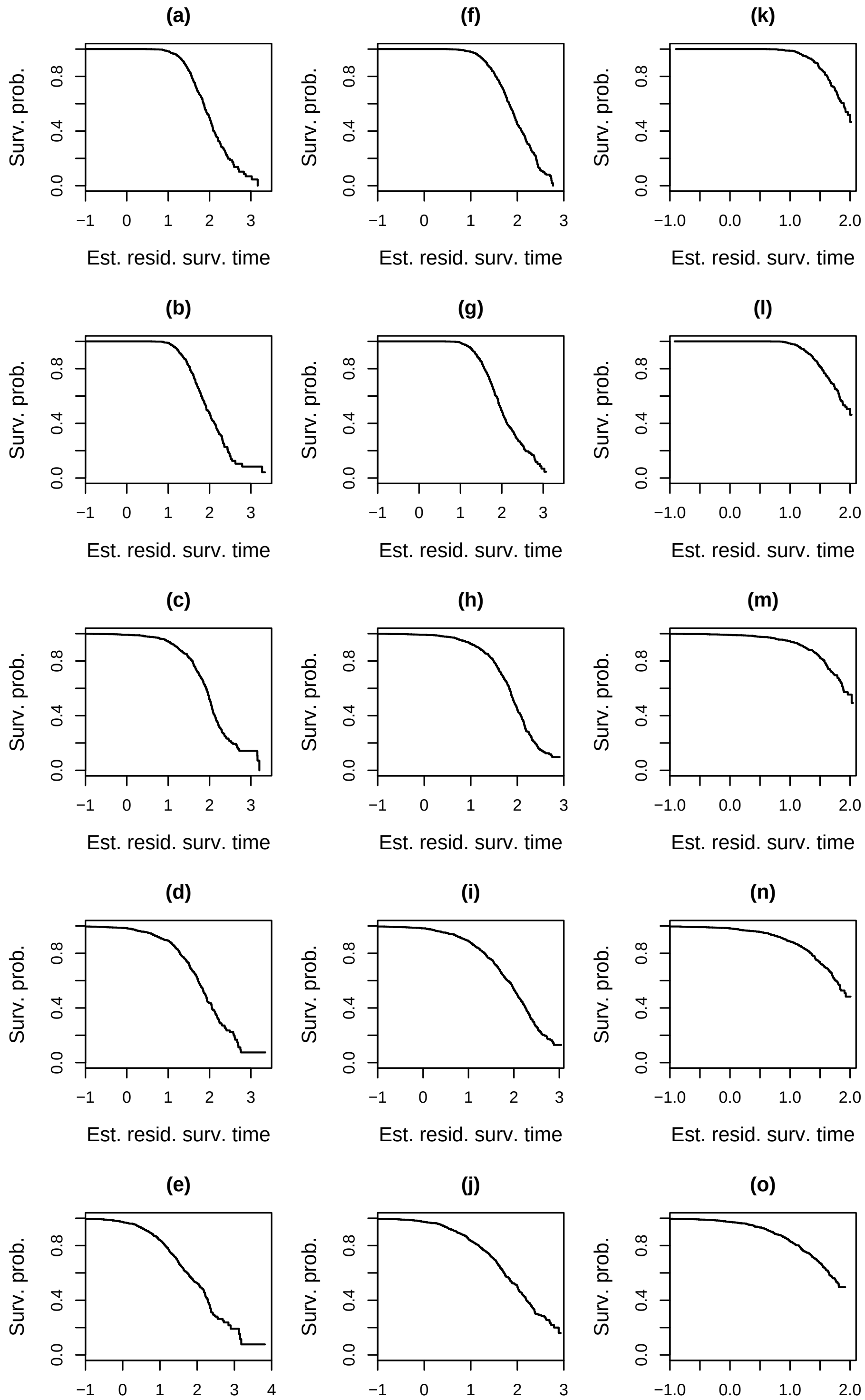

Figure 2.1: Kaplan-Meier curves of the estimated residual survival time $(T - \hat{\beta}_n X)$ under $\tau = 1$. (a)-(e): $X \sim N(0,1)$ with $\zeta \sim N(0, 0.5^2)$, $\zeta \sim \mathit{Gumbel}(-0.5\gamma, 0.5)$, $\zeta \sim \mathit{Laplace}(0, 0.5)$, $\zeta \sim \mathit{Logistic}(0, 0.5)$ and $\zeta \sim T(0, \mathit{df} = 30)$, respectively; (f)-(j): $X \sim U(-2, 2)$ with the same corresponding error distributions under (a)-(e); (k)-(o): $X \sim U(-1, 1)$ with the same corresponding error distributions under (a)-(e).

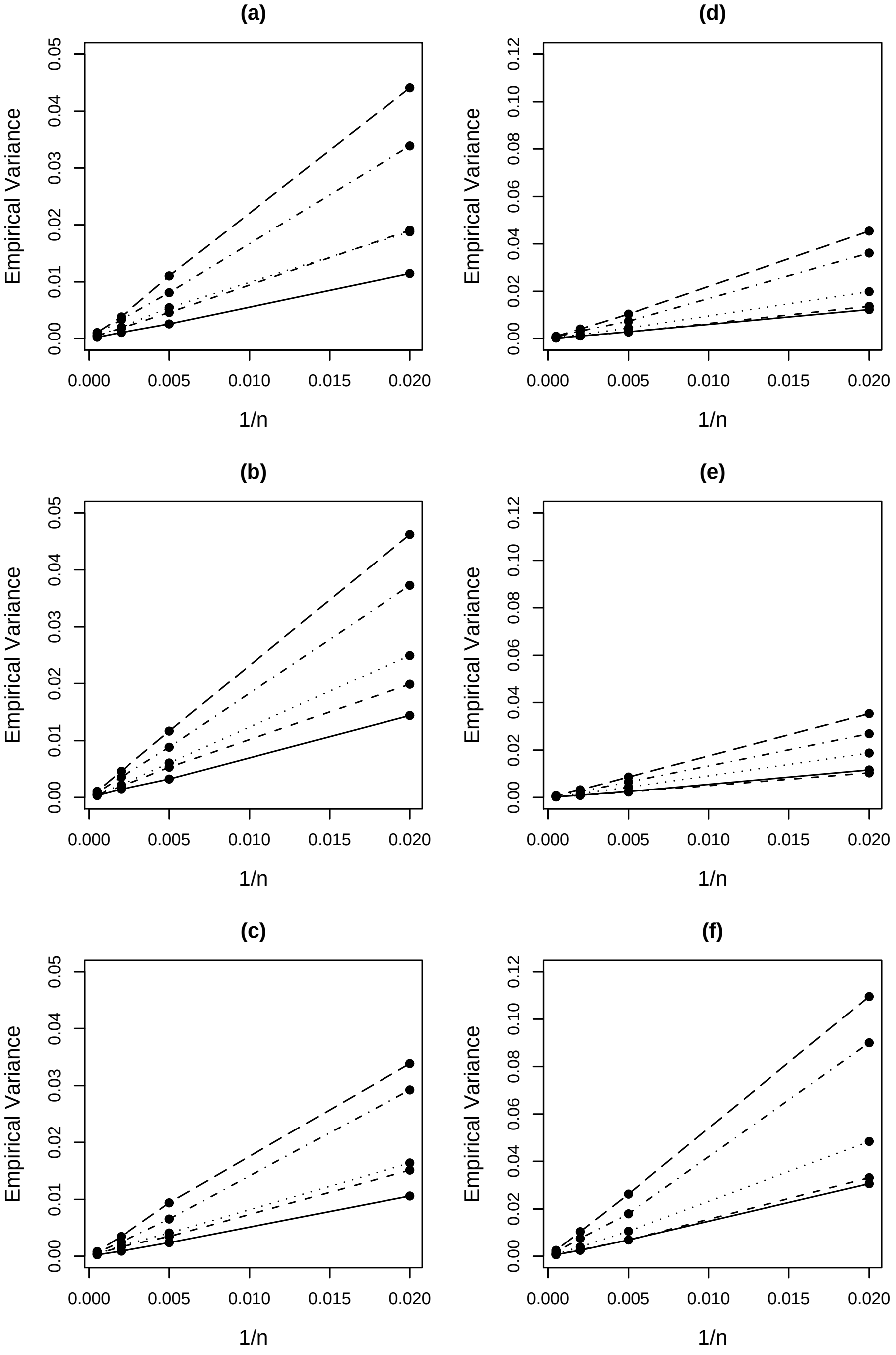

Figure 2.2: Empirical variances of the intercept and slope estimators with $\tau = 3$. (a)-(c): the intercept estimators under $X \sim N(0,1)$, $X \sim U(-2,2)$ and $X \sim U(-1,1)$, respectively; (e)-(f): the slope estimators under $X \sim N(0,1)$, $X \sim U(-2,2)$ and $X \sim U(-1,1)$, respectively. Solid line: $\zeta \sim N(0, 0.5^2)$; dashed line: $\zeta \sim Gumbel(-0.5\gamma, 0.5)$; dotted line: $\zeta \sim Laplace(0, 0.5)$; dotted dash line: $\zeta \sim Logistic(0, 0.5)$; and long dashed line: $\zeta \sim T(0, df = 30)$.

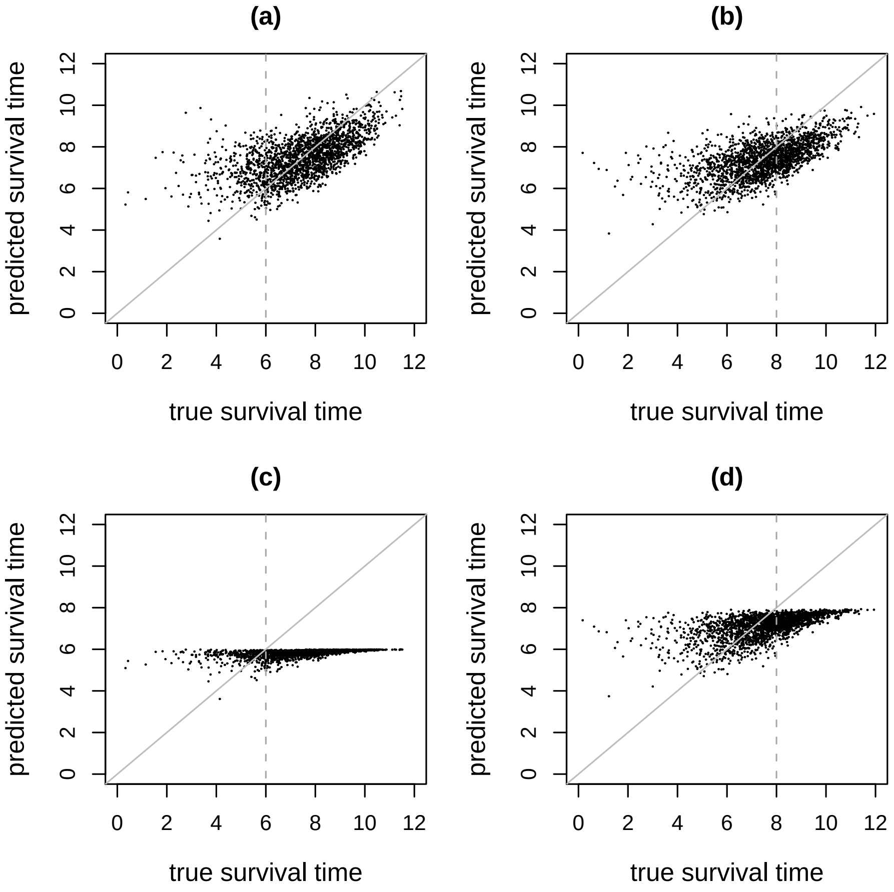

Figure 2.3: The predicted survival time versus the true survival time for the data generated from model (2.23) with $X \sim N(0, 1)$ under the identity transformation. (a): semiparametric linear model with $\tau = -2$; (b): semiparametric linear model with $\tau = 0$; (c): Cox model with $\tau = -2$; and (d): Cox model with $\tau = 0$. A constant 8 was added to shift all the simulated survival times to positive values.

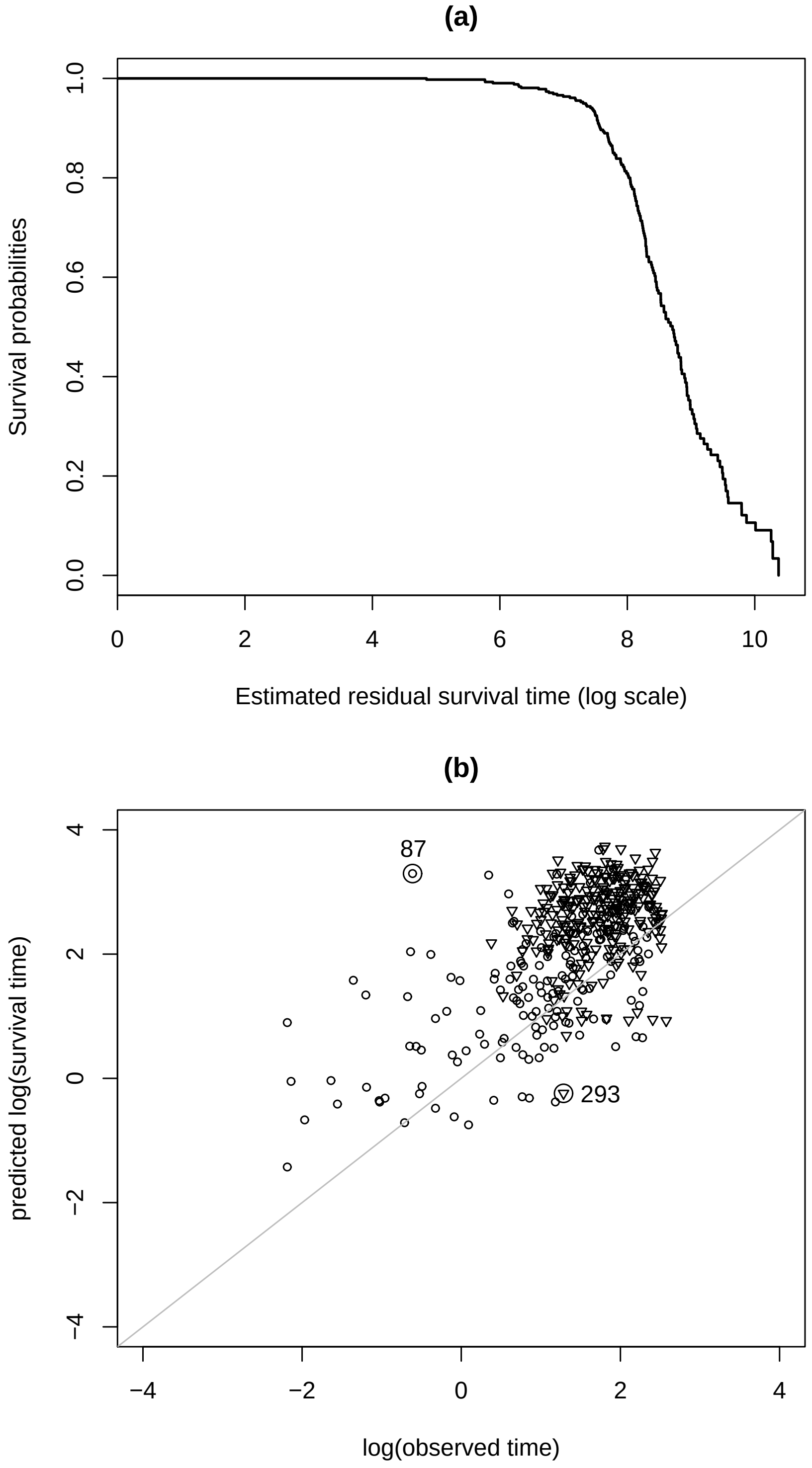

Figure 2.4: (a): Kaplan-Meier curve of the estimated residual survival time for the PBC data. (b): Predicted survival time versus the observed time points (both in the logarithm scale) for the PBC data by fitting the accelerated failure time model via leave-one-out cross-validation. The circles correspond to the individuals who failed and the triangles correspond to the individuals who were censored. Subject 87 and 293 are two potential outliers.

CHAPTER III

Sieve Maximum Likelihood Estimation Using B-Spline Smoothing for the Slope Estimators in the AFT Model

3.1 Introduction

The transformed linear regression model (Kalbfleisch and Prentice 2002) relates the failure time under a monotone transformation to the covariates directly, it provides a straightforward interpretation of the data and serves as an attractive alternative to the Cox proportional hazards model in many applications.

The challenge in analyzing the semiparametric linear model comes from the presence of censoring in failure time data. Several estimators of the slope parameters have been proposed in the literature since late 70's. Prentice (1978) proposed the rank estimators based on the well-known weighted log-rank statistics. Another major estimating method is based on a modified least-squares estimator to accommodate censoring, namely, the Buckley-James estimator (Buckley and James 1979). Ritov (1990), Tsiatis (1990), Lai and Ying (1991) and Ying (1993), among others, studied the asymptotic properties of the rank-based and Buckley-James estimators.

In spite of appealing theoretical properties, estimators from the above estimating methods do not appear to be widely used in applications. One major reason is that these semiparametric estimating functions are discrete and can be non-monotone, which leads to potential multiple solutions and is difficult to solve numerically. Recently, Jin et al. (2003) and Jin et al. (2006) proposed new approaches to approximate the rank-based estimator and least-squares estimators through linear programming, together with the re-sampling procedures to estimate the limiting covariance matrices. However, none of the estimators from the above estimating methods are fully semiparametric efficient. It is also known that the linear programming with re-sampling technique is computational intensive and infeasible for large samples. Recently, Zeng and Lin (2007) developed an efficient estimator for the AFT model by maximizing a kernel-smoothed profile likelihood function for the regression parameters. Their approach, however, is restricted to the log-transformed linear model so far and may not be very flexible to accommodate more general monotone transformations. Furthermore, their kernel smoothing procedure is neither intuitive nor straightforward and thus difficult to implement in practice.

Two decades ago, Ritov and Wellner (1988) derived the semiparametric efficient score functions for the slope parameters in the linear regression model, which involve the derivative of the density function (or the hazard function) of the error term. We propose a new approach by directly maximizing the log likelihood function in a sieve space in which the log hazard function of the error term is approximated by B-splines. Numerically, the estimator can be easily obtained by the Newton-Raphson algorithm or any gradient-based search algorithms. We show that the proposed estimator is consistent and asymptotically normal, and the limiting covariance matrix reaches the semiparametric efficiency bound, which can be estimated either by inverting the information matrix based on the efficient score function of the slope parameters, or by inverting the observed information matrix of all parameters, taking into account that we are also estimating the “nuisance” parameters in the sieve space for the log hazard function.

The organization of this chapter is as follows. Section 3.2 describes the new proposed estimation approach. The regularity conditions and asymptotic properties of the proposed estimators are provided in Section 3.3. In Section 3.4 we conduct extensive simulation studies for different covariate and error distributions, censoring rates and the sample sizes. In Section 3.5 we provide applications to two major medical studies. In Section 3.6, we introduce an extended general theorem on the asymptotic normality of semiparametric M-estimators, where the nuisance parameters can be a function of the parameters of interest. The general theorem is crucial in the proof of asymptotic normality and semiparametric efficiency of the proposed estimators given in Theorem 3.3.2. We provide some discussion in Section 3.7. Technical details are provided in the Appendix.

3.2 Estimating Methods in Linear Models with Censored Data

3.2.1 Accelerated Failure Time Model and Existing Estimators

Suppose that the failure time transformed by a fixed monotone transformation $h(\cdot)$ is linearly related to a set of covariates, where the failure time is subject to right censoring. Let T_i denote the transformed failure time and C_i denote the transformed censoring time by the same transformation for subject i, $i = 1, \cdots, n$. Let $Y_i = \min(T_i, C_i)$ and $\Delta_i = I(T_i \leq C_i)$. Then the semiparametric linear model we consider here can be written as

$$T_i = X_i'\beta_0 + e_{0,i}, \quad i = 1, \cdots, n, \tag{3.1}$$

where the errors $e_{0,i}$ are independent and identically distributed (not necessarily with mean zero) with an unspecified distribution. When $h(\cdot) = \log(\cdot)$, this model corresponds to the well-known accelerated failure time model. Here we assume that (X_i, C_i), $i = 1, \ldots, n$, are i.i.d. and independent of $e_{0,i}$. This is a common assumption and yet a reasonable one in practice for linear models with censored survival data.

As mentioned in Section 3.1, there are two major classes of existing methods for estimating the parameter β_0 in model (3.1). One class is based on the Buckley-James estimator (Buckley and James 1979; Ritov 1990; Lai and Ying 1991), which solves the following estimating equation for β:

$$\sum_{i=1}^{n}(X_i - \bar{X})\left\{\Delta_i\ s(Y_i - X_i'\beta) + (1-\Delta_i)\frac{1}{1-\hat{F}(Y_i - X_i'\beta)}\int_{Y_i - X_i'\beta}^{\infty} s(t)\ d\hat{F}(t)\right\} = 0, \tag{3.2}$$

where $\hat{F}$ is the Kaplan-Meier estimator (or a modified Kaplan-Meier estimator) of F, the distribution function of $e_0 = T - X'\beta_0$, and s is some function in $L_2^0(F)$. When s is the identity function, the above equation reduces to the classical estimating equation of Buckley and James (1979). Equation (3.2) can be solved iteratively.

Another class is the weighted rank based method (Prentice 1978; Wei et al. 1990; Tsiatis 1990; Ying 1993), which solves the following estimating equation for β:

$$\sum_{i=1}^{n}\int\left\{X_i - \frac{\sum_{j=1}^{n} X_j I(Y_j - X_j'\beta \geq t)}{\sum_{j=1}^{n} I(Y_j - X_j'\beta \geq t)}\right\}\omega(t)\ dN_i(t) = 0, \tag{3.3}$$

where $N_i(t) = \Delta_i I(Y_i - X_i'\beta \leq t)$ is the counting process for subject i and $w(t)$ is a weight function that may also depend on β. Note that both equations (3.2) and (3.3) are discrete. Ritov (1990) showed that these two classes of estimating equations are asymptotically equivalent and that when $s(t) = -\dot{f}(t)/f(t)$, here f is the density function of $e_0 = T - X'\beta_0$ and $\dot{f}$ is the derivative of f, the estimator obtained from equation (3.2) is the fully semiparametric efficient estimator. Ritov and Wellner (1988) showed that when $\omega(t) = \dot{\lambda}_0(t)/\lambda_0(t)$, here λ_0 is the hazard function of the error term e_0 and $\dot{\lambda}_0$ is the derivative of λ_0, estimating equation (3.3) yields the most efficient estimator for β_0. Obviously, efficient estimation from either method needs to estimate the derivative of the density (or the hazard) function of the error term, and developing a nonparametric approach for its estimation in addition to solving the discrete estimating equation does not appear attractive to practitioners.

The non-smoothness of these estimating equations complicates the proofs of asymptotic properties of the estimators and makes numerical implementations less straightforward. Ritov (1990), Tsiatis (1990) and Ying (1993) derived asymptotic normality of the estimators obtained from either equation (3.2) or equation (3.3) under similar regularity conditions by first proving the asymptotic linearity of corresponding estimating equation in a neighborhood of β_0. Several numerical methods have been proposed, which include the simulated annealing approach of Lin and Geyer (1992), the linear programming approach of Lin et al. (1998) and Jin et al. (2003), and most recently, the hybrid Newton-type method of Yu and Nan (2006). But none of them is as nice as the Newton-Raphson method for solving smooth equations.

3.2.2 Sieve Maximum Likelihood Estimation Using B-Spline Smoothing

Instead of solving the discrete efficient estimating equation, a likelihood based approach with certain nonparametric estimation of the log hazard function of the error term using the smoothing technique seems more desirable, since such a procedure maximizes a smooth function of unknown parameters. This falls into the sieve maximum likelihood estimation procedure of Geman and Hwang (1982) where the unknown function in the log likelihood is approximated by a linear span of some known basis functions to form a sieve log likelihood. Then we just need to maximize the sieve log likelihood with respect to the unknown coefficients in the linear span to obtain a sieve maximum likelihood estimator. This also significantly reduces the dimensionality of the optimization since the number of basis functions needed to reasonably approximate the unknown function grows at a much slower rate as the sample size increases.

Spline technique has been extensively used as an effective tool in dimension reduction for nonparametric estimation. Stone (1985, 1986) has proved in theory that a smooth unknown function can be well approximated using splines. Some further convergence results of spline-based sieve estimates have been developed by Shen and Wong (1994). Therefore, we consider the spline smoothed sieve maximum likelihood estimation for the semiparametric linear model with censored data.

Since we assume that e_0 is independent of (C, X), the joint density function of (T, C, X) can then be decomposed as

$$f_{T,C,X}(t, c, x) = f_{e_0,C,X}(t - x'\beta_0, c, x) = f(t - x'\beta_0) f_{C,X}(c, x).$$

It is easy to see that T and C are independent conditional on X under the assumption $e_0 \perp (C, X)$. Hence we have

$$f(t - x'\beta_0) = f_{T|C,X}(t|C = c, X = x) = f_{T|X}(t|X = x).$$

Then the joint density function of (Y, Δ, X) can be written as

$$\begin{aligned} f_{Y,\Delta,X}(y, \delta, x) &= f(y - x'\beta_0)^{\delta}\{1 - F(y - x'\beta_0)\}^{1-\delta} H(y, \delta, x) \\ &= \lambda_0(y - x'\beta_0)^{\delta} \exp\{-\Lambda_0(y - x'\beta_0)\} H(y, \delta, x), \end{aligned} \tag{3.4}$$

where $\Lambda_0(\cdot)$ is the true cumulative hazard function for the error term e_0. $H(y, \delta, x)$ only depends on the conditional distribution of C given X and the marginal distribution of X, and is free of β_0 and λ_0. Thus to simplify the notation, we will ignore the factor H from the likelihood function. Then for i.i.d. observations (Y_i, Δ_i, X_i), $i = 1, \cdots, n$, from equation (3.4) we obtain the log likelihood function for β and λ as

$$l_n(\beta, \lambda) = n^{-1} \sum_{i=1}^{n} \left\{ \Delta_i \log\{\lambda(Y_i - X_i'\beta)\} - \int_{-\infty}^{Y_i} \lambda(t - X_i'\beta)\, dt \right\}. \tag{3.5}$$

The log likelihood given in (3.5) apparently is a semiparametric model, in which β is the finite dimensional parameter of interest and λ is an unknown positive function and treated as an infinite dimensional nuisance parameter. Let $g(t) = \log \lambda(t)$, then the log likelihood function for β and g, using the counting process notation, can be written as

$$(3.6)\quad l_n(\beta, g) = n^{-1} \sum_{i=1}^{n} \left\{ \int g(t - X_i'\beta)\, dN_i(t) - \int I(Y_i \geq t) \exp\{g(t - X_i'\beta)\}\, dt \right\},$$

where $N_i(t) = \Delta_i I(Y_i - X_i'\beta \leq t)$ is the counting process for subject i.

By taking the logarithm of the positive function $\lambda(\cdot)$, the function $g(\cdot)$ is no longer restricted to be positive, which eases the estimation. We now describe the spline-based sieve maximum likelihood estimation for the AFT model. Suppose a and b are lower and upper bounds of the observed residual time:

$$\{Y_i - X_i'\beta : \beta \in \mathcal{B},\ i = 1, \cdots, n\},$$

where $\mathcal{B}$ is the parameter space of β. Assuming the transformed observation time Y is bounded away from $-\infty$. Under the regularity conditions stated in section 3.3, we have $-\infty < a < b < \infty$. Let $a = t_0 < t_1 < \cdots < t_{K_n} < t_{K_n+1} = b$ be a partition of $[a, b]$ with K subintervals $I_{Kj} = [t_j, t_{j+1}),\ j = 0, \cdots, K-1$ and $I_{KK} = [t_K, t_{K+1}]$, where $K \equiv K_n \simeq n^{\nu}$ with $0 < \nu < 1/2$ being a positive number such that $\max_{1 \leq j \leq K+1} |t_j - t_{j-1}| = O(n^{-\nu})$. Denote the set of partition points by $T_{K_n} = \{t_1, \cdots, t_{K_n}\}$. Let $\mathcal{S}_n(T_{K_n}, K_n, l)$ be the space of polynomial splines of order $l \geq 1$ consisting of function s satisfying (i) the restriction of s to I_{Kj} is a polynomial of order l (or equivalently, of degree $l-1$) for $l \leq K$; (ii) for $l \geq 2$ and $0 \leq l' \leq l-2$, s is l' times continuously differentiable on $[a, b]$. This definition is phrased after Stone (1985), which is a descriptive version of Schumaker (1981, page 108, Definition 4.1).

According to Schumaker (1981, page 117, Corollary 4.10), there exists a local basis $\{B_j, 1 \leq j \leq q_n\}$, so called B-splines, for $\mathcal{S}_n(T_{K_n}, K_n, l)$, where $q_n = K_n + l$. These basis functions are nonnegative and sum up to one at each point in $[a, b]$, and each B_j is zero outside the interval $[t_j, t_{j+l}]$. Thus for any $s \in \mathcal{S}_n(T_{K_n}, K_n, l)$, we can write

$$s(t) = \sum_{j=1}^{q_n} \gamma_j B_j(t). \tag{3.7}$$

Let $\gamma = \{\gamma_j : 1 \leq j \leq q_n\}$ be the collection of all the coefficients in the representation (3.7). Under suitable smoothness assumptions, $g_0(\cdot) = \log \lambda_0(\cdot)$ can be well approximated by some function in $\mathcal{S}_n(T_{K_n}, K_n, l)$. Therefore, we seek a member of $\mathcal{S}_n(T_{K_n}, K_n, l)$ together with a value of $\beta \in \mathcal{B}$ that maximizes the log likelihood function. Specifically, let $\hat{\theta}_n = (\hat{\beta}_n, \hat{\gamma}_n)$ with $\hat{\gamma}_n = \{\hat{\gamma}_{n,j} : 1 \leq j \leq q_n\}$ be the value that maximizes

$$\begin{aligned} l_n(\beta, \gamma) &= n^{-1} \sum_{i=1}^{n} \Bigg[\int \sum_{j=1}^{q_n} \gamma_j B_j(t - X_i'\beta)\, dN_i(t) \\ &\qquad - \int I(Y_i \geq t) \exp\Bigg\{ \sum_{j=1}^{q_n} \gamma_j B_j(t - X_i'\beta) \Bigg\}\, dt \Bigg]. \end{aligned} \tag{3.8}$$

Taking the first derivatives of $l_n(\beta, \gamma)$ with respect to β and γ and setting them to zero, we obtain the following estimating equations:

$$\begin{aligned} \frac{\partial l_n(\beta, \gamma)}{\partial \beta} &= -n^{-1} \sum_{i=1}^{n} \Bigg[\int X_i \sum_{j=1}^{q_n} \gamma_j \dot{B}_j(t - X_i'\beta)\, dN_i(t) \\ &\qquad - \int I(Y_i \geq t) X_i \sum_{j=1}^{q_n} \gamma_j \dot{B}_j(t - X_i'\beta) \exp\Bigg\{ \sum_{j=1}^{q_n} \gamma_j B_j(t - X_i'\beta) \Bigg\}\, dt \Bigg] \\ &= 0, \end{aligned} \tag{3.9}$$

$$\begin{aligned} \frac{\partial l_n(\beta, \gamma)}{\partial \gamma_j} &= n^{-1} \sum_{i=1}^{n} \Bigg[\int B_j(t - X_i'\beta)\, dN_i(t) \\ &\qquad - \int I(Y_i \geq t) B_j(t - X_i'\beta) \exp\Bigg\{ \sum_{j=1}^{q_n} \gamma_j B_j(t - X_i'\beta) \Bigg\}\, dt \Bigg] \\ &= 0, \quad j = 1, \cdots, q_n. \end{aligned} \tag{3.10}$$

Since the integrals in (3.9) and (3.10) are univariate integrals, their numerical implementation can be easily done by one-dimensional Gaussian-quadrature method that would not increase the computing cost much. Newton-Raphson algorithm or any gradient-based search algorithms can be applied to solve the above equations for all parameters $\theta = (\beta, \gamma)$, e.g.,

$$\theta^{(m+1)} = \theta^{(m)} - H(\theta^{(m)})^{-1} \cdot S(\theta^{(m)})$$

where $\theta^{(m)} = (\beta^{(m)}, \gamma^{(m)})$ is the parameter estimate from the mth iteration, and

$$S(\theta) = \begin{pmatrix} \frac{\partial l_n(\beta,\gamma)}{\partial \beta} \\ \frac{\partial l_n(\beta,\gamma)}{\partial \gamma} \end{pmatrix}, \quad H(\theta) = \begin{pmatrix} \frac{\partial^2 l_n(\beta,\gamma)}{\partial \beta \partial \beta'} & \frac{\partial^2 l_n(\beta,\gamma)}{\partial \beta \partial \gamma'} \\ \frac{\partial^2 l_n(\beta,\gamma)}{\partial \gamma \partial \beta'} & \frac{\partial^2 l_n(\beta,\gamma)}{\partial \gamma \partial \gamma'} \end{pmatrix}$$

is the score function and Hessian matrix of the parameter θ, respectively.

In order to make statistical inferences of β_0, we have to approximate the sampling distribution of $\hat{\beta}_n$. As stated in the next section, the distribution of $\hat{\beta}_n$ can be approximated by a normal distribution when the sample size is large enough. One way to estimate the variance matrix of $\hat{\beta}_n$ is to approximate the (inverse of the) information matrix under the efficient score function for β_0, given in Theorem 3.3.2, by plugging in the estimated parameters $(\hat{\beta}_n, \hat{\lambda}_n(\cdot))$. Another way we suggest is to invert the observed information matrix from the last Newton-Raphson iteration, taking into account that we are also estimating the "nuisance" parameter γ. This approach was also suggested by Huang (1999) in the variance estimation of $\hat{\beta}_n$ in the partly linear additive Cox model. As mentioned in Huang (1999), there is no theoretical justification for the second variance estimator so far, but heuristics based on the finite-dimensional parametric model and simulations indicate that this estimator also works well.

3.3 Asymptotic Results

This section states the asymptotic properties of the proposed estimators. Denote $\epsilon_\beta = Y - X'\beta$ and $\epsilon_0 = Y - X'\beta_0$. We assume the following regularity conditions:

(C.1) The true parameter β_0 belongs to the interior of a compact set $\mathcal{B} \subseteq \mathbb{R}^d$.

(C.2) (a) The covariate X takes values in a bounded subset $\mathcal{X} \subseteq \mathbb{R}^d$; (b) $E(XX')$ is nonsingular.

(C.3) Assume there is a truncation time $\tau < \infty$ such that, for some constant δ, $P(\epsilon_0 > \tau | X) \geq \delta > 0$ almost surely with respect to the probability measure of X. This implies that $\Lambda_0(\tau) < \infty$.

(C.4) The error e_0's density f and its derivative $\dot{f}$ are bounded and

$$\int \left(\dot{f}(t)/f(t)\right)^2 f(t)\ dt < \infty.$$

(C.5) The conditional density of C given X and its derivative $\dot{g}_{C|X}$ are uniformly bounded for all possible values of X. That is,

$$\sup_{x \in \mathcal{X}} g_{C|X}(t|X = x) \leq K_1, \quad \sup_{x \in \mathcal{X}} |\dot{g}_{C|X}(t|X = x)| \leq K_2$$

for all $t \leq \tau$ with some constants $K_1, K_2 > 0$, where τ is the truncation time defined in Condition C.4.

(C.6) Let $\mathcal{H}^p$ denote the collection of functions h on $[a, b]$ whose kth derivative $h^{(k)}$ is bounded, and satisfies the Lipschitz continuity condition with exponent m:

$$|h^{(k)}(s) - h^{(k)}(t)| \leq L|s - t|^m \text{ for } s, t \in [a, b],$$

where k is a positive integer and $m \in (0, 1]$ such that $p = k + m \geq 3$, and $L \in (0, \infty)$ is a constant. The true log hazard function $g_0 = \log \lambda_0$ belongs to $\mathcal{H}^p$, where $[a, b]$ is a bounded interval.

(C.7) For some $\eta \in (0,1)$, $u' Var(X|\epsilon_0)u \geq \eta u' E(XX'|\epsilon_0)u$ $a.s.$ for all $u \in \mathbb{R}^d$.

These conditions are usually satisfied in practice. Condition C.1 is a common regularity assumption that has been imposed in many papers; see e.g. Lai and Ying (1991). Conditions C.2(a) and C.3-C.4 were also assumed in Tsiatis (1990), in particular Condition C.3 implies that only the observations for which the observed residual time $\epsilon_{0,i} = Y_i - X_i'\beta_0$, $1 \leq i \leq n$ is no more than τ, are used in the log-likelihood. Conditions C.1-C.3 guarantee that the observed residual time

$$\{Y_i - X_i'\beta = \epsilon_{0,i} - X_i'(\beta - \beta_0) : \beta \in \mathcal{B}, i = 1, \cdots, n\}$$

included in the likelihood is within some bounded interval $[a, b]$. Condition C.5 implies Condition B in Tsiatis (1990). Here we make a stronger assumption for $g_{C|X}$: in addition to the boundedness of $g_{C|X}$ itself, we assume it has a uniformly bounded derivative. Condition C.6 is the smoothness condition imposed on the underlying hazard function of the error term, which is needed for the B-spline smoothing. In the situation when we just need to control the approximation error rate of g_0 itself, only $p \geq 1$ is required. In this particular problem, however we have to also control the approximation error rates of both the first and second derivatives of g_0, which will be clearly demonstrated in the proof of Theorem 3.3.2 later, thus we require a stronger Lipschitz condition with $p \geq 3$ here. Finally, Condition C.7 was also proposed for the panel count data model in Wellner and Zhang (2007). As noted in their Remark 3.4, this Condition C.7 can be justified in many applications when Condition C.2(b) is satisfied.

For any $g_1, g_2 \in \mathcal{H}^p$, define the norm

$$\|g_1 - g_2\|_2 = \left\{ \int_a^b (g_1(t) - g_2(t))^2 \, d\Lambda_0(t) \right\}^{1/2}. \tag{3.11}$$

For any $\theta_1 = (\beta_1, g_1)$ and $\theta_2 = (\beta_2, g_2)$ in the space of $\Theta^p = \mathcal{B} \times \mathcal{H}^p$, define the following distance

$$d(\theta_1, \theta_2) = \left\{|\beta_1 - \beta_2|^2 + \|g_1 - g_2\|_2^2\right\}^{1/2}, \tag{3.12}$$

where $|\beta_1 - \beta_2|$ is the Euclidean distance.

Denote $\mathcal{S}_n$ as an abbreviation for $\mathcal{S}_n(T_{K_n}, K_n, l)$, the spline space of order l with K_n interior knots. Let $\mathcal{H}_n^p = \{h : h \in \mathcal{S}_n \cap \mathcal{H}^p\}$. Clearly we have $\mathcal{H}_n^p \subseteq \mathcal{H}_{n+1}^p \subseteq \cdots \subseteq \mathcal{H}^p$ for all $n \geq 1$. Denote $\Theta_n^p = \mathcal{B} \times \mathcal{H}_n^p$ and the sieve estimator $\hat{\theta}_n = (\hat{\beta}_n, \hat{g}_n)$ is the maximizer of the empirical log-likelihood $n^{-1} l_n(\theta; Z)$ over the sieve space Θ_n^p, where $Z = (Y, \Delta, X)$. The following theorem gives the convergence rate of the proposed estimator $\hat{\theta}_n$ to the true parameter $\theta_0 = (\beta_0, g_0)$.

Theorem 3.3.1. Let $K_n = O(n^\nu)$, where ν satisfies the restriction $\frac{1}{2(1+p)} < \nu < \frac{1}{2p}$ with p being the smoothness parameter defined in Condition C.6. Suppose conditions C.1-C.7 hold and the failure time T satisfies model (3.1), then

$$d(\hat{\theta}_n, \theta_0) = O_p\{n^{-\min(p\nu, (1-\nu)/2)}\},$$

where $d(\cdot, \cdot)$ is defined in (3.12).

This theorem implies that if $\nu = 1/(1+2p)$, $d(\hat{\theta}_n, \theta_0) = O_p(n^{-p/(1+2p)})$ which is the optimal convergence rate in the nonparametric regression setting. Although the overall convergence rate is lower than $n^{-1/2}$, the next theorem states that the proposed estimator of the regression parameter is still asymptotically normal and semiparametrically efficient.

Theorem 3.3.2. Suppose the conditions given in Theorem 3.3.1 hold, then

$$n^{\frac{1}{2}}(\hat{\beta}_n - \beta_0) = n^{-\frac{1}{2}} I^{-1}(\beta_0) \sum_{i=1}^{n} l_{\beta_0}^*(Y_i, \Delta_i, X_i) + o_p(1) \rightarrow N(0, I^{-1}(\beta_0)) \tag{3.13}$$

in distribution, where $I(\beta_0) = E l_{\beta_0}^*(Y, \Delta, X)^{\otimes 2}$ and $l_{\beta_0}^*(Y, \Delta, X)$ is the efficient score function for the censored linear model derived by Ritov and Wellner (1988) with the following form

$$l^*_{\beta_0}(Y, \Delta, X) = \int \{X - E(X|Y - X'\beta_0 \geq t)\}\left\{-\frac{\dot{\lambda}_0}{\lambda_0}(t)\right\} dM(t),$$

here $M(t)$ is the failure counting process martingale defined as

$$M(t) = \Delta I(Y - X'\beta_0 \leq t) - \int_{-\infty}^{t} I(Y - X'\beta_0 \geq s)\lambda_0(s)\, ds.$$

Because $\hat{\beta}_n$ achieves this information lower bound and is asymptotically linear, it is asymptotically efficient among all the regular estimators. We defer all the detailed proofs of Theorems 3.3.1 and 3.3.2 to the Appendix.

3.4 Simulation Studies

Numerous simulation studies are carried out to evaluate the finite sample performance of the proposed method. In the first set of simulations, we generate failure times from the following model:

$$\log T = \beta_0 X + e_0,$$

where X follows a Bernoulli distribution with 0.3 success probability and the true slope parameter is $\beta_0 = 0$. We consider four error distributions: mixture of two normal distributions: $0.5N(0,1) + 0.5N(-1, 0.5^2)$; standard extreme-value distribution; Gumbel distribution with location parameter -0.5γ and scale parameter 0.5 with γ being the Euler constant, denoted as Gumbel$(-0.5\gamma, 0.5)$; and Weibull distribution with shape parameter 3 and scale parameter 1, denoted as Weibull(3,1). We generate censoring times (logarithm transformed) from the uniform$[c_1, c_2]$ distribution, where c_1, c_2 are chosen to produce two different censoring rates: 25% and 50%. We also include the cases without censoring as references. The sample size is set to $n = 400$.

We choose to use cubic B-splines (i.e. of order 4) to approximate the log hazard function. Three different numbers of interior knots for the B-splines are tried, which are 1, 2 and 3. The results are quite similar and we just present the case with 1 interior knot. We perform the sieve maximum likelihood analysis and obtain the estimates of the slope parameter using the Newton-Raphson algorithm which updates (β, γ) iteratively. We stop the iterations when the change of the parameter estimates or the gradient value is less than a pre-specified tolerance number (10^{-5} in our simulations). We use both methods proposed in section 3.2.2 to estimate the variance of $\hat{\beta}_n$.

For efficiency comparisons, we also include the log-rank and Gehan-weighted estimators using the R package "rankreg" by Jin et al. (2003), as well as the Buckley-James estimator by Buckley and James (1979). We calculate the theoretical semiparametric efficiency bound $I^{-1}(\beta_0)$, and scale it by the sample size, i.e., $\sigma^* = \sqrt{I^{-1}(\beta_0)/n}$, which serves as the reference standard error under the fully efficient situation. The result based on 1000 simulated datasets for each scenario is summarized in Table 3.1.

The proposed parameter estimator is virtually unbiased. The variance estimators based on two methods, denoted as ^{1}SEE and ^{2}SEE both capture the variability of the parameter nicely and the confidence intervals have proper coverage probabilities. It is known that the log-rank estimator is asymptotically efficient under the standard extreme-value error distribution. We observe that the proposed estimators have similar variances as the log-rank estimators and are more efficient than both Gehan-weighted and Buckley-James estimators. Under both the mixture normal and the Gumbel errors, the proposed estimators are more efficient than all three other estimators, especially the log-rank and Buckley-James estimators. Finally for the Weibull(3,1) error, the proposed estimators are quite similar to the Gehan-weighted and Buckley-James estimators in terms of the efficiency and are slightly more efficient than the log-rank estimators. Under all error distributions, the standard errors of the proposed estimators are close to the theoretical standard errors calculated from the efficient score function.

In addition to the slope parameter estimator $\hat{\beta}_n$ that is of our main interest, as a by-product we also obtain the estimate of hazard function

$$\hat{\lambda}_n(t) = \exp\{\sum_{j=1}^{q_n} \hat{\gamma}_{n,j} B_j(t)\}$$

of the error term. Figure 3.1 plots the estimated hazard function over time for each of the four error distributions under the 25% censoring rate in the solid line, and the dashed line plots the true hazard function over time. Two lines are close to each other in all cases except for the mixture normal error case, which indicates reasonable estimations of the hazard functions.

In the second set of simulation studies, the failure times are generated from the model

$$\log T = 2 + X_1 + X_2 + e_0,$$

where X_1 is Bernoulli with success probability 0.5, X_2 is independent normal with mean 0 and standard deviation 0.5. This is the same model used by Jin et al. (2006) and Zeng and Lin (2007). We consider six error distributions: standard normal; standard extreme-value; mixtures of $N(0,1)$ and $N(0,3^2)$ with mixing probabilities (0.5,0.5) and (0.95,0.05), denoted by $0.5N(0,1) + 0.5N(0,3^2)$ and $0.95N(0,1) + 0.05N(0,3^2)$, respectively; Gumbel$(-0.5\gamma, 0.5)$ and $0.5N(0,1)+0.5N(-1,0.5^2)$ which are considered in the first set of simulations. The first four distributions were also considered by Zeng and Lin (2007). We do not choose the two Weibull distributions in Zeng and Lin (2007), namely Weibull(2,1) and Weibull(0.5,1). Although Zeng and Lin (2007) stated that the Weibull distributions in their simulations pertain to exponential of the errors, their simulation code actually reflected that the error themselves were generated from the Weibull distributions. However, those two Weibull distributions do not satisfy the bounded information condition in C.4, which implies that $I(\beta_0)$, the information matrix under the semiparametric efficiency, is unbounded and thus the likelihood based estimator has a convergence rate faster than $n^{-1/2}$. Similarly to Zeng and Lin (2007), the censoring times are generated from the uniform$[0, c]$ distribution, where c is chosen to produce a 25% censoring rate. We set the sample size n to 200, 400 and 600.

We choose cubic B-splines with one interior knot for $n = 200$ and 400, two interior knots for $n = 600$, and perform the same search algorithm as in the first set of simulations. Log-rank and Gehan-weighted estimators are included for efficiency comparisons. Table 3.2 summarizes the results of these studies based on 1000 simulated datasets. The bias of the proposed estimators of β_1 and β_2 are negligible. Both variance estimation procedures yield nice standard error estimates of the parameter estimators and the 95% confidence intervals have proper coverage probabilities, especially when the sample size is large. For the $N(0, 1)$ error and the two mixture of normal errors that are also considered in Zeng and Lin (2007), the proposed estimators are more efficient than the log-rank estimators and have similar variances to the Gehan-weighted estimators, especially when sample size is large. For the standard extreme-value error, the proposed estimators are more efficient than the Gehan-weighted estimator and similar to the log-rank estimators, which is known to be the most efficient estimators under this particular error. For the Gumbel$(-0.5\gamma, 0.5)$ and $0.5N(0, 1) + 0.5N(-1, 0.5^2)$ errors, the proposed estimators are most efficient compared to the other two estimators. Under all six error distributions, when sample size is large, the standard errors of the proposed estimators are quite close to the theoretical standard errors under the fully efficient situations.

To explicitly visualize the problems of the parameter estimations under two Weibull errors that were used in Zeng and Lin (2007), we also conduct simulations under these two error distributions using the same sample sizes as theirs, i.e., $n = 100$, 200 and 400. Table 3.3 summarizes the simulation results. All estimators are practically unbiased. For Weibull(0.5,1), the proposed estimators yield much smaller standard errors than the log-rank and Gehan-weighted estimators. Moreover, for all three estimating methods, the product of the sample size and the variance of the parameter estimators (denoted as $n\mathrm{SE}^2$ in Table 3.3) are diminishing as n increases, which are supposed to be around a constant for a root-n consistent estimator. For Weibull(2,1), the proposed estimators also yield smaller standard errors than the other two estimators, although not much. The $n\mathrm{SE}^2$ values are also decreasing as n increases for the proposed estimators. The findings indicate that the proposed estimators have a faster convergence rate than the usual $n^{-1/2}$ rate under both Weibull errors, and all three estimators have a faster convergence rate than $n^{-1/2}$ under the Weibull(0.5,1) error. As mentioned earlier, the covariance matrix $I^{-1}(\beta_0)$ is singular under these two Weibull distributions.

3.5 Examples

We first use the Stanford heart transplantation data (Miller and Halpern 1982) as an illustrative example. This dataset was also reanalyzed by Jin et al. (2006) using their proposed least squares estimators. Following their analysis, we consider the same two models: the first one regressing the base-10 logarithm of the survival time on the patients' age at transplant and the T5 mismatch score for the 157 patients with complete records on the T5 measure, and the second one regressing the base-10 logarithm of the survival time on age and age^2 for the 152 patients who survived for at least 10 years after heart transplantation. For the first model, 102 patients were deceased and 55 were still alive by February 1980, and for the second model, 97 patients were deceased and 55 patients were censored. We fit these two models using the proposed method with five cubic B-spline basis functions (i.e. one interior knot).

We report the parameter estimates and the standard error estimates in Table 3.4 and compare them with the Gehan-weighted estimators reported by Jin et al. (2006) and the Buckley-James estimators reported by Miller and Halpern (1982). For the first model, the parameter estimates for the age effect are fairly similar among all estimators and the standard error estimate from the proposed method tends to be smaller, while the parameter estimates for the T5 mismatch score vary across different estimators with none of them being significant at the 0.05 level. The disparities of the parameter estimates for the T5 may due to the reason that the AFT model with age and T5 as covariates does not fit the data ideally, which was pointed out in Miller and Halpern (1982). For the second model with age and age^2 being the covariates, both parameter estimates are very close across all methods and the standard error estimates from the proposed method are the smallest.

We consider the well-known Mayo primary biliary cirrhosis (PBC) study (Fleming and Harrington 1991, app. D.1) as the second example. The dataset contains information about the survival time and prognostic factors for 418 patients. Jin et al. (2003), Jin et al. (2006) and Zeng and Lin (2007) fitted the accelerated failure time model with five covariates, namely age, log(albumin), log(bilirubin), edema, and log(protime). They used the rank-based, least squares and kernel-smoothed profile likelihood estimators and reported the slope estimators with their estimated standard errors for the five covariates. We fit the same model using the proposed method with one and two interior knots. The parameter estimates and the standard error estimates are similar and we report the result associated with one interior knot in Table 3.5. Our parameter estimates are close to the Gehan-weighted estimates of Jin et al. (2003), the least squares estimates of Jin et al. (2006) and the kernel smoothed maximal likelihood estimates of Zeng and Lin (2007), while our standard error estimates tend to be smaller.

Since the conditional survival function at time t given covariates X follows

$$\begin{aligned} S(t|X) &= P(T \geq t|X) = P(\log T - X'\beta_0 \geq \log t - X'\beta_0|X) \\ &= P(e_0 \geq \log t - X'\beta_0|X) = P(e_0 \geq \log t - X'\beta_0), \quad \text{since } e_0 \perp X \\ &= \exp\{-\Lambda_0(\log t - X'\beta_0)\}, \end{aligned}$$

a natural estimator of $S(t|X)$ is

$$\hat{S}_n(t|X) = \exp\{-\hat{\Lambda}_n(\log t - X'\hat{\beta}_n)\}, \tag{3.14}$$

where $\hat{\Lambda}_n(t) = \int_{-\infty}^{t} \hat{\lambda}_n(s)\, ds$ with $\hat{\lambda}_n(s) = \exp\big\{\sum_{j=1}^{q_n} \hat{\gamma}_{n,j} B_j(s)\big\}$. Then we can also estimate the marginal survival function for a subgroup by averaging the conditional survival function estimates of that subgroup. Figure 3.2 shows the estimated survival curves for the PBC data in two groups: Edema=0 versus Edema=1. The model-based estimates agree fairly well with the Kaplan-Meier estimates except the right tail of the group with Edema=1. One reason for this lack of agreement is that there are only 50 patients in the Edema=1 group and even less number of deaths by the end of the follow-up time.

3.6 An extended general theorem on the asymptotic normality of semiparametric M-estimators

In this section, we extend the general theorem introduced by Wellner and Zhang (2007), which deals with the asymptotic normality of semiparametric M-estimators of regression parameters when convergence rate of the estimator for nuisance parameters is slower than $n^{-1/2}$. In their theorem, the parameters of interest and the nuisance parameters are assumed to be separate. We consider a more general setting when the nuisance parameters can be a function of the parameters of interest. The theorem is crucial in the proof of asymptotic normality of our proposed estimators given in Theorem 3.3.2.

Some empirical process notation will be used from now on. We denote $Pf = \int f(x)\, dP(x)$ and $\mathbb{P}_n f = n^{-1}\sum_{i=1}^n f(X_i)$, with P being a probability measure, and denote $\mathbb{G}_n f = n^{1/2}(\mathbb{P}_n - P)f$. Given i.i.d. observations $X_1, X_2, \cdots, X_n$, we try to estimate the unknown parameters $(\beta, \Lambda(\cdot,\beta))$ by maximizing an objective function $n^{-1}\sum_{i=1}^n m(\beta, \Lambda(\cdot,\beta); X_i) = \mathbb{P}_n m(\beta, \Lambda(\cdot,\beta); X)$, where β is the parameter of interest with finite dimension and Λ is an infinite-dimensional nuisance parameter and can be a function of β. If the objective function m is the log-likelihood function of a single observation, then the estimator becomes the semiparametric maximum likelihood estimator. Here we adopt the similar notation in Wellner and Zhang (2007).

Let $\theta = (\beta, \Lambda(\cdot,\beta))$, where $\beta \in \mathbb{R}^d$ and Λ is an infinite-dimensional parameter in the class $\mathcal{F}$. For any fixed $\Lambda \in \mathcal{F}$, let $\{\Lambda_\eta : \eta \text{ in a neighborhood of } 0 \in \mathbb{R}\}$ be a smooth curve in $\mathcal{F}$ running through Λ at $\eta = 0$, i.e., $\Lambda_{\eta=0} = \Lambda$. Let

$$\mathbf{H} = \{h : h = \frac{\partial \Lambda_\eta}{\partial \eta}|_{\eta=0}, \Lambda_\eta \in \mathcal{F}\}.$$

For all $\Lambda(\cdot,\beta) \in \mathcal{F}$, assume $\Lambda_\beta^{(k)}(\cdot,\beta)$, the kth derivative with respect to β exists with $k \geq 2$. Then since for a small δ, we have $\Lambda(\cdot,\beta+\delta) - \Lambda(\cdot,\beta) = \dot{\Lambda}_\beta(\cdot,\beta)\delta + o(\delta)$, by the definition of functional derivatives, it follows that

$$\begin{aligned}
& \lim_{\delta\to 0}\frac{1}{\delta}\big\{m(\beta,\Lambda(\cdot,\beta+\delta);x)-m(\beta,\Lambda(\cdot,\beta);x)\big\} \\
= & \lim_{\delta\to 0}\frac{1}{\delta}\big\{m(\beta,\Lambda(\cdot,\beta)+\dot\Lambda_\beta(\cdot,\beta)\delta+o(\delta);x)-m(\beta,\Lambda(\cdot,\beta)+\dot\Lambda_\beta(\cdot,\beta)\delta;x)\big\} \\
& \quad +\lim_{\delta\to 0}\frac{1}{\delta}\big\{m(\beta,\Lambda(\cdot,\beta)+\dot\Lambda_\beta(\cdot,\beta)\delta;x)-m(\beta,\Lambda(\cdot,\beta);x)\big\} \\
= & \lim_{\delta\to 0}\dot m_2(\beta,\Lambda(\cdot,\beta)+\dot\Lambda_\beta(\cdot,\beta)\delta;x)[o(\delta)/\delta]+\dot m_2(\beta,\Lambda(\cdot,\beta);x)[\dot\Lambda_\beta(\cdot,\beta)] \\
= & \dot m_2(\beta,\Lambda(\cdot,\beta);x)[\dot\Lambda_\beta(\cdot,\beta)],
\end{aligned}$$

where the subscript 2 indicates that the derivatives are taking with respect to the second argument of the function and the last equality holds because

$$\lim_{\delta\to 0}\dot m_2(\beta,\Lambda(\cdot,\beta)+\dot\Lambda_\beta(\cdot,\beta)\delta;x)[o(\delta)/\delta]=0.$$

Similarly we have

$$\begin{aligned}
& \lim_{\delta\to 0}\frac{1}{\delta}\big\{\dot m_2(\beta,\Lambda(\cdot,\beta+\delta);x)[h]-\dot m_2(\beta,\Lambda(\cdot,\beta);x)[h]\big\} \\
& \quad =\ddot m_{22}(\beta,\Lambda(\cdot,\beta);x)[h,\dot\Lambda_\beta(\cdot,\beta)]
\end{aligned}$$

and

$$\begin{aligned}
& \lim_{\delta\to 0}\frac{1}{\delta}\big\{\dot m_2(\beta,\Lambda(\cdot,\beta);x)[\dot\Lambda_\beta(\cdot,\beta+\delta)]-\dot m_2(\beta,\Lambda(\cdot,\beta);x)[\dot\Lambda_\beta(\cdot,\beta)]\big\} \\
& \quad =\dot m_2(\beta,\Lambda(\cdot,\beta);x)[\ddot\Lambda_\beta(\cdot,\beta)].
\end{aligned}$$

Thus according to the chain rule of the functional derivatives, we define

$$
\begin{aligned}
\dot{m}_\beta(\beta, \Lambda(\cdot,\beta); x) &= \nabla_\beta\ m(\beta, \Lambda(\cdot,\beta); x) \\
&= \dot{m}_1(\beta, \Lambda(\cdot,\beta); x) + \dot{m}_2(\beta, \Lambda(\cdot,\beta); x)[\dot{\Lambda}_\beta(\cdot,\beta)], \\
\dot{m}_\Lambda(\beta, \Lambda(\cdot,\beta); x)[h] &= \dot{m}_2(\beta, \Lambda(\cdot,\beta); x)[h] = \left.\frac{\partial m(\beta, (\Lambda+\eta h)(\cdot,\beta); x)}{\partial \eta}\right|_{\eta=0}, \\
\ddot{m}_{\beta\beta}(\beta, \Lambda(\cdot,\beta); x) &= \nabla_\beta^2\ m(\beta, \Lambda(\cdot,\beta); x) = \nabla_\beta\ \dot{m}_\beta(\beta, \Lambda(\cdot,\beta); x) \\
&= \ddot{m}_{11}(\beta, \Lambda(\cdot,\beta); x) + \ddot{m}_{12}(\beta; \Lambda(\cdot,\beta); x)[\dot{\Lambda}_\beta(\cdot,\beta)] \\
&\quad + \ddot{m}_{21}(\beta, \Lambda(\cdot,\beta); x)[\dot{\Lambda}_\beta(\cdot,\beta)] \\
&\quad + \ddot{m}_{22}(\beta, \Lambda(\cdot,\beta); x)[\dot{\Lambda}_\beta(\cdot,\beta), \dot{\Lambda}_\beta(\cdot,\beta)] \\
&\quad + \dot{m}_2(\beta, \Lambda(\cdot,\beta); x)[\ddot{\Lambda}_{\beta\beta}(\cdot,\beta)], \\
\ddot{m}_{\beta\Lambda}(\beta, \Lambda(\cdot,\beta); x)[h] &= \left.\frac{\partial \dot{m}_\beta(\beta, (\Lambda+\eta h)(\cdot,\beta); x)}{\partial \eta}\right|_{\eta=0} \\
&= \ddot{m}_{12}(\beta, \Lambda(\cdot,\beta); x)[h(\cdot,\beta)] \\
&\quad + \ddot{m}_{22}(\beta, \Lambda(\cdot,\beta); x)[\dot{\Lambda}_\beta(\cdot,\beta), h(\cdot,\beta)] \\
&\quad + \dot{m}_2(\beta, \Lambda(\cdot,\beta); x)[\dot{h}_\beta(\cdot,\beta)], \\
\ddot{m}_{\Lambda\beta}(\beta, \Lambda(\cdot,\beta); x)[h] &= \nabla_\beta\ \dot{m}_2(\beta, \Lambda(\cdot,\beta); x)[h(\cdot,\beta)] \\
&= \ddot{m}_{21}(\beta, \Lambda(\cdot,\beta); x)[h(\cdot,\beta)] \\
&\quad + \ddot{m}_{22}(\beta, \Lambda(\cdot,\beta); x)[h(\cdot,\beta), \dot{\Lambda}_\beta(\cdot,\beta)] \\
&\quad + \dot{m}_2(\beta, \Lambda(\cdot,\beta); x)[\dot{h}_\beta(\cdot,\beta)], \\
\ddot{m}_{\Lambda\Lambda}(\beta, \Lambda(\cdot,\beta); x)[h_1, h_2] &= \ddot{m}_{22}(\beta, \Lambda(\cdot,\beta); x)[h_1, h_2] \\
&= \left.\frac{\nabla^2\ m(\beta, \Lambda_{\eta_j}(\cdot,\beta); x)}{\partial \eta_1 \partial \eta_2}\right|_{\eta_j=0, j=1,2}.
\end{aligned}
$$

As noted before, the subscript 1 or 2 in the derivatives indicates the derivatives are taking to the first or the second argument of the function, and h inside of the square brackets is a function denoting the direction of the functional derivative with respect to Λ. Note that for the second derivatives $\ddot{m}_{\beta\Lambda}$ and $\ddot{m}_{\Lambda\beta}$, we implicitly require the direction h to be a differentiable function with respect to β. Similarly as in Wellner and Zhang (2007), we also define

$$\begin{aligned}
\dot{S}_\beta(\beta, \Lambda(\cdot, \beta)) &= P\dot{m}_\beta(\beta, \Lambda(\cdot, \beta); X), \\
\dot{S}_\Lambda(\beta, \Lambda(\cdot, \beta))[h] &= P\dot{m}_\Lambda(\beta, \Lambda(\cdot, \beta); X)[h], \\
\dot{S}_{\beta,n}(\beta, \Lambda(\cdot, \beta)) &= \mathbb{P}_n\dot{m}_\beta(\beta, \Lambda(\cdot, \beta); X), \\
\dot{S}_{\Lambda,n}(\beta, \Lambda(\cdot, \beta))[h] &= \mathbb{P}_n\dot{m}_\Lambda(\beta, \Lambda(\cdot, \beta); X)[h], \\
\ddot{S}_{\beta\beta}(\beta, \Lambda(\cdot, \beta)) &= P\ddot{m}_{\beta\beta}(\beta, \Lambda(\cdot, \beta); X), \\
\ddot{S}_{\Lambda\Lambda}(\beta, \Lambda(\cdot, \beta))[h, h] &= P\ddot{m}_{\Lambda\Lambda}(\beta, \Lambda(\cdot, \beta); X)[h, h],
\end{aligned}$$

and

$$\ddot{S}_{\beta\Lambda}(\beta, \Lambda(\cdot, \beta))[h] = \ddot{S}'_{\Lambda\beta}(\beta, \Lambda(\cdot, \beta))[h] = P\ddot{m}_{\beta\Lambda}(\beta, \Lambda(\cdot, \beta); X)[h].$$

Furthermore, for $\mathbf{h} = (h_1, h_2, \cdots, h_d)' \in \mathbf{H}^d$, we denote

$$\begin{aligned}
&\dot{m}_\Lambda(\beta, \Lambda(\cdot, \beta); x)[\mathbf{h}] = (\dot{m}_\Lambda(\beta, \Lambda(\cdot, \beta); x)[h_1], \cdots, \dot{m}_\Lambda(\beta, \Lambda(\cdot, \beta); x)[h_d])', \\
&\ddot{m}_{\beta\Lambda}(\beta, \Lambda(\cdot, \beta); x)[\mathbf{h}] = (\ddot{m}_{\beta\Lambda}(\beta, \Lambda(\cdot, \beta); x)[h_1], \cdots, \ddot{m}_{\beta\Lambda}(\beta, \Lambda(\cdot, \beta); x)[h_d]), \\
&\ddot{m}_{\Lambda\beta}(\beta, \Lambda(\cdot, \beta); x)[\mathbf{h}] = (\ddot{m}_{\Lambda\beta}(\beta, \Lambda(\cdot, \beta); x)[h_1], \cdots, \ddot{m}_{\Lambda\beta}(\beta, \Lambda(\cdot, \beta); x)[h_d])', \\
&\ddot{m}_{\Lambda\Lambda}(\beta, \Lambda(\cdot, \beta); x)[\mathbf{h}, h] = (\ddot{m}_{\Lambda\Lambda}(\beta, \Lambda(\cdot, \beta); x)[h_1, h], \cdots, \ddot{m}_{\Lambda\Lambda}(\beta, \Lambda(\cdot, \beta); x)[h_d, h])',
\end{aligned}$$

and define correspondingly

$$\begin{aligned}
\dot{S}_\Lambda(\beta, \Lambda(\cdot, \beta))[\mathbf{h}] &= P\dot{m}_\Lambda(\beta, \Lambda(\cdot, \beta); X)[\mathbf{h}], \\
\dot{S}_{\Lambda,n}(\beta, \Lambda(\cdot, \beta))[\mathbf{h}] &= \mathbb{P}_n\dot{m}_\Lambda(\beta, \Lambda(\cdot, \beta); X)[\mathbf{h}], \\
\ddot{S}_{\beta\Lambda}(\beta, \Lambda(\cdot, \beta))[\mathbf{h}] &= P\ddot{m}_{\beta\Lambda}(\beta, \Lambda(\cdot, \beta); X)[\mathbf{h}], \\
\ddot{S}_{\Lambda\beta}(\beta, \Lambda(\cdot, \beta))[\mathbf{h}] &= P\ddot{m}_{\Lambda\beta}(\beta, \Lambda(\cdot, \beta); X)[\mathbf{h}], \\
\ddot{S}_{\Lambda\Lambda}(\beta, \Lambda(\cdot, \beta))[\mathbf{h}, h] &= P\ddot{m}_{\Lambda\Lambda}(\beta, \Lambda(\cdot, \beta); X)[\mathbf{h}, h].
\end{aligned}$$

In order to have the asymptotic normality result for the M-estimator $\hat{\beta}_n$, the assumptions we need to make look similar to those in Wellner and Zhang (2007) but all the derivatives with respect to β involve the chain rule and contain more components (defined clearly in the pervious paragraph). We list the following assumptions:

A1. (Consistency and rate of convergence) $|\hat{\beta}_n - \beta_0| = o_p(1)$ and $\|\hat{\Lambda}_n - \Lambda_0\| = O_p(n^{-\gamma})$ for some $\gamma > 0$ and some norm $\|\cdot\|$.

A2. $\dot{S}_\beta(\beta_0, \Lambda_0(\cdot, \beta_0)) = 0$ and $\dot{S}_\Lambda(\beta_0, \Lambda_0(\cdot, \beta_0))[h] = 0$ for all $h \in \mathbf{H}$.

A3. (Positive information) There exists an $\mathbf{h}^* = (h_1^*, \cdots, h_d^*)'$, where $h_j^* \in \mathbf{H}$ for $j = 1, \cdots, d$, such that

$$\ddot{S}_{\beta\Lambda}(\beta_0, \Lambda_0(\cdot, \beta_0))[h] - \ddot{S}_{\Lambda\Lambda}(\beta_0, \Lambda_0(\cdot, \beta_0))[\mathbf{h}^*, h] = 0$$

for all $h \in \mathbf{H}$. Furthermore, the matrix

$$\begin{aligned} A &= -\ddot{S}_{\beta\beta}(\beta_0, \Lambda_0(\cdot, \beta_0)) + \ddot{S}_{\Lambda\beta}(\beta_0, \Lambda_0(\cdot, \beta_0))[\mathbf{h}^*] \\ &= -P\big\{\ddot{m}_{\beta\beta}(\beta_0, \Lambda_0(\cdot, \beta_0); X) - \ddot{m}_{\Lambda\beta}(\beta_0, \Lambda_0(\cdot, \beta_0); X)[\mathbf{h}^*]\big\} \end{aligned}$$

is nonsingular.

A4. The estimator $(\hat{\beta}_n, \hat{\Lambda}_n)$ satisfies

$$\dot{S}_{\beta,n}(\hat{\beta}_n, \hat{\Lambda}_n(\cdot, \hat{\beta}_n)) = o_p(n^{-1/2}) \text{ and } \dot{S}_{\Lambda,n}(\hat{\beta}_n, \hat{\Lambda}_n(\cdot, \hat{\beta}_n))[\mathbf{h}^*] = o_p(n^{-1/2}).$$

A5. (Stochastic equicontinuity) For any $\delta_n \downarrow 0$ and $C > 0$,

$$\begin{aligned} \sup_{|\beta-\beta_0| \le \delta_n, \|\Lambda - \Lambda_0\| \le Cn^{-\gamma}} & |\sqrt{n}(\dot{S}_{\beta,n} - \dot{S}_\beta)(\beta, \Lambda(\cdot, \beta)) \\ & -\sqrt{n}(\dot{S}_{\beta,n} - \dot{S}_\beta)(\beta_0, \Lambda_0(\cdot, \beta_0))| = o_p(1) \end{aligned}$$

and

$$\begin{aligned} \sup_{|\beta-\beta_0| \le \delta_n, \|\Lambda - \Lambda_0\| \le Cn^{-\gamma}} & |\sqrt{n}(\dot{S}_{\Lambda,n} - \dot{S}_\Lambda)(\beta, \Lambda(\cdot, \beta))[\mathbf{h}^*] \\ & -\sqrt{n}(\dot{S}_{\Lambda,n} - \dot{S}_\Lambda)(\beta_0, \Lambda_0(\cdot, \beta_0))[\mathbf{h}^*]| = o_p(1). \end{aligned}$$

A6. (Smoothness of the model) For some $\alpha > 1$ satisfying $\alpha\gamma > 1/2$, and for (β, Λ) in a neighborhood of $(\beta_0, \Lambda_0) : \{(\beta, \Lambda) : |\beta - \beta_0| \leq \delta_n, \|\Lambda - \Lambda_0\| \leq Cn^{-\gamma}\}$,

$$\begin{aligned}&\big|\dot{S}_\beta(\beta, \Lambda(\cdot, \beta)) - \dot{S}_\beta(\beta_0, \Lambda_0(\cdot, \beta_0))\\ &\quad - \ddot{S}_{\beta\beta}(\beta_0, \Lambda_0(\cdot, \beta_0))(\beta - \beta_0) - \ddot{S}_{\beta\Lambda}(\beta_0, \Lambda_0(\cdot, \beta_0))[\Lambda(\cdot, \beta) - \Lambda_0(\cdot, \beta_0)]\big|\\ &\qquad = o(|\beta - \beta_0|) + O(\|\Lambda - \Lambda_0\|^\alpha)\end{aligned}$$

and

$$\begin{aligned}&\big|\dot{S}_\Lambda(\beta, \Lambda(\cdot, \beta))[\mathbf{h}^*] - \dot{S}_\Lambda(\beta_0, \Lambda_0(\cdot, \beta_0))[\mathbf{h}^*]\\ &\quad - \ddot{S}_{\Lambda\beta}(\beta_0, \Lambda_0(\cdot, \beta_0))[\mathbf{h}^*](\beta - \beta_0) - \ddot{S}_{\Lambda\Lambda}(\beta_0, \Lambda_0(\cdot, \beta_0))[\mathbf{h}^*, \Lambda(\cdot, \beta) - \Lambda_0(\cdot, \beta_0)]\big|\\ &\quad = o(|\beta - \beta_0|) + O(\|\Lambda - \Lambda_0\|^\alpha).\end{aligned}$$

The following theorem is an extension to the Theorem 6.1 in Wellner and Zhang (2007) where the infinite dimensional parameter Λ is a function of the finite-dimensional parameter β.

Theorem 3.6.1. Suppose that assumptions A1-A6 hold. Then

$$\begin{aligned}\sqrt{n}(\hat{\beta}_n - \beta_0) &= A^{-1}\sqrt{n}\,\mathbb{P}_n m^*(\beta_0, \Lambda_0(\cdot, \beta_0); X) + o_p(1)\\ &\rightarrow_d N(0, A^{-1}B(A^{-1})'),\end{aligned}$$

where $m^*(\beta_0, \Lambda_0(\cdot, \beta_0); x) = \dot{m}_\beta(\beta_0, \Lambda_0(\cdot, \beta_0); x) - \dot{m}_\Lambda(\beta_0, \Lambda_0(\cdot, \beta_0); x)[\mathbf{h}^*]$,

$$B = Pm^*(\beta_0, \Lambda_0(\cdot, \beta_0); X)^{\otimes 2} = P\{m^*(\beta_0, \Lambda_0(\cdot, \beta_0); X)m^*(\beta_0, \Lambda_0(\cdot, \beta_0); X)'\},$$

and A is given in assumption A3.

Proof. The proof follows along the proof for the Theorem 6.1 of Wellner and Zhang (2007). Assumptions A1 and A5 yield

$$\sqrt{n}(\dot{S}_{\beta,n} - \dot{S}_\beta)(\hat{\beta}_n, \hat{\Lambda}_n(\cdot, \hat{\beta}_n)) - \sqrt{n}(\dot{S}_{\beta,n} - \dot{S}_\beta)(\beta_0, \Lambda_0(\cdot, \beta_0)) = o_p(1).$$

Since $\dot{S}_{\beta,n}(\hat{\beta}_n, \hat{\Lambda}_n(\cdot, \hat{\beta}_n)) = o_p(n^{-1/2})$ by A4 and $\dot{S}_\beta(\beta_0, \Lambda(\cdot, \beta_0)) = 0$ by A2, we have

$$\sqrt{n}\dot{S}_\beta(\hat{\beta}_n, \hat{\Lambda}_n(\cdot, \hat{\beta}_n)) + \sqrt{n}\dot{S}_{\beta,n}(\beta_0, \Lambda(\cdot, \beta_0)) = o_p(1).$$

Similarly,

$$\sqrt{n}\dot{S}_\Lambda(\hat{\beta}_n, \hat{\Lambda}_n(\cdot, \hat{\beta}_n))[\mathbf{h}^*] + \sqrt{n}\dot{S}_{\Lambda,n}(\beta_0, \Lambda(\cdot, \beta_0))[\mathbf{h}^*] = o_p(1).$$

Combining these equalities and assumption A6 yields

$$\begin{aligned} & \ddot{S}_{\beta\beta}(\beta_0, \Lambda_0(\cdot, \beta_0))(\hat{\beta}_n - \beta_0) + \ddot{S}_{\beta\Lambda}(\beta_0, \Lambda_0(\cdot, \beta_0))[\hat{\Lambda}_n(\cdot, \hat{\beta}_n) - \Lambda_0(\cdot, \beta_0)] \\ & \quad + \dot{S}_{\beta,n}(\beta_0, \Lambda_0(\cdot, \beta_0)) + o(|\hat{\beta}_n - \beta_0|) + O(\|\hat{\Lambda}_n - \Lambda_0\|^\alpha) \\ = \; & o_p(n^{-1/2}) \end{aligned} \tag{3.15}$$

and

$$\begin{aligned} & \ddot{S}_{\Lambda\beta}(\beta_0, \Lambda_0(\cdot, \beta_0))[\mathbf{h}^*](\hat{\beta}_n - \beta_0) + \ddot{S}_{\Lambda\Lambda}(\beta_0, \Lambda_0(\cdot, \beta_0))[\mathbf{h}^*, \hat{\Lambda}_n(\cdot, \hat{\beta}_n) - \Lambda_0(\cdot, \beta_0)] \\ & \quad + \dot{S}_{\Lambda,n}(\beta_0, \Lambda_0(\cdot, \beta_0))[\mathbf{h}^*] + o(|\hat{\beta}_n - \beta_0|) + O(\|\hat{\Lambda}_n - \Lambda_0\|^\alpha). \\ = \; & o_p(n^{-1/2}). \end{aligned} \tag{3.16}$$

Since $\alpha > 1$ with $\alpha\gamma > 1/2$, the rate of convergence assumption in A1 implies that $\sqrt{n}O(\|\hat{\Lambda}_n - \Lambda_0\|^\alpha) = o_p(1)$, then (3.15) – (3.16) together with A3 yields

$$\begin{aligned} & (\ddot{S}_{\beta\beta}(\beta_0, \Lambda_0(\cdot, \beta_0)) - \ddot{S}_{\Lambda\beta}(\beta_0, \Lambda_0(\cdot, \beta_0))[\mathbf{h}^*])(\hat{\beta}_n - \beta_0) + o(|\hat{\beta}_n - \beta_0|) \\ = \; & -(\dot{S}_{\beta,n}(\beta_0, \Lambda_0(\cdot, \beta_0)) - \dot{S}_{\Lambda,n}(\beta_0, \Lambda_0)[\mathbf{h}^*]) + o_p(n^{-1/2}), \end{aligned}$$

that is,

$$-(A + o(1))(\hat{\beta}_n - \beta_0) = -\mathbb{P}_n m^*(\beta_0, \Lambda_0(\cdot, \beta_0); X) + o_p(n^{-1/2}).$$

This yields

$$\begin{aligned} \sqrt{n}(\hat{\beta}_n - \beta_0) & = (A + o(1))^{-1}\sqrt{n}\mathbb{P}_n m^*(\beta_0, \Lambda_0(\cdot, \beta_0); X) + o_p(1) \\ & \to_d N(0, A^{-1}B(A^{-1})'). \end{aligned}$$

In some situation, if in addition to the consistency, a convergence rate (possibly sub-optimal) for the parameter estimator $\hat{\beta}_n$ is known, then instead of considering the previous assumptions A_1, A_5 and A_6, we assume the following three modified conditions:

A1'. (Rate of convergence) $|\hat{\beta}_n - \beta_0| + \|\hat{\Lambda}_n - \Lambda_0\| = O_p(n^{-\gamma})$ for some $\gamma > 0$ and some norm $\| \cdot \|$.

A5'. (Stochastic equicontinuity) For any $C > 0$,

$$\sup_{|\beta-\beta_0|+\|\Lambda-\Lambda_0\|\leq Cn^{-\gamma}} |\sqrt{n}(\dot{S}_{\beta,n} - \dot{S}_{\beta})(\beta, \Lambda(\cdot,\beta)) \\ -\sqrt{n}(\dot{S}_{\beta,n} - \dot{S}_{\beta})(\beta_0, \Lambda_0(\cdot,\beta_0))| = o_p(1)$$

and

$$\sup_{|\beta-\beta_0|+\|\Lambda-\Lambda_0\|\leq Cn^{-\gamma}} |\sqrt{n}(\dot{S}_{\Lambda,n} - \dot{S}_{\Lambda})(\beta, \Lambda(\cdot,\beta))[\mathbf{h}^*] \\ -\sqrt{n}(\dot{S}_{\Lambda,n} - \dot{S}_{\Lambda})(\beta_0, \Lambda_0(\cdot,\beta_0))[\mathbf{h}^*]| = o_p(1).$$

A6'. (Smoothness of the model) For some $\alpha > 1$ satisfying $\alpha\gamma > 1/2$, and for (β, Λ) in a neighborhood of $(\beta_0, \Lambda_0) : \{(\beta, \Lambda) : |\beta - \beta_0| + \|\Lambda - \Lambda_0\| \leq Cn^{-\gamma}\}$,

$$\Big|\dot{S}_{\beta}(\beta, \Lambda(\cdot,\beta)) - \dot{S}_{\beta}(\beta_0, \Lambda_0(\cdot,\beta_0)) \\ - \ddot{S}_{\beta\beta}(\beta_0, \Lambda_0(\cdot,\beta_0))(\beta - \beta_0) - \ddot{S}_{\beta\Lambda}(\beta_0, \Lambda_0(\cdot,\beta_0))[\Lambda(\cdot,\beta) - \Lambda_0(\cdot,\beta_0)]\Big| \\ = O\{(|\beta - \beta_0| + \|\Lambda - \Lambda_0\|)^{\alpha}\}$$

and

$$\Big|\dot{S}_{\Lambda}(\beta, \Lambda(\cdot,\beta))[\mathbf{h}^*] - \dot{S}_{\Lambda}(\beta_0, \Lambda_0(\cdot,\beta_0))[\mathbf{h}^*] \\ - \ddot{S}_{\Lambda\beta}(\beta_0, \Lambda_0(\cdot,\beta_0))[\mathbf{h}^*](\beta - \beta_0) - \ddot{S}_{\Lambda\Lambda}(\beta_0, \Lambda_0(\cdot,\beta_0))[\mathbf{h}^*, \Lambda(\cdot,\beta) - \Lambda_0(\cdot,\beta_0)]\Big| \\ = O\{(|\beta - \beta_0| + \|\Lambda - \Lambda_0\|)^{\alpha}\}.$$

These modified assumptions are easier to verify when a (sub-optimal) convergence rate for $\hat{\beta}_n$ is known, and we still have the asymptotic normality result for $\hat{\beta}_n$ under these modified assumptions, which is summarized in the corollary below.

Corollary 3.6.2. Suppose that assumptions A1', A2-A4, and A5'-A6' hold. Then

$$\begin{aligned}\sqrt{n}(\hat{\beta}_n - \beta_0) &= A^{-1}\sqrt{n}\,\mathbb{P}_n m^*(\beta_0, \Lambda_0(\cdot, \beta_0); X) + o_p(1)\\ &\to_d N(0, A^{-1}B(A^{-1})'),\end{aligned}$$

where $m^*(\beta_0, \Lambda_0(\cdot, \beta_0); x)$ and B are given in Theorem 3.6.1 and A is given in assumption A3.

Proof. The proof is almost identical to the proof of the Theorem 3.6.1 except that equations (3.15) and (3.16) become

$$\begin{aligned}&\ddot{S}_{\beta\beta}(\beta_0, \Lambda_0(\cdot, \beta_0))(\hat{\beta}_n - \beta_0) + \ddot{S}_{\beta\Lambda}(\beta_0, \Lambda_0(\cdot, \beta_0))[\hat{\Lambda}_n(\cdot, \hat{\beta}_n) - \Lambda_0(\cdot, \beta_0)]\\ &\quad + \dot{S}_{\beta,n}(\beta_0, \Lambda_0(\cdot, \beta_0)) + O\{(|\beta - \beta_0| + \|\Lambda - \Lambda_0\|)^{\alpha}\}\\ &= o_p(n^{-1/2})\end{aligned} \tag{3.17}$$

and

$$\begin{aligned}&\ddot{S}_{\Lambda\beta}(\beta_0, \Lambda_0(\cdot, \beta_0))[\mathbf{h}^*](\hat{\beta}_n - \beta_0) + \ddot{S}_{\Lambda\Lambda}(\beta_0, \Lambda_0(\cdot, \beta_0))[\mathbf{h}^*, \hat{\Lambda}_n(\cdot, \hat{\beta}_n) - \Lambda_0(\cdot, \beta_0)]\\ &\quad + \dot{S}_{\Lambda,n}(\beta_0, \Lambda_0(\cdot, \beta_0))[\mathbf{h}^*] + O\{(|\beta - \beta_0| + \|\Lambda - \Lambda_0\|)^{\alpha}\}\\ &= o_p(n^{-1/2}).\end{aligned} \tag{3.18}$$

Since $\alpha > 1$ with $\alpha\gamma > 1/2$, the rate of convergence assumption in A1' implies that $\sqrt{n}O\{(|\beta - \beta_0| + \|\Lambda - \Lambda_0\|)^{\alpha}\} = O_p(n^{1/2-\alpha\gamma}) = o_p(1)$, then (3.17) – (3.18) together with A3 yields

$$\begin{aligned}&(\ddot{S}_{\beta\beta}(\beta_0, \Lambda_0(\cdot, \beta_0)) - \ddot{S}_{\Lambda\beta}(\beta_0, \Lambda_0(\cdot, \beta_0))[\mathbf{h}^*])(\hat{\beta}_n - \beta_0)\\ &= -(\dot{S}_{\beta,n}(\beta_0, \Lambda_0(\cdot, \beta_0)) - \dot{S}_{\Lambda,n}(\beta_0, \Lambda_0)[\mathbf{h}^*]) + o_p(n^{-1/2}),\end{aligned}$$

that is,

$$-A(\hat{\beta}_n - \beta_0) = -\mathbb{P}_n m^*(\beta_0, \Lambda_0(\cdot, \beta_0); X) + o_p(n^{-1/2}).$$

This yields

$$\begin{aligned}\sqrt{n}(\hat{\beta}_n - \beta_0) &= A^{-1}\sqrt{n}\mathbb{P}_n m^*(\beta_0, \Lambda_0(\cdot, \beta_0); X) + o_p(1) \\ &\rightarrow_d N(0, A^{-1}B(A^{-1})').\end{aligned}$$

3.7 Discussion

Comparing to the existing methods for estimating β in the semiparametric model (3.1), the proposed method has three advantages. Firstly, the estimating functions in (3.9) and (3.10) are smooth functions in contrast to the discrete estimating functions in (3.2) and (3.3). Thus the root search is easier and can be done fast by conventional iterative methods such as the Newton-Raphson algorithm. Secondly, the standard error estimates are obtained directly by inverting either the information matrix under the efficient score function for the slope parameters or the observed information matrix of all parameters, which are both more computationally tractable compared to the re-sampling techniques. Thirdly, the proposed estimator achieves the semiparametric efficiency bound. The kernel-smoothed profile likelihood estimator proposed by Zeng and Lin (2007) has also been claimed to achieve the semiparametric efficiency bound, however, their approach is less intuitive and not easy to implement numerically. Moreover, their method is restricted to the logarithm transformation only and can be more difficult to implement for general monotone transformations.

From our simulation studies, it looks that the standard errors of either the Gehan-weighted or the Log-rank estimators are close to the standard errors of the proposed method. We recommend that when the sample size is small, one can compute both the Gehan-weighted and the Log-rank estimators, and choose the one with a smaller standard error. While for the large sample size, the proposed method is preferred since it not only yields statistical efficient estimators but also estimates the variance matrix faster.

By providing a statistically efficient and computationally feasible estimating procedures, this work makes the semiparametric linear model a more viable alternative to the Cox proportional hazards model. In some applications, censoring occurs just because the equipment cannot detect values under or above certain thresholds. For such type of censored data, the semiparametric linear model is particularly more attractive since in this context, the concept of hazard is not relevant. The left censored data can be turned to the right censored data by using $-T$ and $-C$ as the failure and censoring variables when applying the semiparametric linear model.

3.8 Appendix: Proofs of the Technical Results

This section contains the proofs for Theorems 3.3.1 and 3.3.2. Some empirical process theorems developed in van der Vaart and Wellner (1996) and van der Vaart (1998) will be heavily involved. We use the symbol $\lesssim$ to denote that the left hand side is bounded above by a constant times the right hand side and $\gtrsim$ to denote that the left hand side is bounded below by a constant times the right hand side. For notational simplicity, we drop the superscript * in the outer probability measure P^* whenever an outer probability applies.

3.8.1 Technical Lemmas

We first introduce several lemmas that will be used for the proofs of Theorems 3.3.1 and 3.3.2.

Lemma A.1. Under Conditions C.1-C.3 and C.6, the log-likelihood

$$l(\beta, g; Z) = \Delta g(Y - X'\beta) - \int 1(Y \geq t) \exp\{g(t - X'\beta)\}\, dt$$

has bounded and continuous first and second derivatives of $l(\beta, g; Z)$ with respect to $\beta \in \mathcal{B}$ and $g \in \mathcal{H}^p$.

Proof: This result follows by direct calculations. By the definition of the functional derivatives in section 3.6, we can obtain all the first and second derivatives of $l(\beta, g; Z)$ with respect to β and g as follows:

$$\begin{aligned}
&\dot{l}_\beta(\beta, g; Z) \\
&\quad = -X\bigg\{\Delta\dot{g}(Y - X'\beta) - \int 1(Y \geq t)\exp\{g(t - X'\beta)\}\dot{g}(t - X'\beta)\ dt\bigg\}, \\
&\dot{l}_g(\beta, g; Z)[h] = \frac{\partial}{\partial\eta} l(\beta, g + \eta h; Z)|_{\eta=0} \\
&\quad = \Delta h(Y - X'\beta) - \int 1(Y \geq t)\exp\{g(t - X'\beta)\}h(t - X'\beta)\ dt, \\
&\ddot{l}_{\beta\beta}(\beta, g; Z) \\
&\quad = XX'\bigg\{\Delta\ddot{g}(Y - X'\beta) - \int 1(Y \geq t)\exp\{g(t - X'\beta)\}\big[\ddot{g}(t - X'\beta) \\
&\qquad + \dot{g}^2(t - X'\beta)\big]\ dt\bigg\}, \\
&\ddot{l}_{\beta g}(\beta, g; Z)[h] = \ddot{l}'_{g\beta}(\beta, g; Z)[h] \\
&\quad = -X\bigg\{\Delta\dot{h}(Y - X'\beta) - \int 1(Y \geq t)\exp\{g(t - X'\beta)\}\big[\dot{h}(t - X'\beta) \\
&\qquad + \dot{g}(t - X'\beta)h(t - X'\beta)\big]\ dt\bigg\}, \\
&\ddot{l}_{gg}(\beta, g; Z)[h_1, h_2] \\
&\quad = -\int 1(Y \geq t)\exp\{g(t - X'\beta)\}h_1(t - X'\beta)h_2(t - X'\beta)\ dt,
\end{aligned}$$

where $h \in \mathbf{H} = \{h : h = \frac{\partial g_\eta}{\partial\eta}|_{\eta=0}, g_\eta \in \mathcal{H}^p\}$. All the above derivatives are continuous and bounded due to Conditions C.1-C.3 and C.6.

Lemma A.2. For $g_0 \in \mathcal{H}^p$, there exists a $g_{0,n} \in \mathcal{H}_n^p$ such that

$$\|g_{0,n} - g_0\|_\infty = O(n^{-p\nu}).$$

Proof: This is a direct result according to Corollary 6.21 of Schumaker (1981), that is, there exists a $g_{0,n} \in \mathcal{H}_n^p$ such that $\|g_{0,n} - g_0\|_\infty = O(q_n^{-p}) = O(n^{-p\nu})$.

Lemma A.3. Let $\theta_{0,n} = (\beta_0, g_{0,n})$ with $g_{0,n}$ being defined in Lemma A.2. Denote $\mathcal{F}_n = \{l(\theta; Z) - l(\theta_{0,n}; Z) : \theta \in \Theta_n^p\}$. Assume that Conditions C.1-C.3 and C.6 hold, then the ε-bracketing number associated with $\| \cdot \|_\infty$ norm for $\mathcal{F}_n$ is bounded by $(1/\varepsilon)^{cq_n+d}$, i.e., $N_{[\,]}(\varepsilon, \mathcal{F}_n, \| \cdot \|_\infty) \lesssim (1/\varepsilon)^{cq_n+d}$ for some constant $c > 0$.

Proof: By the calculation of Shen and Wong (1994) in page 597, denote $\lceil x \rceil$ as the ceiling of the number x, then for any $\varepsilon > 0$, there exists a set of brackets $\{[g_i^L, g_i^U] : i = 1, 2, \cdots, \lceil (1/\varepsilon)^{c_1 q_n} \rceil\}$ such that for any $g \in \mathcal{H}_n^p$, $g_i^L(t) \le g(t) \le g_i^U(t)$ for some $1 \le i \le \lceil (1/\varepsilon)^{c_1 q_n} \rceil$ and all $t \in [a, b]$, and $\|g_i^U - g_i^L\|_\infty \le \varepsilon$. Since $\mathcal{B} \subseteq \mathbb{R}^d$ is compact, $\mathcal{B}$ can be covered by $\lceil c_2 (1/\varepsilon)^d \rceil$ balls with radius ε; that is, for any $\beta \in \mathcal{B}$, there exists an $1 \le s \le \lceil c_2 (1/\varepsilon)^d \rceil$ such that $|\beta - \beta_s| \le \varepsilon$ and hence $|x'(\beta - \beta_s)| \le C\varepsilon$ for any $x \in \mathcal{X}$ because of Condition C.2(a), where $C > 0$ is a constant. This indicates that $t - x'\beta \in [t - x'\beta_s - C\varepsilon, t - x'\beta_s + C\varepsilon]$ for any x and t. Assume $g_i^L(t - x'\beta_s + c_1^{i,t}\varepsilon)$ and $g_i^U(t - x'\beta_s + c_2^{i,t}\varepsilon)$ are the minimum and maximum values of g_i^L and g_i^U within the interval $[t - x'\beta_s - C\varepsilon, t - x'\beta_s + C\varepsilon]$, where $c_1^{i,t}$ and $c_2^{i,t}$ are two constants that only depend on g_i^L, g_i^U and t with $|c_1^{i,t}|, |c_2^{i,t}| \le C$. So we have

$$g_i^L(t - x'\beta_s + c_1^{i,t}\varepsilon) \le g_i^L(t - x'\beta) \le g(t - x'\beta) \le g_i^U(t - x'\beta) \le g_i^U(t - x'\beta_s + c_2^{i,t}\varepsilon).$$

Hence we can construct a set of brackets

$$\left\{[m_{i,s}^L(Z), m_{i,s}^U(Z)] : i = 1, \cdots, \lceil (1/\varepsilon)^{c_1 q_n} \rceil;\ s = 1, \cdots, \lceil c_2 (1/\varepsilon)^d \rceil\right\}$$

that for any $m(\theta; Z) \in \mathcal{F}_n$, there exists a pair (i, s) such that for any sample point

Z, $m(\theta;Z) \in [m^L_{i,s}(Z), m^U_{i,s}(Z)]$, where

$$\begin{aligned} m^L_{i,s}(Z) &= \left\{\Delta g^L_i(Y - X'\beta_s + c^{i,Y}_1\varepsilon) - \int 1(Y \geq t)\exp\{g^U_i(t - X'\beta_s + c^{i,t}_2\varepsilon)\}\, dt\right\} \\ &\quad - l(\theta_{0,n};Z) \\ &= \left\{\Delta g^L_i(\epsilon_s + c^{i,Y}_1\varepsilon) - \int_a^b 1(\epsilon_0 \geq t)\exp\{g^U_i(t_s + c^{i,t}_2\varepsilon)\}\, dt\right\} - l(\theta_{0,n};Z) \end{aligned}$$

and

$$\begin{aligned} m^U_{i,s}(Z) &= \left\{\Delta g^U_i(Y - X'\beta_s + c^{i,Y}_2\varepsilon) - \int 1(Y \geq t)\exp\{g^L_i(t - X'\beta_s + c^{i,t}_1\varepsilon)\}\, dt\right\} \\ &\quad - l(\theta_{0,n};Z) \\ &= \left\{\Delta g^U_i(\epsilon_s + c^{i,Y}_2\varepsilon) - \int_a^b 1(\epsilon_0 \geq t)\exp\{g^L_i(t_s + c^{i,t}_1\varepsilon)\}\, dt\right\} - l(\theta_{0,n};Z), \end{aligned}$$

with

$$\epsilon_s = Y - X'\beta_s,\ \epsilon_0 = Y - X'\beta_0 \text{ and } t_s = t - X'(\beta_s - \beta_0) \tag{3.19}$$

for notational simplicity. It then follows that

$$\begin{aligned} |m^U_{i,s}(Z) - m^L_{i,s}(Z)| &\leq |g^U_i(\epsilon_s + c^{i,Y}_2\varepsilon) - g^L_i(\epsilon_s + c^{i,Y}_1\varepsilon)| \\ &\qquad + \int_a^b |\exp\{g^U_i(t_s + c^{i,t}_2\varepsilon)\} - \exp\{g^L_i(t_s + c^{i,t}_1\varepsilon)\}|\, dt \\ &= A_1 + A_2. \end{aligned}$$

For A_1, by subtracting and adding the terms $g(\epsilon_s + c^{i,Y}_2\varepsilon)$ and $g(\epsilon_s + c^{i,Y}_1\varepsilon)$ and applying the Taylor expansion to g at $\epsilon_s + c^{i,Y}_1\varepsilon$, we have

$$\begin{aligned} A_1 &\leq |g^U_i(\epsilon_s + c^{i,Y}_2\varepsilon) - g(\epsilon_s + c^{i,Y}_2\varepsilon)| + |g(\epsilon_s + c^{i,Y}_2\varepsilon) - g(\epsilon_s + c^{i,Y}_1\varepsilon)| + \\ &\quad + |g(\epsilon_s + c^{i,Y}_1\varepsilon) - g^L_i(\epsilon_s + c^{i,Y}_1\varepsilon)| \\ &\leq \|g^U_i - g\|_\infty + |\dot{g}(\epsilon_s + \tilde{c}\varepsilon)(c^{i,Y}_2 - c^{i,Y}_1)\varepsilon| + \|g - g^L_i\|_\infty \\ &\lesssim \|g^U_i - g^L_i\|_\infty + |(c^{i,Y}_2 - c^{i,Y}_1)|\varepsilon \\ &\lesssim \varepsilon + 2C\varepsilon \lesssim \varepsilon, \end{aligned}$$

where the third inequality holds because $\|g_i^U - g\|_\infty, \|g - g_i^L\|_\infty \le \|g_i^U - g_i^L\|_\infty$ and $\dot{g}$ is bounded. For A_2, by using the similar arguments as for A_1, we have

$$\begin{aligned} A_2 &\le \int_a^b \{|\exp\{g_i^U(t_s + c_2^{i,t}\varepsilon)\} - \exp\{g(t_s + c_2^{i,t}\varepsilon)\}| \\ &\quad + |\exp\{g(t_s + c_2^{i,t}\varepsilon)\} - \exp\{g(t_s + c_1^{i,t}\varepsilon)\}| \\ &\quad + |\exp\{g(t_s + c_1^{i,t}\varepsilon)\} - \exp\{g_i^L(t_s + c_1^{i,t}\varepsilon)\}|\} \, dt \\ &= \int_a^b \{|\exp\{\tilde{g}_i^U(t_s + c_2^{i,t}\varepsilon)\}(g_i^U - g)(t_s + c_2^{i,t}\varepsilon)| \\ &\quad + |\exp\{g(t_s + \tilde{c}\varepsilon)\}(c_2^{i,t} - c_1^{i,t})\varepsilon| \\ &\quad + |\exp\{\tilde{g}_i^L(t_s + c_1^{i,t}\varepsilon)\}(g_i^L - g)(t_s + c_1^{i,t}\varepsilon)|\} \, dt \\ &\lesssim \|g_i^U - g\|_\infty + |(c_2^{i,t} - c_1^{i,t})\varepsilon| + \|g - g_i^L\|_\infty \lesssim \varepsilon. \end{aligned}$$

The first equality above holds because $\tilde{g}_i^U = g + \xi(g_i^U - g)$ for some $0 < \xi < 1$ and thus $|\tilde{g}_i^U(t)| \le |g(t)| + \varepsilon$, which is bounded in $[a, b]$, and similarly for $\tilde{g}_i^L$. Hence $\|m_i^U - m_i^L\|_\infty \lesssim \varepsilon$ and the ε-bracketing number associated with $\|\cdot\|_\infty$ norm for the class $\mathcal{F}_n$ follows

$$N_{[\,]}(\varepsilon, \mathcal{F}_n, \|\cdot\|_\infty) \le (1/\varepsilon)^{c_1 q_n} c_2 (1/\varepsilon)^d \lesssim (1/\varepsilon)^{c_1 q_n + d}.$$

Lemma A.4. Let $h_j^*(t) = -\dot{g}_0(t)P(X_j|\epsilon_0 \ge t)$, $j = 1, \cdots, d$. This is the least favorable direction for the score function of the nonparametric component g_0, which will be shown in the proof of Theorem 3.3.2. Assume Conditions C.1-C.6 hold, then there exists an $h_{j,n}^* \in \mathcal{H}_n^2$ such that $\|h_{j,n}^* - h_j^*\|_\infty = O(n^{-2\nu})$.

Proof: By Conditions C.4-C.5, the conditional density of ϵ_0 given X, i.e.,

$$f_{\epsilon_0|X}(t|X = x) = f(t)\bar{G}_{C|X}(t + x'\beta_0|X = x) + g_{C|X}(t + x'\beta_0|X = x)\bar{F}(t)$$

is uniformly bounded for all $x \in \mathcal{X}$, and its derivative with respect to t, that is,

$$\begin{aligned} \dot{f}_{\varepsilon_0|X}(t|X = x) &= \dot{f}(t)\bar{G}_{C|X}(t + x'\beta_0|X = x) - f(t)g_{C|X}(t + x'\beta_0|X = x) \\ &\quad + \dot{g}_{C|X}(t + x'\beta_0|X = x)\bar{F}(t) - g_{C|X}(t + x'\beta_0|X = x)f(t) \end{aligned}$$

is also uniformly bounded. Hence the density of ϵ_0

$$f_{\epsilon_0}(t) = \int_{\mathcal{X}} f_{\epsilon_0|X}(t|X = x) f_X(x)\ dx$$

and its derivative

$$\dot{f}_{\epsilon_0}(t) = \int_{\mathcal{X}} \dot{f}_{\epsilon_0|X}(t|X = x) f_X(x)\ dx$$

are bounded. Thus the first and second derivatives of $P(\epsilon_0 \geq t)$, i.e., $-f_{\epsilon_0}(t)$ and $-\dot{f}_{\epsilon_0}(t)$, are both bounded. In addition, under Condition C.2(a), the first and second derivatives of $P[X1(\epsilon_0 \geq t)]$ with respect to t, i.e.,

$$\frac{dP[X1(\epsilon_0 \geq t)]}{dt} = -\int_{\mathcal{X}} x f_X(x) f_{\epsilon_0|X}(t|X = x)\ dx$$

and

$$\frac{d^2P[X1(\epsilon_0 \geq t)]}{dt^2} = -\int_{\mathcal{X}} x f_X(x) \dot{f}_{\epsilon_0|X}(t|X = x)\ dx$$

are also bounded. Therefore, $P[X|\epsilon_0 \geq t] = P[X1(\epsilon_0 \geq t)]/P(\epsilon_0 \geq t)$ has a bounded second derivative with respect to t for $t \leq \tau$, where τ is the truncation time defined in Condition C.3. Thus $P[X|\epsilon_0 \geq t] \in \mathcal{H}^2$. Moreover, since $g_0 \in \mathcal{H}^p$ for $p \geq 3$, we have $\dot{g}_0(t) \in \mathcal{H}^{p-1}$ with $p - 1 \geq 2$. Thus according to Corollary 6.21 of Schumaker (1981), there exists an $h^*_{j,n} \in \mathcal{H}_n^{\min(p-1,2)} = \mathcal{H}_n^2$ such that $\|h^*_j - h^*_{j,n}\|_\infty = O(q_n^{-2}) = O(n^{-2\nu})$.

Lemma A.5. Let h^*_j, $j = 1, \cdots, d$ be the function defined in Lemma A.4 and denote the class of function

$$\mathcal{F}_n^j(\eta) = \{\dot{l}_g(\theta; z)[h^*_j - h] : \theta \in \Theta_n^p, h \in \mathcal{H}_n^2 \text{ and } d(\theta, \theta_0) \leq \eta, \|h - h^*_j\|_\infty \leq \eta\}.$$

Assume Conditions C.1-C.6 hold, then $N_{[\,]}(\varepsilon, \mathcal{F}_n^j(\eta), \|\cdot\|_\infty) \lesssim (\eta/\varepsilon)^{cq_n+d}$ for some constant $c > 0$.

Proof: We shall use the similar argument for the bracketing number of the class

$\mathcal{F}_n$ in Lemma A.3. First, define the classes of functions

$$\begin{aligned}
\mathcal{H}_n^p(\eta) &= \{g \in \mathcal{H}_n^p, \|g - g_0\|_2 \leq \eta\}, \\
\mathcal{H}_{n,j}^2(\eta) &= \{h \in \mathcal{H}_n^2, \|h - h_j^*\|_\infty \leq \eta\}, \quad \text{and} \\
\mathcal{B}(\eta) &= \{\beta \in \mathcal{B} \subseteq R^d, |\beta - \beta_0| \leq \eta\},
\end{aligned}$$

then it follows by the calculation of Shen and Wong (1994) in page 597 that

$$N_{[\,]}(\varepsilon, \mathcal{H}_n^p(\eta), \|\cdot\|_\infty) \leq (\eta/\varepsilon)^{c_1 q_n} \text{ and } N_{[\,]}(\varepsilon, \mathcal{H}_{n,j}^2(\eta), \|\cdot\|_\infty) \leq (\eta/\varepsilon)^{c_2 q_n}$$

for some constants c_1, $c_2 > 0$. In addition, since $\mathcal{B} \subseteq R^d$ is compact, the covering number for $\mathcal{B}(\eta)$ follows $N(\varepsilon, \mathcal{B}(\eta), \|\cdot\|_\infty) \leq c_3(\eta/\varepsilon)^d$.

Then as in the proof of Lemma A.3, let ϵ_s, ϵ_0 and t_s be defined in (3.19) for notational simplicity. g_i^L and g_i^U are functions that bracket g with $\|g_i^U - g_i^L\|_\infty \leq \varepsilon$; h_k^L and h_k^U are functions that bracket h with $\|h_k^U - h_k^L\|_\infty \leq \varepsilon$; and β_s satisfies $t - x'\beta \in [t - x'\beta_s - C\varepsilon, t - x'\beta_s + C\varepsilon] = [t_s - C\varepsilon, t_s + C\varepsilon]$ for any x, t and some constant $C > 0$. Moreover, let $c_1^{j,k,t}$ and $c_2^{j,k,t}$ be two constants that $(h_j^* - h_k^U)(t_s + c_1^{j,k,t}\varepsilon)$ and $(h_j^* - h_k^L)(t_s + c_2^{j,k,t}\varepsilon)$ are the minimum and maximum values of $(h_j^* - h_k^U)$ and $(h_j^* - h_k^L)$ in $[t_s - C\varepsilon, t_s + C\varepsilon]$, respectively; and let $c_3^{i,t}$ and $c_4^{i,t}$ be two constants that $g_i^L(t_s + c_3^{i,t}\varepsilon)$ and $g_i^U(t_s + c_4^{i,t}\varepsilon)$ are the minimum and maximum values of g_i^L and g_i^U in $[t_s - C\varepsilon, t_s + C\varepsilon]$, respectively. Then we can construct a set of brackets for $\mathcal{F}_n^j(\eta)$ as

$$\left\{[d_{i,k,s}^L(Z),\ d_{i,k,s}^U(Z)] : 1 \leq i \leq \lceil(\eta/\varepsilon)^{c_1 q_n}\rceil;\ 1 \leq k \leq \lceil(\eta/\varepsilon)^{c_2 q_n}\rceil;\ 1 \leq s \leq \lceil c_3(\eta/\varepsilon)^d\rceil\right\}$$

that for any $\dot{l}_g(\theta; Z)[h_j^* - h] \in \mathcal{F}_n^j(\eta)$, there exists a triplet (i, k, s) such that

$$\dot{l}_g(\theta; Z)[h_j^* - h] \in [d_{i,k,s}^L(Z),\ d_{i,k,s}^U(Z)]$$

for any sample point Z, where

$$
\begin{aligned}
d^L_{i,k,s}(Z) &= \Delta(h^*_j - h^U_k)(\epsilon_s + c_1^{j,k,Y}\varepsilon) \\
&- \int_a^b 1(\epsilon_0 \geq t) e^{g_i^U(t_s + c_4^{i,t}\varepsilon)}(h^*_j - h^L_k)(t_s + c_2^{j,k,t}\varepsilon)1\{(h^*_j - h)(t_s + c_2^{j,k,t}\varepsilon) \geq 0\}\ dt \\
&- \int_a^b 1(\epsilon_0 \geq t) e^{g_i^L(t_s + c_3^{i,t}\varepsilon)}(h^*_j - h^L_k)(t_s + c_2^{j,k,t}\varepsilon)1\{(h^*_j - h)(t_s + c_2^{j,k,t}\varepsilon) < 0\}\ dt
\end{aligned}
$$

and

$$
\begin{aligned}
d^U_{i,k,s}(Z) &= \Delta(h^*_j - h^L_k)(\epsilon_s + c_2^{j,k,Y}\varepsilon) \\
&- \int_a^b 1(\epsilon_0 \geq t) e^{g_i^L(t_s + c_3^{i,t}\varepsilon)}(h^*_j - h^U_k)(t_s + c_1^{j,k,t}\varepsilon)1\{(h^*_j - h)(t_s + c_1^{j,k,t}\varepsilon) \geq 0\}\ dt \\
&- \int_a^b 1(\epsilon_0 \geq t) e^{g_i^U(t_s + c_4^{i,t}\varepsilon)}(h^*_j - h^U_k)(t_s + c_1^{j,k,t}\varepsilon)1\{(h^*_j - h)(t_s + c_1^{j,k,t}\varepsilon) < 0\}\ dt.
\end{aligned}
$$

Then it follows that

$$
\begin{aligned}
&|d^U_{i,k,s}(Z) - d^L_{i,k,s}(Z)| \leq |(h^*_j - h^L_k)(\epsilon_s + c_2^{j,k,t}\varepsilon) - (h^*_j - h^U_k)(\epsilon_s + c_1^{j,k,t}\varepsilon)| \\
&+ \int_a^b \Big| e^{g_i^U(t_s + c_4^{i,t}\varepsilon)}(h^*_j - h^L_k)(t_s + c_2^{j,k,t}\varepsilon)1\{(h^*_j - h)(t_s + c_2^{j,k,t}\varepsilon) \geq 0\} \\
&- e^{g_i^L(t_s + c_3^{i,t}\varepsilon)}(h^*_j - h^U_k)(t_s + c_1^{j,k,t}\varepsilon)1\{(h^*_j - h)(t_s + c_1^{j,k,t}\varepsilon) \geq 0\}\Big|\ dt \\
&+ \int_a^b \Big| e^{g_i^L(t_s + c_3^{i,t}\varepsilon)}(h^*_j - h^L_k)(t_s + c_2^{j,k,t}\varepsilon)1\{(h^*_j - h)(t_s + c_2^{j,k,t}\varepsilon) < 0\} \\
&- e^{g_i^U(t_s + c_4^{i,t}\varepsilon)}(h^*_j - h^U_k)(t_s + c_1^{j,k,t}\varepsilon)1\{(h^*_j - h)(t_s + c_1^{j,k,t}\varepsilon) < 0\}\Big|\ dt \\
&= A_1 + A_2 + A_3.
\end{aligned}
$$

For term A_1, by subtracting and adding the terms $h^L_k(\epsilon_s + c_1^{j,k,t}\varepsilon)$, $h(\epsilon_s + c_1^{j,k,t}\varepsilon)$ and $h(\epsilon_s + c_2^{j,k,t}\varepsilon)$, and applying the Taylor expansions for h^*_j and h, we have

$$
\begin{aligned}
A_1 \ \leq\ & |h^*_j(\epsilon_s + c_2^{j,k,t}\varepsilon) - h^*_j(\epsilon_s + c_1^{j,k,t}\varepsilon)| + |h^U_k(\epsilon_s + c_1^{j,k,t}\varepsilon) - h^L_k(\epsilon_s + c_1^{j,k,t}\varepsilon)| \\
& + |h^L_k(\epsilon_s + c_1^{j,k,t}\varepsilon) - h(\epsilon_s + c_1^{j,k,t}\varepsilon)| + |h(\epsilon_s + c_1^{j,k,t}\varepsilon) - h(\epsilon_s + c_2^{j,k,t}\varepsilon)| \\
& + |h(\epsilon_s + c_2^{j,k,t}\varepsilon) - h^L_k(\epsilon_s + c_2^{j,k,t}\varepsilon)| \\
\lesssim\ & |c_2^{j,k,t} - c_1^{j,k,t}|\varepsilon + \|h^U_k - h^L_k\|_\infty + \|h^L_k - h\|_\infty + |c_2^{j,k,t} - c_1^{j,k,t}|\varepsilon + \|h - h^L_k\|_\infty \\
\lesssim\ & |c_2^{j,k,t} - c_1^{j,k,t}|\varepsilon + \|h^U_k - h^L_k\|_\infty \lesssim \varepsilon.
\end{aligned}
$$

Next for A_2, first, by subtracting and adding the term $e^{g_i^U(t_s+c_4^{k,t}\varepsilon)}(h_j^*-h_k^U)(t_s+c_1^{j,k,t}\varepsilon)$ we have

$$\begin{aligned}
A_2' &= \int_a^b |e^{g_i^U(t_s+c_4^{i,t}\varepsilon)}(h_j^*-h_k^L)(t_s+c_2^{j,k,t}\varepsilon) - e^{g_i^L(t_s+c_3^{i,t}\varepsilon)}(h_j^*-h_k^U)(t_s+c_1^{j,k,t}\varepsilon)|\ dt \\
&\leq \int_a^b e^{g_i^U(t_s+c_4^{i,t}\varepsilon)}\big|(h_j^*-h_k^L)(t_s+c_2^{j,k,t}\varepsilon) - (h_j^*-h_k^U)(t_s+c_1^{j,k,t}\varepsilon)\big|\ dt \\
&\qquad + \int_a^b \big|\{e^{g_i^U(t_s+c_4^{i,t}\varepsilon)} - e^{g_i^L(t_s+c_3^{i,t}\varepsilon)}\}(h_j^*-h_k^U)(t_s+c_1^{j,k,t}\varepsilon)\big|\ dt \\
&= A_4 + A_5.
\end{aligned}$$

By the above result for A_1, since $\|g_i^U\|_\infty \leq \|g\|_\infty + \|g_i^U - g\|_\infty \leq \|g\|_\infty + \varepsilon$, which is bounded, we have $A_4 \leq \|A_1\|_\infty \int_a^b e^{g_i^U(t_s+c_4^{i,t}\varepsilon)}\ dt \lesssim \varepsilon$. Then for A_5, using the similar argument for A_1, it follows that

$$\begin{aligned}
A_5 &\leq \|h_j^*-h_k^U\|_\infty \int_a^b \big\{|e^{g_i^U(t_s+c_4^{i,t}\varepsilon)} - e^{g(t_s+c_4^{i,t}\varepsilon)}| + |e^{g(t_s+c_4^{i,t}\varepsilon)} - e^{g(t_s+c_3^{i,t}\varepsilon)}| \\
&\qquad + |e^{g(t_s+c_3^{i,t}\varepsilon)} - e^{g_i^L(t_s+c_3^{i,t}\varepsilon)}|\big\}\ dt \\
&\leq (\|h_j^*-h\|_\infty + \|h-h_k^U\|_\infty)\bigg\{\int_a^b \big[e^{\tilde{g}_i^U(t_s+c_4^{i,t}\varepsilon)}|(g_i^U-g)(t_s+c_4^{i,t}\varepsilon)| \\
&\qquad + e^{g(t_s+\tilde{c}\varepsilon)}|(c_4^{i,t}-c_3^{i,t})\varepsilon| + e^{\tilde{g}_i^L(t_s+c_3^{i,t}\varepsilon)}|(g-g_i^L)(t_s+c_3^{i,t}\varepsilon)|\big]\ dt\bigg\} \\
&\lesssim (\eta+\varepsilon)\{\|g_i^U-g\|_\infty + |c_4^{i,t}-c_3^{i,t}|\varepsilon + \|g-g_i^L\|_\infty\} \\
&\lesssim (\eta+\varepsilon)\varepsilon \lesssim \varepsilon,
\end{aligned}$$

where $\tilde{g}_i^U = g + \xi_1(g_i^U - g)$ and $\tilde{g}_i^L = g + \xi_2(g_i^L - g)$ for some constants $\xi_1, \xi_2 \in (0,1)$, and $\tilde{c} = c_3^{i,t} + \xi(c_4^{i,t} - c_3^{i,t})$ for some $\xi \in (0,1)$. Thus $A_2' \lesssim \varepsilon$. Therefore,

$$\begin{aligned} A_2 &\leq \int_a^b |e^{g_i^U(t_s + c_4^{i,t}\varepsilon)}(h_j^* - h_k^L)(t_s + c_2^{j,k,t}\varepsilon) - e^{g_i^L(t_s + c_3^{i,t}\varepsilon)}(h_j^* - h_k^U)(t_s + c_1^{j,k,t}\varepsilon)| \\ &\qquad \cdot 1\{(h_j^* - h)(t_s + c_2^{j,k,t}\varepsilon) \geq 0\} \ dt \\ &\quad + \int_a^b e^{g_i^L(t_s + c_3^{i,t}\varepsilon)} |(h_j^* - h_k^U)(t_s + c_1^{j,k,t}\varepsilon)| \\ &\qquad \cdot |1\{(h_j^* - h)(t_s + c_2^{j,k,t}\varepsilon) \geq 0\} - 1\{(h_j^* - h)(t_s + c_1^{j,k,t}\varepsilon) \geq 0\}| \ dt \\ &\lesssim A_2' + \int_a^b |1\{(h_j^* - h)(t_s + c_2^{j,k,t}\varepsilon) \geq 0\} - 1\{(h_j^* - h)(t_s + c_1^{j,k,t}\varepsilon) \geq 0\}| \ dt \\ &\lesssim \varepsilon + |(c_2^{j,k,t} - c_1^{j,k,t})\varepsilon| \lesssim \varepsilon, \end{aligned}$$

where the second inequality holds since $\|g_i^L\|_\infty \leq \|g\|_\infty + \|g_i^L - g\|_\infty \leq \|g\|_\infty + \varepsilon$ and thus is bounded, then

$$\|e^{g_i^L(t_s + c_3^{i,t}\varepsilon)}(h_j^* - h_k^U)(t_s + c_1^{j,k,t}\varepsilon)\|_\infty \leq \|e^{g_i^L}\|_\infty(\|h_j^* - h\|_\infty + \|h - h_k^U\|_\infty) \lesssim (\eta + \varepsilon),$$

which is bounded. Finally, by using the same argument as for A_2, we can also show that $A_3 \lesssim \varepsilon$. Hence $\|d_{i,k,s}^U(Z) - d_{i,k,s}^L(Z)\|_\infty \lesssim \varepsilon$. Therefore, the ε-bracketing number for the class $\mathcal{F}_n^j(\eta)$ is bounded by $(\eta/\varepsilon)^{c_1 q_n}(\eta/\varepsilon)^{c_2 q_n} c_3 (\eta/\varepsilon)^d$, that is,

$$N_{[\,]}(\varepsilon, \mathcal{F}_n^j(\eta), \|\cdot\|_\infty) \lesssim (\eta/\varepsilon)^{(c_1+c_2)q_n + d}.$$

Lemma A.6. For $j = 1, \cdots, d$, define the two classes of functions as

$$\mathcal{F}_{n,j}^\beta(\eta) = \{\dot{l}_{\beta_j}(\theta; z) - \dot{l}_{\beta_j}(\theta_0; z) : \theta \in \Theta_n^p, \dot{g} \in \mathcal{H}_n^{p-1} \text{ and } d(\theta, \theta_0) \leq \eta, \|\dot{g} - \dot{g}_0\|_2 \leq \eta\},$$

and

$$\mathcal{F}_{n,j}^g(\eta) = \{\dot{l}_g(\theta; z)[h_j^*] - \dot{l}_g(\theta_0; z)[h_j^*] : \theta \in \Theta_n^p \text{ and } d(\theta, \theta_0) \leq \eta\},$$

where $\dot{l}_{\beta_j}(\theta; Z)$ is the jth element of $\dot{l}_\beta(\theta; Z)$ and h_j^* is defined in Lemma A.4. Assume Conditions C.1-C.6 hold, then $N_{[\,]}(\varepsilon, \mathcal{F}_{n,j}^\beta(\eta), \|\cdot\|_\infty) \lesssim (\eta/\varepsilon)^{c_1 q_n + d}$ and $N_{[\,]}(\varepsilon, \mathcal{F}_{n,j}^g(\eta), \|\cdot\|_\infty) \lesssim (\eta/\varepsilon)^{c_2 q_n + d}$ for some constants $c_1, c_2 > 0$.

Proof: First, for the following classes of functions

$$\begin{aligned}\mathcal{H}_n^p(\eta) &= \{g \in \mathcal{H}_n^p, \|g - g_0\|_2 \leq \eta\},\\ \mathcal{H}_n^{p-1}(\eta) &= \{\dot{g} \in \mathcal{H}_n^{p-1}, \|\dot{g} - \dot{g}_0\|_2 \leq \eta\} \text{ and}\\ \mathcal{B}(\eta) &= \{\beta \in \mathcal{B} \subseteq R^d, |\beta - \beta_0| \leq \eta\},\end{aligned}$$

by the same argument in the proof of Lemma A.5, we have

$$N_{[\ ]}(\varepsilon, \mathcal{H}_n^p(\eta), \|\cdot\|_\infty) \leq (\eta/\varepsilon)^{c_1 q_n}, N_{[\ ]}(\varepsilon, \mathcal{H}_n^{p-1}(\eta), \|\cdot\|_\infty) \leq (\eta/\varepsilon)^{c_2 q_n},$$

and $N(\varepsilon, \mathcal{B}(\eta), \|\cdot\|_\infty) \leq c_3(\eta/\varepsilon)^d$ for some constants $c_1, c_2, c_3 > 0$.

Then using the similar argument as in the proof of Lemma A.5, let g_i^L and g_i^U be functions that bracket g with $\|g_i^U - g_i^L\|_\infty \leq \varepsilon$; $\dot{g}_k^L$ and $\dot{g}_k^U$ be functions that bracket $\dot{g}$ with $\|\dot{g}_k^U - \dot{g}_k^L\|_\infty \leq \varepsilon$; and let β_s satisfy $t - x'\beta \in [t - x'\beta_s - C\varepsilon, t - x'\beta_s + C\varepsilon] = [t_s - C\varepsilon, t_s + C\varepsilon]$ for any x, t and some constant $C > 0$. Moreover, let $c_1^{k,t}$ and $c_2^{k,t}$ be two constants that $\dot{g}_k^L(t_s + c_1^{k,t}\varepsilon)$ and $\dot{g}_k^U(t_s + c_2^{k,t}\varepsilon)$ are the minimum and maximum values of $\dot{g}_k^L$ and $\dot{g}_k^U$ in $[t_s - C\varepsilon, t_s + C\varepsilon]$, respectively; let $c_3^{i,t}$ and $c_4^{i,t}$ be two constants that $g_i^L(t_s + c_3^{i,t}\varepsilon)$ and $g_i^U(t_s + c_4^{i,t}\varepsilon)$ are the minimum and maximum values of g_i^L and g_i^U in $[t_s - C\varepsilon, t_s + C\varepsilon]$, respectively; and let $c_5^{j,t}$ and $c_6^{j,t}$ be two constants that $h_j^*(t_s + c_5^{j,t}\varepsilon)$ and $h_j^*(t_s + c_6^{j,t}\varepsilon)$ are the minimum and maximum values of h_j^* in $[t_s - C\varepsilon, t_s + C\varepsilon]$. Then we can similarly construct a set of brackets for $\mathcal{F}_{n,j}^\beta(\eta)$ as

$$\left\{[u_{i,k,s}^L(Z),\ u_{i,k,s}^U(Z)] : 1 \leq i \leq \lceil(\eta/\varepsilon)^{c_1 q_n}\rceil;\ 1 \leq k \leq \lceil(\eta/\varepsilon)^{c_2 q_n}\rceil;\ 1 \leq s \leq \lceil c_3(\eta/\varepsilon)^d\rceil\right\}$$

that for any element in $\mathcal{F}_{n,j}^\beta(\eta)$, there exists a triplet (i, k, s) such that

$$\dot{l}_{\beta_j}(\theta; Z) - \dot{l}_{\beta_j}(\theta_0; Z) \in [u_{i,k,s}^L(Z),\ u_{i,k,s}^U(Z)]$$

for any sample point Z (without loss of generality, assume X_j is nonnegative, for negative X_j, just switch the lower and upper brackets), where

$$\begin{aligned} u^L_{i,k,s}(Z) = -X_j \Big\{ & \Delta \dot{g}^U_k(\epsilon_s + c_2^{k,Y}\varepsilon) \\ & - \int_a^b 1(\epsilon_0 \geq t) e^{g_i^L(t_s + c_3^{i,t}\varepsilon)} \dot{g}^L_k(t_s + c_1^{k,t}\varepsilon) 1\{\dot{g}(t_s + c_1^{k,t}\varepsilon) \geq 0\}\, dt \\ & - \int_a^b 1(\epsilon_0 \geq t) e^{g_i^U(t_s + c_4^{i,t}\varepsilon)} \dot{g}^L_k(t_s + c_1^{k,t}\varepsilon) 1\{\dot{g}(t_s + c_1^{k,t}\varepsilon) < 0\}\, dt \Big\} - \dot{l}_{\beta_j}(\beta_0, g_0; Z) \end{aligned}$$

and

$$\begin{aligned} u^U_{i,k,s}(Z) = -X_j \Big\{ & \Delta \dot{g}^L_k(\epsilon_s + c_1^{k,Y}\varepsilon) \\ & - \int_a^b 1(\epsilon_0 \geq t) e^{g_i^U(t_s + c_4^{i,t}\varepsilon)} \dot{g}^U_k(t_s + c_2^{k,t}\varepsilon) 1\{\dot{g}(t_s + c_2^{k,t}\varepsilon) \geq 0\}\, dt \\ & - \int_a^b 1(\epsilon_0 \geq t) e^{g_i^L(t_s + c_3^{i,t}\varepsilon)} \dot{g}^U_k(t_s + c_2^{k,t}\varepsilon) 1\{\dot{g}(t_s + c_2^{k,t}\varepsilon) < 0\}\, dt \Big\} - \dot{l}_{\beta_j}(\beta_0, g_0; Z), \end{aligned}$$

with ϵ_s and t_s being defined in (3.19). Also we can construct a bracket for $\mathcal{F}^g_{n,j}(\eta)$ as

$$\left\{[v^L_{i,s}(Z),\ v^U_{i,s}(Z)] : 1 \leq i \leq \lceil (\eta/\varepsilon)^{c_1 q_n} \rceil;\ 1 \leq s \leq \lceil c_3(\eta/\varepsilon)^d \rceil \right\}$$

that for any element in $\mathcal{F}^g_{n,j}(\eta)$, there exists a pair (i, s) such that

$$\dot{l}_g(\theta; Z)[h^*_j] - \dot{l}_g(\theta_0; Z)[h^*_j] \in [v^L_{i,s}(Z),\ v^U_{i,s}(Z)]$$

for any sample point Z, where

$$\begin{aligned} v^L_{i,s}(Z) = \Big\{ & \Delta h^*_j(\epsilon_s + c_5^{j,Y}\varepsilon) \\ & - \int_a^b 1(\epsilon_0 \geq t) e^{g_i^U(t_s + c_4^{i,t})} h^*_j(t_s + c_6^{j,t}\varepsilon) 1\{h^*_j(t_s + c_6^{j,t}\varepsilon) \geq 0\}\, dt \\ & - \int_a^b 1(\epsilon_0 \geq t) e^{g_i^L(t_s + c_3^{i,t})} h^*_j(t_s + c_6^{j,t}\varepsilon) 1\{h^*_j(t_s + c_6^{j,t}\varepsilon) < 0\}\, dt \Big\} - \dot{l}_g(\theta_0; Z)[h^*_j] \end{aligned}$$

and

$$
\begin{aligned}
v^U_{i,s}(Z) = \Big\{ & \Delta h^*_j(\epsilon_s + c_6^{j,Y}\varepsilon) \\
& - \int_a^b 1(\epsilon_0 \geq t) e^{g_i^L(t_s + c_3^{i,t})} h^*_j(t_s + c_5^{j,t}\varepsilon) 1\{h^*_j(t_s + c_5^{j,t}\varepsilon) \geq 0\}\ dt \\
& - \int_a^b 1(\epsilon_0 \geq t) e^{g_i^U(t_s + c_4^{i,t})} h^*_j(t_s + c_5^{j,t}\varepsilon) 1\{h^*_j(t_s + c_5^{j,t}\varepsilon) < 0\}\ dt \Big\} - \dot{l}_g(\theta_0; Z)[h^*_j].
\end{aligned}
$$

By using the similar argument for the proof of Lemma A.5, we can show that $\|u^U_{i,k,s}(Z) - u^L_{i,k,s}(Z)\|_\infty \lesssim \varepsilon$ and $\|v^U_{i,s}(Z) - v^L_{i,s}(Z)\|_\infty \lesssim \varepsilon$. Therefore,

$$
N_{[\,]}(\varepsilon, \mathcal{F}^\beta_{n,j}(\eta), \|\cdot\|_\infty) \lesssim (\eta/\varepsilon)^{c_1 q_n} (\eta/\varepsilon)^{c_2 q_n} (\eta/\varepsilon)^d = (\eta/\varepsilon)^{(c_1 + c_2) q_n + d},
$$

and

$$
N_{[\,]}(\varepsilon, \mathcal{F}^g_{n,j}(\eta), \|\cdot\|_\infty) \lesssim (\eta/\varepsilon)^{c_1 q_n} (\eta/\varepsilon)^d = (\eta/\varepsilon)^{c_1 q_n + d}.
$$

3.8.2 Proof of Theorem 3.3.1

We shall apply Theorem 1 of Shen and Wong (1994) to derive the convergence rate. First, we verify their Condition C1. Since $Pl(\beta, g; Z)$ is maximized at (β_0, g_0), its first derivative at (β_0, g_0) is equal to 0. By Lemma A.1 that all the second derivatives of $l(\beta, g; Z)$ are continuous and bounded, so the Taylor expansion yields

$$
\begin{aligned}
& Pl(\beta, g; Z) - Pl(\beta_0, g_0; Z) \\
= \ & \frac{1}{2} P\big\{(\beta - \beta_0)' \ddot{l}_{\beta\beta}(\beta_0, g_0; Z)(\beta - \beta_0) + 2(\beta - \beta_0)' \ddot{l}_{\beta g}(\beta_0, g_0; Z)[g - g_0] \\
& \quad + \ddot{l}_{gg}(\beta_0, g_0; Z)[g - g_0, g - g_0]\big\} + o(d^2(\theta, \theta_0)) \\
= \ & A + o(d^2(\theta, \theta_0)),
\end{aligned}
\tag{3.20}
$$

where $\theta = (\beta, g) \in \Theta^p_n$. By the model assumption we have $P\{\dot{l}_g(\beta_0, g_0; Z)[h] | X\} = 0$ for all $h \in \{h = \frac{\partial g_\eta}{\partial \eta}|_{\eta=0}, g_\eta \in \mathcal{H}^p\}$. The result holds in conditional expectation because the log-likelihood $l(\beta,g;Z)$ is from the conditional density of (Y,Δ,X) with conditioning on X. Taking h to be $\ddot{g}_0$ and $\dot{g}-\dot{g}_0$ respectively, we have

$$P\left\{\Delta\ddot{g}_0(Y-X'\beta_0)-\int 1(Y\geq t)\exp\{g_0(t-X'\beta_0)\}\ddot{g}_0(t-X'\beta_0)\,dt\middle|X\right\}=0$$

and

$$P\left\{\Delta(\dot{g}-\dot{g}_0)(Y-X'\beta_0)-\int 1(Y\geq t)\exp\{g_0(t-X'\beta_0)\}(\dot{g}-\dot{g}_0)(t-X'\beta_0)\,dt\middle|X\right\}=0.$$

Then it follows

$$\begin{array}{lrcl}
 & A & = & P\left\{\frac{1}{2}\int_a^\tau 1(\epsilon_0\geq t)\exp\{g_0(t)\}\big\{-[\dot{g}_0(t)(\beta-\beta_0)'X]^2\right.\\
 & & & \left.\quad+\,2\dot{g}_0(t)(\beta-\beta_0)'X(g-g_0)(t)-(g-g_0)^2(t)\big\}\,dt\right\}\\
 & & = & -P\left\{\frac{1}{2}\int_a^\tau 1(\epsilon_0\geq t)\exp\{g_0(t)\}\big[\dot{g}_0(t)(\beta-\beta_0)'X-(g-g_0)(t)\big]^2\,dt\right\}\\
(3.21) & & = & -\frac{1}{2}\int_a^\tau \exp\{g_0(t)\}P\big\{1(\epsilon_0\geq t)\big[\dot{g}_0(t)(\beta-\beta_0)'X-(g-g_0)(t)\big]^2\big\}\,dt.
\end{array}$$

The integrand is from a to τ $(\leq b)$ because of Condition C.3 and C.6. Denote $s_0(t)=-\dot{g}_0(t)$, $s_1(t;\epsilon_0,X)=1(\epsilon_0\geq t)(\beta-\beta_0)'X$, and $s_2(t;\epsilon_0)=1(\epsilon_0\geq t)$, then

$$\begin{array}{lcl}
 & & P\big\{1(\epsilon_0\geq t)[\dot{g}_0(t)(\beta-\beta_0)'X-(g-g_0)(t)]^2\big\}\\
 & = & P\big\{[s_0(t)s_1(t;\epsilon_0,X)+(g-g_0)(t)s_2(t;\epsilon_0)]^2\big\}\\
(3.22) & \geq & s_0^2(t)P(s_1^2)+(g-g_0)^2(t)P(s_2^2)-2|s_0(t)(g-g_0)(t)P(s_1s_2)|
\end{array}$$

Using the same argument in Wellner and Zhang (2007), page 2126, under Condition C.7, we have $[P(s_1s_2)]^2\leq(1-\eta)P(s_1^2)P(s_2^2)$ for some $\eta\in(0,1)$. It follows

$$\begin{array}{rcl}
(3.22) & \geq & s_0^2(t)P(s_1^2)+(g-g_0)^2(t)P(s_2^2)\\
 & & \quad-\,(1-\eta)^{\frac{1}{2}}\cdot 2|s_0(t)\{P(s_1^2)\}^{\frac{1}{2}}|\cdot|(g-g_0)(t)\{P(s_2^2)\}^{\frac{1}{2}}|\\
 & \geq & \{1-(1-\eta)^{\frac{1}{2}}\}\{s_0^2(t)P(s_1^2)+(g-g_0)^2(t)P(s_2^2)\}\\
 & \gtrsim & \dot{g}_0^2(t)(\beta-\beta_0)'P[1(\epsilon_0\geq t)XX'](\beta-\beta_0)+(g-g_0)^2(t)P[1(\epsilon_0\geq t)].
\end{array}$$

Hence, we have

$$\begin{aligned}(3.21) \quad &\lesssim -\Big\{(\beta-\beta_0)'\Big[\int_a^\tau \exp\{g_0(t)\}\dot{g}_0^2(t)P[1(\epsilon_0 \geq t)XX']\ dt\Big](\beta-\beta_0) \\ &\quad + \int_a^\tau \exp\{g_0(t)\}(g-g_0)^2(t)P[1(\epsilon_0 \geq t)]\ dt\Big\} \\ &= -(A_1+A_2).\end{aligned}$$

For A_1, Condition C.3 implies that

$$P[1(\epsilon_0 \geq t)XX'] = P[XX'P(\epsilon_0 \geq t|X)] \geq P[XX'P(\epsilon_0 \geq \tau|X)] \geq \delta P(XX').$$

Then Condition C.2(b) yields that $P(XX')$ is positive definite and thus its smallest eigenvalue $\lambda_1 > 0$. In addition, $\int_a^\tau \exp\{g_0(t)\}\dot{g}_0^2(t)\ dt$ is bounded away from zero since $\exp\{g_0(t)\}, \dot{g}_0^2(t) \geq 0$ but not $\equiv 0$ for $t \in [a,\tau]$. Hence it follows that

$$A_1 \gtrsim (\beta-\beta_0)'P(XX')(\beta-\beta_0) \geq \lambda_1|\beta-\beta_0|^2 \gtrsim |\beta-\beta_0|^2.$$

For A_2, Condition C.3 with the fact that $\exp\{g_0(t)\} = d\Lambda_0(t)$ yields

$$A_2 \geq P(\epsilon_0 \geq \tau)\int_a^\tau (g-g_0)^2(t)\ d\Lambda_0(t) \gtrsim \|g-g_0\|_2^2.$$

Therefore

$$(3.21) \lesssim -(|\beta-\beta_0|^2 + \|g-g_0\|_2^2) = -d^2(\theta,\theta_0),$$

and thus

$$Pl(\beta,g;Z) - Pl(\beta_0,g_0;Z) \lesssim -d^2(\theta,\theta_0) + o(d^2(\theta,\theta_0)) \lesssim -d^2(\theta,\theta_0),$$

i.e. $P(l(\theta_0;Z) - Pl(\theta;Z)) \gtrsim d^2(\theta,\theta_0)$ for any $\theta \in \Theta_n^p$, which implies that

$$\inf_{\{d(\theta,\theta_0)\geq\varepsilon,\theta\in\Theta_n^p\}} P(l(\theta_0;Z) - l(\theta;Z)) \gtrsim \varepsilon^2.$$

Hence Condition C1 of Shen and Wong (1994) in page 583 holds with the constant $\alpha = 1$ in their notation.

Next we verify the Condition C2 of Shen and Wong (1994). Denote $\epsilon_\beta = Y - X'\beta$ and $t_\beta = t - X'(\beta - \beta_0)$ for notational simplicity. It follows that

$$\begin{aligned}
& [l(\theta; Z) - l(\theta_0; Z)]^2 \\
= \ & \left\{\Delta g(\epsilon_\beta) - \int_a^b 1(\epsilon_0 \geq t) e^{g(t_\beta)}\, dt - \Delta g_0(\epsilon_0) + \int_a^b 1(\epsilon_0 \geq t) e^{g_0(t)}\, dt\right\}^2 \\
\lesssim \ & \Delta[g(\epsilon_\beta) - g_0(\epsilon_0)]^2 + \left\{\int_a^b 1(\epsilon_0 \geq t)[e^{g(t_\beta)} - e^{g_0(t)}]\, dt\right\}^2 \\
\lesssim \ & \Delta[g(\epsilon_\beta) - g_0(\epsilon_0)]^2 + \int_a^b [e^{g(t_\beta)} - e^{g_0(t)}]^2\, dt \\
\lesssim \ & \Delta[g(\epsilon_\beta) - g(\epsilon_0)]^2 + \Delta[g(\epsilon_0) - g_0(\epsilon_0)]^2 \\
& \quad + \int_a^b [e^{g(t_\beta)} - e^{g(t)}]^2\, dt + \int_a^b [e^{g(t)} - e^{g_0(t)}]^2\, dt \\
= \ & I_1 + I_2 + I_3 + I_4,
\end{aligned}$$

where the second inequality holds because of the Cauchy-Schwartz inequality

$$\begin{aligned}
& \left\{\int_a^b 1(\epsilon_0 \geq t)[e^{g(t_\beta)} - e^{g_0(t)}]\, dt\right\}^2 \leq \left\{\int_a^b 1(\epsilon_0 \geq t)\, dt\right\}\left\{\int_a^b [e^{g(t_\beta)} - e^{g_0(t)}]^2\, dt\right\} \\
& \leq (b-a)\int_a^b [e^{g(t_\beta)} - e^{g_0(t)}]^2\, dt,
\end{aligned}$$

and the third inequality holds by subtracting and adding the terms $g(\epsilon_0)$ and $e^{g(t)}$. For I_1, since $\dot{g} \in \mathcal{H}_n^{p-1}$ is bounded, applying Taylor expansion for g at ϵ_0 we get

$$\begin{aligned}
PI_1 &= P\big\{\Delta[g(\epsilon_0 - X'(\beta - \beta_0)) - g(\epsilon_0)]^2\big\} \\
&\leq P[\dot{g}(\epsilon_0 - X'(\tilde{\beta} - \beta_0))X'(\beta - \beta_0)]^2 \\
&\lesssim P[X'(\beta - \beta_0)]^2 = (\beta - \beta_0)'P(XX')(\beta - \beta_0) \\
&\leq \lambda_d|\beta - \beta_0|^2 \lesssim |\beta - \beta_0|^2,
\end{aligned}$$

where λ_d is the largest eigenvalue of $P(XX')$. For I_2, since the density function for $(Y, \Delta = 1, X)$ is

$$f_{Y,\Delta,X}(y, 1, x) = \lambda_0(y - x'\beta_0)e^{-\Lambda_0(y - x'\beta_0)}\bar{G}_{C|X}(y|X = x)f_X(x),$$

it follows that

$$
\begin{aligned}
& PI_2 = P[\Delta(g-g_0)^2(\epsilon_0)] \\
(3.23) \quad = \;& \int_{\mathcal{X}} \left\{ \int_a^b (g(t)-g_0(t))^2 \lambda_0(t) e^{-\Lambda_0(t)} \bar{G}_{C|X}(t+x'\beta_0|X=x)\ dt \right\} f_X(x)\ dx \\
\leq \;& \int_{\mathcal{X}} \left\{ \int_a^b (g(t)-g_0(t))^2 \lambda_0(t)\ dt \right\} f_X(x)\ dx \\
= \;& \int_{\mathcal{X}} \|g-g_0\|_2^2 \cdot f_X(x)\ dx = \|g-g_0\|_2^2.
\end{aligned}
$$

Then for I_3, since $g \in \mathcal{H}_n^p$ is bounded, it follows that

$$
\begin{aligned}
PI_3 &= P\int_a^b [e^{g(t-X'(\beta-\beta_0))} - e^{g(t)}]^2\ dt \\
&= P\int_a^b e^{2g(t-X'(\tilde{\beta}-\beta_0))}[X'(\beta-\beta_0)]^2\ dt \\
&\lesssim \int_a^b P[X'(\beta-\beta_0)]^2\ dt \\
&\lesssim (\beta-\beta_0)'P[XX'](\beta-\beta_0) \lesssim |\beta-\beta_0|^2.
\end{aligned}
$$

Finally for I_4, by the Taylor expansion for $e^{g(t)}$ at g_0, we have

$$
\begin{aligned}
PI_4 &= \int_a^b [e^{g(t)} - e^{g_0(t)}]^2\ dt \\
&\leq \int_a^b e^{2\tilde{g}(t)}(g(t)-g_0(t))^2\ dt \\
&= \int_a^b e^{2\tilde{g}(t)-g_0(t)}(g(t)-g_0(t))^2\ d\Lambda_0(t) \\
&\lesssim \int_a^b (g(t)-g_0(t))^2\ d\Lambda_0(t) = \|g-g_0\|_2^2,
\end{aligned}
$$

where $\tilde{g}(t) = g_0(t) + \xi(g-g_0)(t)$ for some $0 < \xi < 1$ and hence is bounded. Therefore we have

$$P(l(\theta;Z) - l(\theta_0;Z))^2 \lesssim |\beta-\beta_0|^2 + \|g-g_0\|_2^2 = d^2(\theta,\theta_0)$$

for any $\theta \in \Theta_n^p$, which implies that

$$\sup_{\{d(\theta,\theta_0)\leq\varepsilon, \theta\in\Theta_n^p\}} \mathrm{Var}(l(\theta_0;Z) - l(\theta;Z)) \leq \sup_{\{d(\theta,\theta_0)\leq\varepsilon, \theta\in\Theta_n^p\}} P(l(\theta_0;Z) - l(\theta;Z))^2 \lesssim \varepsilon^2.$$

So Condition C2 of Shen and Wong (1994) in page 583 holds with the constant $\beta = 1$ in their notation.

Finally, we verify the Condition C3 in Shen and Wong (1994). By lemma A.3, for $\mathcal{F}_n = \{l(\theta; Z) - l(\theta_{0,n}; Z) : \theta \in \Theta_n^p\}$, we have $N_{[\ ]}(\varepsilon, \mathcal{F}_n, \|\cdot\|_\infty) \lesssim (1/\varepsilon)^{cq_n+d}$. Then by the fact that the covering number is bounded by the bracketing number, it follows that

$$H(\varepsilon, \mathcal{F}_n, \|\cdot\|_\infty) = \log N(\varepsilon, \mathcal{F}_n, \|\cdot\|_\infty) \lesssim (cq_n + d)\log(1/\varepsilon) \lesssim n^\nu \log(1/\varepsilon).$$

So Condition C3 of Shen and Wong (1994) in page 583 holds with the constants $2r_0 = \nu$ and $r = 0^+$ in their notation.

Therefore, the constant τ in Theorem 1 of Shen and Wong (1994), page 584, is $\frac{1-\nu}{2} - \frac{\log\log n}{2\log n}$. Since $\frac{\log\log n}{2\log n} \to 0$ as $n \to 0$, we can pick a $\tilde{\nu}$ slightly greater than ν such that $\frac{1-\tilde{\nu}}{2} \le \frac{1-\nu}{2} - \frac{\log\log n}{2\log n}$ for n large. We still denote the $\tilde{\nu}$ as ν and then $\tau = \frac{1-\nu}{2}$. Then by definition $\hat{\theta}_n$ maximizes the empirical log-likelihood $\mathbb{P}_n l(\theta; Z)$ over the sieve space Θ_n^p, so $\hat{\theta}_n$ satisfies the inequality (1.1) in Shen and Wong (1994) with $\eta_n = 0$. By Lemma A.2, there exists an $g_{0,n} \in \mathcal{H}_n^p$ such that $\|g_{0,n} - g_0\|_\infty = O(n^{-p\nu})$. Moreover, by the Taylor expansion for $P[l(\beta_0, g_0; Z) - l(\beta, g; Z)]$ in (3.20) and plugging in $\theta = \theta_{0,n} = (\beta_0, g_{0,n})$, the Kullback-Leilber pseudodistance of $\theta_{0,n} = (\beta_0, g_{0,n})$ and $\theta_0 = (\beta_0, g_0)$ follows

$$\begin{aligned}
K(\theta_{0,n}, \theta_0) &= P[l(\theta_0; Z) - l(\theta_{0,n}; Z)] \\
&= -P\big\{\ddot{l}_{gg}(\beta_0, g_0)[g_{0,n} - g_0, g_{0,n} - g_0]\big\} + o(\|g_{0,n} - g_0\|_2^2) \\
&= P\bigg\{\int_a^b 1(\epsilon_0 \ge t)\exp\{g_0(t)\}(g_{0,n}(t) - g_0(t))^2\, dt\bigg\} + o(\|g_{0,n} - g_0\|_2^2) \\
&\le \int_a^b (g_{0,n}(t) - g_0(t))^2\, d\Lambda_0(t) + o(\|g_{0,n} - g_0\|_2^2) \\
&= \|g_{0,n} - g_0\|_2^2 + o(\|g_{0,n} - g_0\|_2^2) = O(n^{-2p\nu}),
\end{aligned}$$

where the last equality holds because $\|g_{0,n} - g_0\|_2 \leq \|g_{0,n} - g_0\|_\infty = O(n^{-p\nu})$. Therefore $K^{1/2}(\theta_{0,n}, \theta_0) = O(n^{-p\nu})$. Thus by Theorem 1 of Shen and Wong (1994), we obtain the convergence rate for $\hat{\theta}_n$ as

$$d(\hat{\theta}_n, \theta_0) = O_p\{\max(n^{-(1-\nu)/2}, n^{-p\nu}, n^{-p\nu})\} = O_p\{n^{-\min(p\nu,(1-\nu)/2)}\}.$$

This completes the proof of our Theorem 3.3.1.

3.8.3 Proof of Theorem 3.3.2

We prove the theorem by checking the assumptions A1', A2-A4 and A5'-A6' of Corollary 3.6.2. Here the criterion function of a single observation is the log-likelihood function $l(\beta, g; Z)$. So instead of using m, we use l to denote the criterion function. All the first and second derivatives of $l(\beta, g; Z)$ with respect to β and g are calculated in Lemma A.1. By Theorem 3.3.1, note that assumption A1' holds with $\gamma = \min(p\nu, (1-\nu)/2)$ and the norm $\|\cdot\|_2$ defined in (3.11). A2 automatically holds by the model assumption. For A3, we need to find an $\mathbf{h}^* = (h_1^*, \cdots, h_d^*)'$ such that

$$\begin{aligned} &\ddot{S}_{\beta g}(\beta_0, g_0)[h] - \ddot{S}_{gg}(\beta_0, g_0)[\mathbf{h}^*, h] \\ &\qquad = P\{\ddot{l}_{\beta g}(\beta_0, g_0; Z)[h] - \ddot{l}_{gg}(\beta_0, g_0; Z)[\mathbf{h}^*, h]\} = 0 \end{aligned}$$

for all $h \in \mathbf{H} = \{h : h = \frac{\partial g_\eta}{\partial \eta}|_{\eta=0}, g_\eta \in \mathcal{H}^p\}$. Note that

$$\begin{aligned} & P\{\ddot{l}_{\beta g}(\beta_0, g_0; Z)[h] - \ddot{l}_{gg}(\beta_0, g_0; Z)[\mathbf{h}^*, h]\} \\ = \;& P\Bigg\{-X\left[\Delta\dot{h}(\epsilon_0) - \int_a^b 1(\epsilon_0 \geq t)\exp\{g_0(t)\}\dot{h}(t)\; dt\right] \\ & \quad + \int_a^b 1(\epsilon_0 \geq t)\exp\{g_0(t)\}h(t)[X\dot{g}_0(t) + \mathbf{h}^*(t)]\; dt\Bigg\}. \end{aligned}$$

As in the proof of Theorem 3.3.1, $P\{\dot{l}_g(\beta_0, g_0; Z)[h]|X\} = 0$ by the model assumption for all $h \in \mathbf{H}$. Therefore, by taking h to be $\dot{h}$, we have

$$\begin{aligned} & P\bigg\{-X\bigg[\Delta\dot{h}(\epsilon_0) - \int_a^b 1(\epsilon_0 \geq t)\exp\{g_0(t)\}\dot{h}(t)\ dt\bigg]\bigg\} \\ = \ & P\bigg\{-X \cdot P\bigg[\Delta\dot{h}(\epsilon_0) - \int_a^b 1(\epsilon_0 \geq t)\exp\{g_0(t)\}\dot{h}(t)\ dt\bigg|X\bigg]\bigg\} \\ = \ & P\{-X \cdot 0\} = 0. \end{aligned}$$

Hence we only need to find a $\mathbf{h}^*$ such that

$$\begin{aligned} & P\bigg\{\int_a^b 1(\epsilon_0 \geq t)\exp\{g_0(t)\}h(t)[X\dot{g}_0(t) + \mathbf{h}^*(t)]\ dt\bigg\} \\ = \ & \int_a^b \exp\{g_0(t)\}h(t)\big\{\dot{g}_0(t)P[1(\epsilon_0 \geq t)X] + \mathbf{h}^*(t)P[1(\epsilon_0 \geq t)]\big\}\ dt = 0. \end{aligned}$$

One obvious choice for $\mathbf{h}^*$ is

$$\mathbf{h}^*(t) = -\dot{g}_0(t)\frac{P[1(\epsilon_0 \geq t)X]}{P[1(\epsilon_0 \geq t)]} = -\dot{g}_0(t)P(X|\epsilon_0 \geq t). \tag{3.24}$$

Then it follows that

$$\begin{aligned} & \dot{l}_\beta(\beta_0, g_0; Z) - \dot{l}_g(\beta_0, g_0; Z)[\mathbf{h}^*] \\ = \ & \Delta\{-\dot{g}_0(Y - X'\beta_0)\}\big\{X - P(X|\epsilon_0 \geq Y - X'\beta_0)\big\} \\ & \quad - \int 1(Y - X'\beta_0 \geq t)\big\{X - P(X|\epsilon_0 \geq t)\big\}\{-\dot{g}_0(t)\}\exp\{g_0(t)\}\ dt \\ = \ & \int \big\{X - P(X|\epsilon_0 \geq t)\big\}\{-\dot{g}_0(t)\}\ dM(t) \\ = \ & \int \big\{X - P(X|Y - X'\beta_0 \geq t)\big\}\{-\dot{g}_0(t)\}\ dM(t), \end{aligned}$$

which is the efficient score function for β_0 with

$$M(t) = \Delta I(Y - X'\beta_0 \leq t) - \int_{-\infty}^t I(Y - X'\beta_0 \geq s)\exp\{g_0(s)\}\ ds.$$

We denote the efficient score function as $l^*_{\beta_0}(Y, \Delta, X)$. Then by the fact that

$$P\dot{l}_\beta(\beta, g; Z) = \int_{\mathcal{Z}} \dot{l}_\beta(\beta, g; z)f(z; \beta, g)\ dz = 0,$$

where $f(z;\beta,g)$ is the density function for $Z=(Y,\Delta,X)$, it follows that for any $h\in\mathbf{H}$,

$$\begin{aligned}
0 &= \frac{\partial}{\partial\eta}P\dot{l}_\beta(\beta,g+\eta h;Z)|_{\eta=0} \\
&= \int_{\mathcal{Z}}\ddot{l}_{\beta g}(\beta,g;z)[h]f(z;\beta,g)\;dz+\int_{\mathcal{Z}}\dot{l}_\beta(\beta,g;z)\dot{f}_g(z;\beta,g)[h]\;dz \\
&= P\ddot{l}_{\beta g}(\beta,g;Z)[h]+\int_{\mathcal{Z}}\dot{l}_\beta(\beta,g;Z)\dot{l}'_g(\beta,g;z)[h]f(z;\beta,g)\;dz \\
&= P\ddot{l}_{\beta g}(\beta,g;Z)[h]+P\big\{\dot{l}_\beta(\beta,g;Z)\dot{l}'_g(\beta,g;Z)[h]\big\},
\end{aligned}$$

and hence $P\ddot{l}_{\beta g}(\beta,g;Z)[h]=-P\big\{\dot{l}_\beta(\beta,g;Z)\dot{l}'_g(\beta,g;Z)[h]\big\}$. Similarly, we have

$$\begin{aligned}
P\ddot{l}_{g\beta}(\beta,g;Z)[h] &= -P\big\{\dot{l}_g(\beta,g;Z)[h]\dot{l}'_\beta(\beta,g;Z)\big\}, \\
P\ddot{l}_{\beta\beta}(\beta,g;Z) &= -P\big\{\dot{l}_\beta(\beta,g;Z)\dot{l}'_\beta(\beta,g;Z)\big\}, \\
P\ddot{l}_{gg}(\beta,g;Z)[h_1,h_2] &= -P\big\{\dot{l}_g(\beta,g;Z)[h_1]\dot{l}'_g(\beta,g;Z)[h_2]\big\}.
\end{aligned}$$

Then together with the fact that $P\{\ddot{l}_{\beta g}(\beta_0,g_0;Z)[\mathbf{h}^*]-\ddot{l}_{gg}(\beta_0,g_0;Z)[\mathbf{h}^*,\mathbf{h}^*]\}=0$, the matrix A in assumption A3 of Theorem 3.6.1 follows

$$\begin{aligned}
A &= P\{-\ddot{l}_{\beta\beta}(\beta_0,g_0;Z)+\ddot{l}_{g\beta}(\beta_0,g_0;Z)[\mathbf{h}^*]\} \\
&= P\{-\ddot{l}_{\beta\beta}(\beta_0,g_0;Z)+\ddot{l}_{g\beta}(\beta_0,g_0;Z)[\mathbf{h}^*] \\
&\qquad +\ddot{l}_{\beta g}(\beta_0,g_0;Z)[\mathbf{h}^*]-\ddot{l}_{gg}(\beta_0,g_0;Z)[\mathbf{h}^*,\mathbf{h}^*]\} \\
&= P\{\dot{l}_\beta(\beta_0,g_0;Z)\dot{l}'_\beta(\beta_0,g_0;Z)-\dot{l}_g(\beta_0,g_0;Z)[\mathbf{h}^*]\dot{l}'_\beta(\beta_0,g_0;Z) \\
&\qquad -\dot{l}_\beta(\beta_0,g_0;Z)\dot{l}'_g(\beta_0,g_0;Z)[\mathbf{h}^*]+\dot{l}_g(\beta_0,g_0;Z)[\mathbf{h}^*]\dot{l}'_g(\beta_0,g_0;Z)[\mathbf{h}^*]\} \\
&= P\{\dot{l}_\beta(\beta_0,g_0;Z)-\dot{l}_g(\beta_0,g_0;Z)[\mathbf{h}^*]\}^{\otimes 2}=Pl^*_{\beta_0}(Y,\Delta,X)^{\otimes 2},
\end{aligned}$$

which is the information matrix for β_0 under semiparametric efficiency and it is non-singular by Conditions C.3-C.5.

To verify A4, we note that the first part automatically holds since $\hat{\beta}_n$ satisfies the score equation (3.9), that is $\dot{S}_{\beta,n}(\hat{\beta}_n,\hat{g}_n)=\mathbb{P}_n\dot{l}_\beta(\hat{\beta}_n,\hat{g}_n;Z)=0$. Next we shall show that

$$\begin{aligned}
&\dot{S}_{g,n}(\hat{\beta}_n, \hat{g}_n)[h_j^*] \\
&\quad = \mathbb{P}_n \left\{ \Delta h_j^*(Y - X'\hat{\beta}_n) - \int 1(Y \geq t) \exp\{\hat{g}_n(t - X'\hat{\beta}_n)\} h_j^*(t - X'\hat{\beta}_n)\, dt \right\} \\
&\quad = o_p(n^{-1/2}),
\end{aligned}$$

where $h_j^*(t) = -\dot{g}_0(t)P(X_j|\epsilon_0 \geq t)$, $j = 1, \cdots, d$, is the jth component of $\mathbf{h}^*(t)$ given in (3.24). According to Lemma A.4, there exists an $h_{j,n}^* \in \mathcal{H}_n^2$, such that $\|h_j^* - h_{j,n}^*\|_\infty = O(n^{-2\nu})$. Then by the score equation (3.10) and the fact that $h_{j,n}^*(t)$ can be written as $h_{j,n}^*(t) = \sum_{k=1}^{q_n} \gamma_{j,k}^* B_k(t)$ for some coefficients $\{\gamma_{j,1}^*, \cdots, \gamma_{j,q_n}^*\}$ and the basis functions $B_k(t)$ of the spline space, it follows that

$$\mathbb{P}_n \left\{ \Delta h_{j,n}^*(Y - X'\hat{\beta}_n) - \int 1(Y \geq t) \exp\{\hat{g}_n(t - X'\hat{\beta}_n) h_{j,n}^*(t - X'\hat{\beta}_n)\}\, dt \right\} = 0.$$

So it suffices to show that for each $1 \leq j \leq d$,

$$I_n = \dot{S}_{g,n}(\hat{\beta}_n, \hat{g}_n)[h_j^* - h_{j,n}^*] = \mathbb{P}_n \dot{l}_g(\hat{\beta}_n, \hat{g}_n; Z)[h_j^* - h_{j,n}^*] = o_p(n^{-1/2}).$$

But I_n can be decomposed as $I_n = I_{1n} + I_{2n}$, where

$$I_{1n} = (\mathbb{P}_n - P)\dot{l}_g(\hat{\beta}_n, \hat{g}_n; Z)[h_j^* - h_{j,n}^*]$$

and

$$I_{2n} = P\big\{\dot{l}_g(\hat{\beta}_n, \hat{g}_n; Z)[h_j^* - h_{j,n}^*] - \dot{l}_g(\beta_0, g_0; Z)[h_j^* - h_{j,n}^*]\big\},$$

because $P\big\{\dot{l}_g(\beta_0, g_0; Z)[h_j^* - h_{j,n}^*]\big\} = 0$. We will show that I_{1n} and I_{2n} are both $o_p(n^{-1/2})$.

First for I_{1n}, according to Lemma A.5, the ε-bracketing number associated with $\|\cdot\|_\infty$ norm for the class $\mathcal{F}_n^j(\eta)$, where

$$\mathcal{F}_n^j(\eta) = \{\dot{l}_g(\theta; z)[h_j^* - h] : \theta \in \Theta_n^p, h \in \mathcal{H}_n^2 \text{ and } d(\theta, \theta_0) \leq \eta, \|h - h_j^*\|_\infty \leq \eta\},$$

is bounded by $(\eta/\varepsilon)^{cq_n+d}$. This implies that

$$\log N_{[\,]}(\varepsilon, \mathcal{F}_n^j(\eta), L_2(P)) \leq \log N_{[\,]}(\varepsilon, \mathcal{F}_n^j(\eta), \|\cdot\|_\infty) \lesssim q_n \log(\eta/\varepsilon),$$

which leads to the bracketing integral

$$J_{[\,]}(\eta, \mathcal{F}_n^j(\eta), L_2(P)) = \int_0^\eta \sqrt{1 + \log N_{[\,]}(\varepsilon, \mathcal{F}_n^j(\eta), L_2(P))}\; d\varepsilon \lesssim q_n^{1/2}\eta.$$

Now we pick η to be $\eta_n = O\{n^{-\min(2\nu,(1-\nu)/2)}\}$, then

$$\|h_j^* - h_{j,n}^*\|_\infty = O(n^{-2\nu}) \leq O\{n^{-\min(2\nu,(1-\nu)/2)}\} = \eta_n,$$

and since $p \geq 3$,

$$d(\hat{\theta}_n, \theta_0) = O_p\{n^{-\min(p\nu,(1-\nu)/2)}\} \leq O_p\{n^{-\min(2\nu,(1-\nu)/2)}\} = \eta_n.$$

Therefore, $\dot{l}_g(\hat{\beta}_n, \hat{g}_n; z)[h_j^* - h_{j,n}^*] \in \mathcal{F}_n^j(\eta_n)$. As in the proof of Theorem 3.3.1, denote $t_\beta = t - X'(\beta - \beta_0)$ for notational simplicity, for any $\dot{l}_g(\theta; Z)[h_j^* - h] \in \mathcal{F}_n^j(\eta_n)$, it then follows that

$$\begin{aligned}
& P\{\dot{l}_g(\theta; Z)[h_j^* - h]\}^2 \\
= \;& P\left\{\Delta(h_j^* - h)(\epsilon_\beta) + \int_a^b 1(\epsilon_0 \geq t)\exp\{g(t_\beta)\}(h_j^* - h)(t_\beta)\; dt\right\}^2 \\
\lesssim \;& P[(h_j^* - h)^2(\epsilon_\beta)] + P\left\{\int_a^b 1(\epsilon_0 \geq t)\exp\{g(t_\beta)\}(h_j^* - h)(t_\beta)\; dt\right\}^2 \\
\lesssim \;& \|h_j^* - h\|_\infty^2 + P\left\{\int_a^b \exp\{2g(t_\beta)\}(h_j^* - h)^2(t_\beta)\; dt\right\} \\
= \;& \|h_j^* - h\|_\infty^2 + \|h_j^* - h\|_\infty^2 \int_a^b P[\exp\{2g(t_\beta)\}]\; dt \\
\lesssim \;& \|h_j^* - h\|_\infty^2 \leq \eta_n^2,
\end{aligned}$$

where the second inequality holds because of the Cauchy-Schwartz inequality and second to last inequality holds because g is bounded. Moreover, by Lemma A.1,

$\|\dot{l}_g(\theta; Z)[h_j^* - h]\|_\infty$ is bounded by some constant $M > 0$. Then by the maximal inequality in Lemma 3.4.2 of van der Vaart and Wellner (1996), it follows that

$$\begin{aligned} E_P\|\mathbb{G}_n\|_{\mathcal{F}_n^j(\eta_n)} &\lesssim J_{[\,]}(\eta_n, \mathcal{F}_n^j(\eta_n), L_2(P))\left(1 + \frac{J_{[\,]}(\eta_n, \mathcal{F}_n^j(\eta_n), L_2(P))}{\eta_n^2\sqrt{n}}M\right) \\ &\lesssim q_n^{1/2}\eta_n + q_n n^{-1/2} \\ &= O\{n^{\nu/2-\min(2\nu,(1-\nu)/2)}\} + O(n^{\nu-1/2}) \\ &= O\{n^{-\min(3\nu/2,1/2-\nu)}\} + O(n^{\nu-1/2}) = o(1), \end{aligned}$$

where the last equality holds because $0 < \nu < 1/2$. Thus by the Markov's inequality,

$$I_{1n} = n^{-1/2}\mathbb{G}_n\dot{l}_g(\hat{\theta}_n; Z)[h_j^* - h_{j,n}^*] = o_p(n^{-1/2}).$$

Next for I_{2n}, the Taylor expansion for $\dot{l}_g(\hat{\theta}_n; Z)[h_j^* - h_{j,n}^*]$ at θ_0 yields

$$\begin{aligned} &\dot{l}_g(\hat{\beta}_n, \hat{g}_n; Z)[h_j^* - h_{j,n}^*] - \dot{l}_g(\beta_0, g_0; Z)[h_j^* - h_{j,n}^*] \\ = \;& (\hat{\beta}_n - \beta_0)'\ddot{l}_{\beta g}(\tilde{\beta}_n, \tilde{g}_n; Z)[h_j^* - h_{j,n}^*] + \ddot{l}_{gg}(\tilde{\beta}_n, \tilde{g}_n; Z)[h_j^* - h_{j,n}^*, \hat{g}_n - g_0], \end{aligned}$$

where $(\tilde{\beta}_n, \tilde{g}_n)$ is the intermediate value between (β_0, g_0) and $(\hat{\beta}_n, \hat{g}_n)$. Then denote $\epsilon_{\tilde{\beta}_n} = Y - X'\tilde{\beta}_n$ and $t_{\tilde{\beta}_n} = t - X'(\tilde{\beta}_n - \beta_0)$ for notational simplicity, it follows that

$$\begin{aligned} &|\ddot{l}_{\beta g}(\tilde{\beta}_n, \tilde{g}_n; Z)[h_j^* - h_{j,n}^*]| \\ = \;& \Bigg|X\Bigg\{\Delta(\dot{h}_j^* - \dot{h}_{j,n}^*)(\epsilon_{\tilde{\beta}_n}) \\ & - \int_a^b 1(\epsilon_0 \geq t)\exp\{\tilde{g}_n(t_{\tilde{\beta}_n})\}\left[(\dot{h}_j^* - \dot{h}_{j,n}^*)(t_{\tilde{\beta}_n}) + \dot{\tilde{g}}_n(t_{\tilde{\beta}_n})(h_j^* - h_{j,n}^*)(t_{\tilde{\beta}_n})\right] dt\Bigg\}\Bigg| \\ \lesssim \;& \|\dot{h}_j^* - \dot{h}_{j,n}^*\|_\infty + \|\dot{h}_j^* - \dot{h}_{j,n}^*\|_\infty\left\{\int_a^b \exp\{\tilde{g}_n(t_{\tilde{\beta}_n})\}\,dt\right\} \\ & + \|h_j^* - h_{j,n}^*\|_\infty\left\{\int_a^b \exp\{\tilde{g}_n(t_{\tilde{\beta}_n})\dot{\tilde{g}}_n(t_{\tilde{\beta}_n})\}\,dt\right\} \\ \lesssim \;& \|\dot{h}_j^* - \dot{h}_{j,n}^*\|_\infty + \|h_j^* - h_{j,n}^*\|_\infty \\ = \;& O(n^{-\nu}) + O(n^{-2\nu}) = O(n^{-\nu}), \end{aligned}$$

where second inequality holds because $\tilde{g}_n$ and its first derivative $\dot{\tilde{g}}_n$ are bounded, and the last equality holds due to the Corollary 6.21 of Schumaker (1981) that $\|\dot{h}_j^* - \dot{h}_{j,n}^*\|_\infty = O(n^{-(2-1)\nu}) = O(n^{-\nu})$. Thus,

$$\begin{aligned} &|(\hat{\beta}_n - \beta_0)'\ddot{l}_{\beta g}(\tilde{\beta}_n, \tilde{g}_n; Z)[h_j^* - h_{j,n}^*]| = |\hat{\beta}_n - \beta_0| \cdot O(n^{-2\nu}) \\ &\quad = O_p\{n^{-\min(p\nu,(1-\nu)/2)}\} \cdot O(n^{-\nu}) = O_p\{n^{-\min((p+1)\nu,(1+3\nu)/2)}\} \end{aligned}$$

And also

$$\begin{aligned} & |\ddot{l}_{gg}(\tilde{\beta}_n, \tilde{g}_n; Z)[h_j^* - h_{j,n}^*, \hat{g}_n - g_0]| \\ = & \left| \int_a^b 1(\epsilon_0 \geq t) \exp\{\tilde{g}_n(t_{\tilde{\beta}_n})\}(h_j^* - h_{j,n}^*)(t_{\tilde{\beta}_n})(\hat{g}_n - g_0)(t_{\tilde{\beta}_n})\ dt \right| \\ \leq & \|h_j^* - h_{j,n}^*\|_\infty \cdot \left\{ \int_a^b \exp\{\tilde{g}_n(t_{\tilde{\beta}_n})\}(\hat{g}_n - g_0)(t_{\tilde{\beta}_n})\ dt \right\} \\ = & \|h_j^* - h_{j,n}^*\|_\infty \cdot I_{3n}, \end{aligned}$$

where by the Cauchy-Schwartz inequality and the boundedness of $\tilde{g}_n$, we have

$$\begin{aligned} \{I_{3n}\}^2 &= \left\{ \int_a^b \exp\{\tilde{g}_n(t_{\tilde{\beta}_n})\}(\hat{g}_n - g_0)(t_{\tilde{\beta}_n})\ dt \right\}^2 \\ &\lesssim \int_a^b \exp\{2\tilde{g}_n(t_{\tilde{\beta}_n})\}(\hat{g}_n - g_0)^2(t_{\tilde{\beta}_n})\ dt \lesssim \|\hat{g}_n - g_0\|_2^2. \end{aligned}$$

Hence $I_{3n} \lesssim \|\hat{g}_n - g_0\|_2$ and

$$\begin{aligned} &|\ddot{l}_{gg}(\tilde{\beta}_n, \tilde{g}_n; Z)[h_j^* - h_{j,n}^*, \hat{g}_n - g_0]| = \|h_j^* - h_{j,n}^*\|_\infty \cdot \|\hat{g}_n - g_0\|_2 \\ &\quad = O(n^{-2\nu}) \cdot O_p\{n^{-\min(p\nu,(1-\nu)/2)}\} = O_p\{n^{-\min((p+2)\nu,(1+3\nu)/2)}\}. \end{aligned}$$

Finally since $\frac{1}{2(1+p)} < \nu < \frac{1}{1+2p}$, it follows that

$$I_{2n} = O\{n^{-\min((p+1)\nu,(1+3\nu)/2)}\} = o(n^{-1/2}).$$

Thus we have shown that $I_n = I_{1n} + I_{2n} = o_p(n^{-1/2})$ and Condition A4 holds.

Now we verify assumption A5'. First by Lemma A.6, the ε-bracketing numbers for the classes of functions

$$\mathcal{F}^{\beta}_{n,j}(\eta) = \{\dot{l}_{\beta_j}(\theta; z) - \dot{l}_{\beta_j}(\theta_0; z) : \theta \in \Theta^p_n, \dot{g} \in \mathcal{H}^{p-1}_n \text{ and } d(\theta, \theta_0) \leq \eta, \|\dot{g} - \dot{g}_0\|_2 \leq \eta\}$$

and

$$\mathcal{F}^{g}_{n,j}(\eta) = \{\dot{l}_{g}(\theta; z)[h^*_j] - \dot{l}_{g}(\theta_0; z)[h^*_j] : \theta \in \Theta^p_n \text{ and } d(\theta, \theta_0) \leq \eta\},$$

are both bounded by $(\eta/\varepsilon)^{cq_n+d}$, which implies that the corresponding ε-bracketing integrals are both bounded by $q_n^{1/2}\eta$, i.e.,

$$J_{[\ ]}(\eta, \mathcal{F}^{\beta}_{n,j}(\eta), L_2(P)) \lesssim q_n^{1/2}\eta \quad \text{and} \quad J_{[\ ]}(\eta, \mathcal{F}^{g}_{n,j}(\eta), L_2(P)) \lesssim q_n^{1/2}\eta.$$

Then for $\dot{l}_{\beta_j}(\theta; z) - \dot{l}_{\beta_j}(\theta_0; z)$, by applying the Cauchy-Schwartz inequality, together with subtracting and adding the terms $\dot{g}(\epsilon_0)$, $e^{g_0(t_\beta)}\dot{g}(t_\beta)$, $e^{g_0(t)}\dot{g}(t_\beta)$ and $e^{g_0(t)}\dot{g}_0(t_\beta)$, we have

$$\begin{aligned}
& \big\{\dot{l}_{\beta_j}(\theta; Z) - \dot{l}_{\beta_j}(\theta_0; Z)\big\}^2 \\
= & \left\{ - \Delta X_j[\dot{g}(\epsilon_\beta) - \dot{g}_0(\epsilon_0)] + X_j \int_a^b 1(\epsilon_0 \geq t)[e^{g(t_\beta)}\dot{g}(t_\beta) - e^{g_0(t)}\dot{g}_0(t)]\ dt \right\}^2 \\
\lesssim & \ \{\Delta[\dot{g}(\epsilon_\beta) - \dot{g}_0(\epsilon_0)]^2\} + \left\{ \int_a^b [e^{g(t_\beta)}\dot{g}(t_\beta) - e^{g_0(t)}\dot{g}_0(t)]^2\ dt \right\} \\
\lesssim & \ \{\Delta[\dot{g}(\epsilon_\beta) - \dot{g}(\epsilon_0)]^2\} + \{\Delta[\dot{g}(\epsilon_0) - \dot{g}_0(\epsilon_0)]^2\} \\
& \quad + \int_a^b \big\{[e^{g(t_\beta))} - e^{g_0(t_\beta)}]^2 + [e^{g_0(t_\beta)} - e^{g_0(t)}]^2\big\}\dot{g}^2(t_\beta)\ dt \\
& \quad + \int_a^b e^{2g_0(t)}\big\{[\dot{g}(t_\beta) - \dot{g}_0(t_\beta)]^2 + e^{2g_0(t)}[\dot{g}_0(t_\beta) - \dot{g}_0(t)]^2\big\}\ dt \\
= & \ B_1 + B_2 + B_3 + B_4.
\end{aligned}$$

For B_1, since $\ddot{g}$ is bounded and the largest eigenvalue of $P(XX')$ satisfies $0 < \lambda_d < \infty$

by Condition C.2(b), it follows that

$$\begin{aligned} PB_1 &\leq P[\ddot{g}(Y - X'\tilde{\beta})X'(\beta - \beta_0)]^2 \\ &\lesssim P[X'(\beta - \beta_0)]^2 \leq \lambda_d |\beta - \beta_0|^2 \\ &\lesssim |\beta - \beta_0|^2 \leq \eta^2 \end{aligned}$$

For B_2, by (3.23) we have

$$\begin{aligned} PB_2 &= P\{\Delta[\dot{g}(\epsilon_0) - \dot{g}_0(\epsilon_0)]^2\} \\ &\leq \int_{\mathcal{X}} \left\{ \int_a^b (\dot{g}(t) - \dot{g}_0(t))^2 \lambda_0(t)\ dt \right\} f_X(x)\ dx \\ &= \|\dot{g} - \dot{g}_0\|_2^2 \int_{\mathcal{X}} f_X(x)\ dx \\ &= \|\dot{g} - \dot{g}_0\|_2^2 \leq \eta^2. \end{aligned}$$

For B_3, by using the Mean Value Theorem, it follows that

$$\begin{aligned} PB_3 &= P\left\{ \int_a^b \left\{ [e^{\tilde{g}(t_\beta)}(g - g_0)(t_\beta)]^2 + [e^{g_0(t - X'(\tilde{\beta} - \beta_0))} X'(\beta - \beta_0)]^2 \right\} \dot{g}^2(t_\beta)\ dt \right\} \\ &\lesssim \int_a^b (g - g_0)^2(t)\ dt + P[X'(\beta - \beta_0)]^2 \\ &\lesssim \|g - g_0\|_2^2 + |\beta - \beta_0|^2 \leq \eta^2, \end{aligned}$$

where $\tilde{g} = g_0 + \xi(g - g_0)$ for some $0 < \xi < 1$ and thus is bounded. The first inequality above holds because of the boundedness of $\tilde{g}$, g_0 and $\dot{g}$. Finally for B_4, since g_0 and $\ddot{g}_0$ are bounded, by the Mean Value Theorem, it follows that

$$\begin{aligned} PB_4 &= P\left\{ \int_a^b e^{2g_0(t)} \left\{ [\dot{g}(t_\beta) - \dot{g}_0(t_\beta)]^2 + e^{2g_0(t)} [\dot{g}_0(t_\beta) - \dot{g}_0(t)]^2 \right\}\ dt \right\} \\ &\lesssim \int_a^b (\dot{g} - \dot{g}_0)^2(t)\ dt + P \int_a^b [\ddot{g}_0(t - X'(\tilde{\beta} - \beta_0)) X'(\beta - \beta_0)]^2\ dt \\ &\lesssim \|\dot{g} - \dot{g}_0\|_2^2 + P[X'(\beta - \beta_0)]^2 \\ &\lesssim \|\dot{g} - \dot{g}_0\|_2^2 + |\beta - \beta_0|^2 \lesssim \eta^2. \end{aligned}$$

Therefore we have $P\{\dot{l}_{\beta_j}(\theta; Z) - \dot{l}_{\beta_j}(\theta_0; Z)\}^2 \lesssim \eta^2$. Using the similar argument, we can show that $P\{\dot{l}_g(\theta; Z)[h_j^*] - \dot{l}_g(\theta_0; Z)[h_j^*]\}^2 \lesssim \eta^2$. By Lemma A.1, we also have $\|\dot{l}_{\beta_j}(\theta; Z) - \dot{l}_{\beta_j}(\theta_0; Z)\|_\infty$ and $\|\dot{l}_g(\theta; Z)[h_j^*] - \dot{l}_g(\theta_0; Z)[h_j^*]\|_\infty$ are both bounded. Now we pick η as $\eta_n = O\{n^{-\min(p\nu,(1-\nu)/2)}\}$, then by the maximal inequality in Lemma 3.4.2 of van der Vaart and Wellner (1996), it follows that

$$\begin{aligned} E_P\|\mathbb{G}_n\|_{\mathcal{F}^\beta_{n,j}(\eta_n)} &\lesssim q_n^{1/2}\eta_n + q_n n^{-1/2} \\ &= O\{n^{\max((\frac{1}{2}-p)\nu,\nu-\frac{1}{2})}\} + O(n^{\nu-\frac{1}{2}}) = o(1), \end{aligned}$$

where the last equality holds since $p \geq 3$ and $\nu < \frac{1}{2}$. Similarly $E_P\|\mathbb{G}_n\|_{\mathcal{F}^g_{n,j}(\eta_n)} = o(1)$. Thus for $\gamma = \min(p\nu, (1-\nu)/2)$ and $Cn^{-\gamma} = O\{n^{-\min(p\nu,(1-\nu)/2)}\} = \eta_n$, by the Markov's inequality,

$$\sup_{|\beta-\beta_0|+\|g-g_0\|_2 \leq Cn^{-\gamma}} \mathbb{G}_n\{\dot{l}_{\beta_j}(\beta, g; Z) - \dot{l}_{\beta_j}(\beta_0, g_0; Z)\} = o_p(1)$$

and

$$\sup_{|\beta-\beta_0|+\|g-g_0\|_2 \leq Cn^{-\gamma}} \mathbb{G}_n\{\dot{l}_g(\beta, g; Z)[h_j^*] - \dot{l}_g(\beta_0, g_0; Z)[h_j^*]\} = o_p(1).$$

This completes the verification of assumption A5'.

Finally, assumption A6' can be verified by using the Taylor expansion. Since the proofs for the two equations in A6' are essentially identical, we just prove the first equation here. In a neighborhood of $(\beta_0, g_0) : \{(\beta, g) : |\beta - \beta_0| + \|g - g_0\|_2 \leq Cn^{-\gamma}\}$ with $\gamma = \min(p\nu, (1-\nu)/2)$, the Taylor expansion for $\dot{l}_\beta(\beta, g; Z)$ yields

$$\begin{aligned} \dot{l}_\beta(\beta, g; Z) &= \dot{l}_\beta(\beta_0, g_0; Z) + \ddot{l}_{\beta\beta}(\tilde{\beta}, \tilde{g}; Z)(\beta - \beta_0) + \ddot{l}_{\beta g}(\tilde{\beta}, \tilde{g}; Z)[g - g_0] \\ &= \dot{l}_\beta(\beta_0, g_0; Z) + \ddot{l}_{\beta\beta}(\beta_0, g_0; Z)(\beta - \beta_0) + \ddot{l}_{\beta g}(\beta_0, g_0; Z)[g - g_0] \\ &\quad + \{\ddot{l}_{\beta\beta}(\tilde{\beta}, \tilde{g}; Z) - \ddot{l}_{\beta\beta}(\beta_0, g_0; Z)\}(\beta - \beta_0) \\ &\quad + \{\ddot{l}_{\beta g}(\tilde{\beta}, \tilde{g}; Z)[g - g_0] - \ddot{l}_{\beta g}(\beta_0, g_0; Z)[g - g_0]\}, \end{aligned}$$

where $(\tilde{\beta}, \tilde{g})$ is an intermediate value between (β_0, g_0) and (β, g). So

$$\begin{aligned}
&P\{\dot{l}_\beta(\beta, g; Z) - \dot{l}_\beta(\beta_0, g_0; Z) - \ddot{l}_{\beta\beta}(\beta_0, g_0; Z)(\beta - \beta_0) - \ddot{l}_{\beta g}(\tilde{\beta}, \tilde{g}; Z)[g - g_0]\} \\
&\quad = P\{\ddot{l}_{\beta\beta}(\tilde{\beta}, \tilde{g}; Z) - \ddot{l}_{\beta\beta}(\beta_0, g_0; Z)\}(\beta - \beta_0) \\
&\qquad + P\{\ddot{l}_{\beta g}(\tilde{\beta}, \tilde{g}; Z)[g - g_0] - \ddot{l}_{\beta g}(\beta_0, g_0; Z)[g - g_0]\}
\end{aligned}$$

Then By Lemma A.1, we have

$$\begin{aligned}
& P|\ddot{l}_{\beta\beta}(\tilde{\beta}, \tilde{g}; Z) - \ddot{l}_{\beta\beta}(\beta_0, g_0; Z)| \\
\leq\ & P|XX'\Delta\{\ddot{\tilde{g}}(\epsilon_{\tilde{\beta}}) - \ddot{g}_0(\epsilon_0)\}| \\
& + P\bigg\{XX'\bigg|\int_a^b 1(\epsilon_0 \geq t)\big\{\exp\{\tilde{g}(t_{\tilde{\beta}})\}\ddot{\tilde{g}}(t_{\tilde{\beta}}) - \exp\{g_0(t)\}\ddot{g}_0(t)\big\}\ dt \\
& + \int_a^b 1(\epsilon_0 \geq t)\big\{\exp\{\tilde{g}(t_{\tilde{\beta}})\}\dot{\tilde{g}}^2(t_{\tilde{\beta}}) - \exp\{g_0(t)\}\dot{g}_0^2(t)\big\}\ dt\bigg|\bigg\} \\
\lesssim\ & P|\Delta\{\ddot{\tilde{g}}(\epsilon_{\tilde{\beta}}) - \ddot{g}_0(\epsilon_0)\}| + P\bigg\{\int_a^b |\exp\{\tilde{g}(t_{\tilde{\beta}})\}\ddot{\tilde{g}}(t_{\tilde{\beta}}) - \exp\{g_0(t)\}\ddot{g}_0(t)|\ dt\bigg\} \\
& + P\bigg\{\int_a^b |\exp\{\tilde{g}(t_{\tilde{\beta}})\}\dot{\tilde{g}}^2(t_{\tilde{\beta}}) - \exp\{g_0(t)\}\dot{g}_0^2(t)|\ dt\bigg\} \\
=\ & C_1 + C_2 + C_3.
\end{aligned}$$

By applying the similar argument that we are using all the time, it can be shown that

$$C_1 \lesssim |\tilde{\beta} - \beta_0| + \|\ddot{\tilde{g}} - \ddot{g}_0\|_2 = O(n^{-\gamma}) + O\{n^{-\min((p-2)\nu, (1-\nu)/2)}\},$$

where the first equality holds since $|\tilde{\beta} - \beta_0| \leq |\beta - \beta_0| = O(n^{-\gamma})$, and by the Corollary 6.21 of Schumaker (1981), for $\tilde{g}$ within the neighborhood of $\|\tilde{g} - g_0\|_2 = O(n^{-\gamma}) = O\{n^{-\min(p\nu, (1-\nu)/2)}\}$, we have $\|\ddot{\tilde{g}} - \ddot{g}_0\|_2 = O\{n^{-\min((p-2)\nu, (1-\nu)/2)}\}$. Similarly, it can be shown that

$$C_2 \lesssim |\tilde{\beta} - \beta_0| + \|\ddot{\tilde{g}} - \ddot{g}_0\|_2 = O(n^{-\gamma}) + O\{n^{-\min((p-2)\nu, (1-\nu)/2)}\}$$

and

$$C_3 \lesssim |\tilde{\beta} - \beta_0| + \|\dot{\tilde{g}} - \dot{g}_0\|_2 = O(n^{-\gamma}) + O\{n^{-\min((p-1)\nu, (1-\nu)/2)}\}.$$

Therefore, $P|\ddot{l}_{\beta\beta}(\tilde{\beta},\tilde{g};Z)-\ddot{l}_{\beta\beta}(\beta_0,g_0;Z)|=O\{n^{-\min((p-2)\nu,(1-\nu)/2)}\}$ and thus

$$\begin{aligned} & P|\ddot{l}_{\beta\beta}(\tilde{\beta},\tilde{g};Z)-\ddot{l}_{\beta\beta}(\beta_0,g_0;Z)|(\beta-\beta_0) \\ = \;& O\{n^{-\min((p-2)\nu,(1-\nu)/2)}\}\cdot O\{n^{-\min(p\nu,(1-\nu)/2)}\} \\ = \;& O\{n^{-\min(2(p-1)\nu,\frac{1}{2}+(p-\frac{5}{2})\nu,1-\nu)}\}=o(n^{-1/2}), \end{aligned}$$

where the last equality holds since for $p\geq 3$, it follows that $2(p-1)\nu>\frac{p-1}{p+1}\geq\frac{1}{2}$, $\frac{1}{2}+(p-\frac{5}{2})\nu>\frac{1}{2}$ and $1-\nu>\frac{1}{2}$. Moreover, followed by the similar argument for $P|\ddot{l}_{\beta\beta}(\tilde{\beta},\tilde{g};Z)-\ddot{l}_{\beta\beta}(\beta_0,g_0;Z)|$, we are able to show that

$$\begin{aligned} & P|\ddot{l}_{\beta g}(\tilde{\beta},\tilde{g};Z)[g-g_0]-\ddot{l}_{\beta g}(\beta_0,g_0;Z)[g-g_0]| \\ = \;& O\{n^{-\min(2(p-1)\nu,\frac{1}{2}+(p-\frac{5}{2})\nu,1-\nu)}\}=o(n^{-1/2}). \end{aligned}$$

Hence,

$$\begin{aligned} & |P\{\dot{l}_{\beta}(\beta,g;Z)-\dot{l}_{\beta}(\beta_0,g_0;Z)-\ddot{l}_{\beta\beta}(\beta_0,g_0;Z)(\beta-\beta_0)-\ddot{l}_{\beta g}(\tilde{\beta},\tilde{g};Z)[g-g_0]\}| \\ = \;& O\{n^{-\min(2(p-1)\nu,\frac{1}{2}+(p-\frac{5}{2})\nu,1-\nu)}\}=O(n^{-\alpha\gamma}), \end{aligned}$$

where $\alpha=\min(2(p-1)\nu,\frac{1}{2}+(p-\frac{5}{2})\nu,1-\nu)/\min(p\nu,\frac{1-\nu}{2})>1$ and $\alpha\gamma>1/2$.

Therefore, we have verified all six assumptions of Corollary 3.6.2 and thus we have

$$\sqrt{n}(\hat{\beta}_n-\beta_0)=A^{-1}\sqrt{n}\mathbb{P}_n l^*(\beta_0,g_0;Z)+o_p(1)\rightarrow N(0,A^{-1}B(A^{-1})'),$$

where $l^*(\beta_0,g_0;Z)=\dot{l}_{\beta}(\beta_0,g_0;Z)-\dot{l}_g(\beta_0,g_0;Z)[\mathbf{h}^*]=\dot{l}_{\beta}(\beta_0,g_0;Z)$ is the efficient score function for β_0 and $A=P\{\dot{l}_{\beta}(\beta_0,g_0;Z)\}^{\otimes 2}=I(\beta_0)$, as shown in the verification of the assumption A3. Hence $A=B$ and $A^{-1}B(A^{-1})'=A^{-1}=I^{-1}(\beta_0)$, and

$$\sqrt{n}\mathbb{P}_n l^*(\beta_0,g_0;Z)=n^{-\frac{1}{2}}\sum_{i=1}^{n} l^*_{\beta_0}(Y_i,\Delta_i,X_i).$$

Thus equation (3.13) holds and we complete the proof of Theorem 3.3.2.

Table 3.1: Summary statistics for the first set of simulation studies. The true slope parameter is $\beta_0 = 0$. SE is the empirical standard error of the parameter estimator, SEE is the mean of the standard error estimators, CP is the coverage probability of the 95% confidence interval, and $\sigma^* = \sqrt{I^{-1}(\beta_0)/n}$ is the sample size scaled theoretical standard error under the fully efficient situation. ^{1}SEE: the standard error estimates by inverting the information matrix based on the efficient score function; ^{2}SEE: the standard error estimates by inverting the observed information matrix of all parameters including the "nuisance" parameters for estimating the log hazard function. (a): $0.5N(0,1)+0.5N(-1,0.5^2)$; (b): standard extreme-value; (c): Gumbel(-0.5γ,0.5); (d): Weibull(3,1).

Err.	Cen.	B-spline MLE				Log-rank		Gehan-weight		B-J		
dist	rate	Bias	SE	^{1}SEE (CP)	^{2}SEE (CP)	Bias	SE	Bias	SE	Bias	SE	σ^*
(a)	.00	.000	.090	.088 (.938)	.090 (.945)	-.003	.117	-.002	.096	-.002	.102	.085
	.25	-.004	.093	.092 (.937)	.093 (.942)	-.007	.114	-.004	.093	-.006	.101	.088
	.50	.001	.110	.105 (.940)	.106 (.946)	.001	.138	.001	.111	.001	.121	.105
(b)	.00	.003	.113	.107 (.934)	.109 (.942)	.002	.110	.002	.124	.002	.137	.109
	.25	.001	.131	.122 (.938)	.125 (.941)	.002	.128	.003	.151	.004	.161	.126
	.50	-.001	.166	.148 (.912)	.152 (.923)	.002	.161	.006	.189	.007	.194	.154
(c)	.00	.000	.057	.060 (.942)	.060 (.951)	.001	.083	.000	.063	.000	.070	.055
	.25	.002	.059	.058 (.946)	.058 (.947)	.002	.084	.002	.066	.001	.073	.058
	.50	.002	.072	.068 (.931)	.068 (.934)	.004	.100	.003	.079	.003	.088	.068
(d)	.00	.000	.037	.037 (.944)	.038 (.954)	.001	.040	.000	.037	.000	.036	.033
	.25	-.001	.039	.037 (.938)	.038 (.954)	-.001	.043	.000	.040	.000	.039	.036
	.50	-.001	.044	.040 (.914)	.042 (.934)	-.001	.048	.000	.044	.000	.043	.039

Table 3.2: Summary statistics for the second set of simulation studies. The true slope parameters are $\beta_1 = 1$ and $\beta_2 = 1$. SE, CP, ^{1}SEE, ^{2}SEE and σ^* have the same interpretation as these abbreviations in Table 3.1. (a): $N(0, 1)$; (b): standard extreme-value; (c): $0.5N(0, 1) + 0.5N(0, 3^2)$; (d): $0.95N(0, 1) + 0.05N(0, 3^2)$; (e): Gumbel($-0.5\gamma$,0.5); (f): $0.5N(0, 1) + 0.5N(-1, 0.5^2)$.

Err.			B-spline MLE				Log-rank		Gehan-weight		
dist	n		Bias	SE	^{1}SEE (CP)	^{2}SEE (CP)	Bias	SE	Bias	SE	σ^*
(a)	200	β_1	.003	.168	.149 (.912)	.155 (.924)	.000	.170	.002	.159	.155
		β_2	.003	.167	.153 (.928)	.156 (.928)	.004	.171	.002	.160	.156
	400	β_1	.006	.110	.108 (.948)	.110 (.950)	.005	.115	.008	.108	.110
		β_2	.001	.110	.109 (.944)	.110 (.945)	.002	.116	.001	.109	.110
	600	β_1	.001	.092	.088 (.939)	.090 (.943)	.001	.096	.002	.093	.090
		β_2	.005	.091	.089 (.945)	.090 (.944)	.005	.097	.003	.092	.090
(b)	200	β_1	-.009	.180	.154 (.894)	.161 (.903)	-.008	.168	-.007	.190	.165
		β_2	.004	.182	.162 (.903)	.163 (.915)	.005	.170	.005	.195	.169
	400	β_1	.000	.126	.113 (.914)	.115 (.923)	-.001	.124	.000	.143	.117
		β_2	.008	.118	.116 (.934)	.116 (.938)	.010	.116	.012	.135	.120
	600	β_1	.001	.102	.093 (.919)	.094 (.923)	.001	.100	.000	.114	.095
		β_2	.011	.098	.095 (.944)	.095 (.945)	.011	.097	.007	.114	.098
(c)	200	β_1	.014	.300	.281 (.930)	.279 (.924)	-.020	.315	-.019	.292	.259
		β_2	.000	.306	.285 (.916)	.282 (.918)	.002	.317	.002	.288	.260
	400	β_1	.034	.199	.206 (.955)	.200 (.949)	.002	.218	-.002	.201	.183
		β_2	-.003	.207	.208 (.949)	.202 (.942)	.009	.228	.012	.202	.184
	600	β_1	.035	.168	.171 (.957)	.165 (.949)	.003	.185	.001	.163	.150
		β_2	-.007	.169	.172 (.956)	.166 (.956)	-.004	.190	-.002	.168	.150
(d)	200	β_1	-.013	.172	.157 (.926)	.164 (.927)	-.010	.181	-.007	.166	.167
		β_2	-.004	.180	.160 (.908)	.164 (.913)	-.005	.184	-.005	.173	.166
	400	β_1	.003	.119	.113 (.944)	.116 (.948)	.004	.126	.006	.117	.118
		β_2	.003	.117	.114 (.942)	.116 (.953)	.004	.126	.003	.115	.118
	600	β_1	-.003	.097	.093 (.948)	.095 (.952)	-.002	.105	.002	.097	.096
		β_2	.001	.096	.094 (.942)	.095 (.944)	.002	.105	.003	.094	.096
(e)	200	β_1	-.002	.081	.077 (.944)	.078 (.946)	-.008	.109	-.006	.086	.079
		β_2	.000	.085	.080 (.929)	.078 (.934)	-.007	.119	-.004	.093	.080
	400	β_1	-.005	.055	.055 (.946)	.055 (.951)	-.003	.079	-.004	.061	.056
		β_2	.003	.055	.056 (.954)	.056 (.950)	.003	.081	.003	.063	.056
	600	β_1	-.003	.047	.045 (.940)	.045 (.938)	.000	.067	-.001	.052	.045
		β_2	-.001	.047	.046 (.944)	.045 (.943)	-.002	.066	-.001	.051	.046
(f)	200	β_1	-.002	.126	.117 (.918)	.120 (.929)	-.002	.159	-.001	.128	.119
		β_2	.000	.133	.120 (.917)	.121 (.926)	.002	.164	.001	.134	.116
	400	β_1	-.002	.087	.084 (.949)	.085 (.950)	.003	.114	.000	.091	.084
		β_2	.004	.086	.086 (.951)	.086 (.953)	.003	.111	.004	.090	.082
	600	β_1	.003	.074	.070 (.929)	.070 (.931)	.005	.101	.001	.074	.069
		β_2	.003	.074	.070 (.936)	.070 (.936)	.009	.104	.004	.075	.067

Table 3.3: Summary statistics for the second set of simulation studies with Weibull(0.5,1) and Weibull(2,1) as the error distributions. The true slope parameters are $\beta_1 = 1$ and $\beta_2 = 1$. $n\text{SE}^2$ is the multiplication of the sample size and the square of SE.

Err.			B-spline MLE					Log-rank			Gehan-weight		
dist	n		Bias	SE	SEE	CP	$n\text{SE}^2$	Bias	SE	$n\text{SE}^2$	Bias	SE	$n\text{SE}^2$
Weibull	100	β_1	-.038	.053	.056	.953	.281	-.004	.312	9.73	.001	.126	1.59
(0.5,1)		β_2	-.004	.053	.058	.985	.281	-.001	.317	10.05	-.002	.136	1.85
	200	β_1	-.011	.018	.017	.933	.065	.050	.184	6.77	.025	.075	1.13
		β_2	-.002	.015	.017	.977	.045	-.006	.194	7.53	-.004	.080	1.28
	400	β_1	-.011	.009	.011	.901	.032	.038	.116	5.38	.017	.043	.740
		β_2	-.001	.009	.011	.981	.032	.000	.113	5.11	.001	.044	.774
Weibull	100	β_1	-.002	.101	.091	.908	1.02	-.008	.125	1.56	-.001	.106	1.12
(2,1)		β_2	-.006	.104	.092	.907	1.08	-.004	.127	1.61	-.007	.107	1.14
	200	β_1	-.005	.067	.062	.925	.898	-.004	.087	1.51	-.002	.074	1.10
		β_2	-.003	.064	.063	.951	.819	-.003	.086	1.48	-.003	.072	1.04
	400	β_1	-.002	.045	.042	.927	.810	.003	.065	1.69	.002	.054	1.17
		β_2	-.002	.044	.042	.936	.774	-.001	.060	1.44	-.002	.052	1.08

Table 3.4: Regression parameter estimates and standard error estimates for $\log_{10}$ of time to death versus age at transplant and T5 mismatch score with $n = 157$ Stanford heart transplant patients. The proposed estimators (B-spline MLE) are compared with Gehan-weighted estimators reported from Jin et al. (2003) and Buckley-James estimators reported from Miller and Halpern (1982).

		B-spline MLE		Gehan-weight		Buckley-James	
	Covariate	Est.	SE	Est.	SE	Est.	SE
M. 1	Age	-0.0237	0.0068	-0.0211	0.0106	-0.015	0.008
	T5	-0.2118	0.1271	-0.0265	0.1507	-0.003	0.134
M. 2	Age	0.1022	0.0245	0.1046	0.0474	0.107	0.037
	Age^2	-0.0016	0.0004	-0.0017	0.0006	-0.0017	0.0005

Table 3.5: Accelerated failure time regression for the Mayor PBC data. The proposed estimators (B-spline MLE) are compared with Gehan-weighted estimators reported from Jin et al. (2003), least squares estimators reported from Jin et al. (2006) and kernel smoothed profile likelihood estimators reported from Zeng and Lin (2007).

	B-spline MLE		Gehan-weight		Least-squares		Kernel MLE (a_n^{opt})	
Parameter	Est.	SE	Est.	SE	Est.	SE	Est.	SE
Age	-0.0295	0.0058	-0.0258	0.0059	-0.0256	0.0063	-0.0286	0.0061
log(albumin)	1.8654	0.4314	1.5906	0.5352	1.6174	0.5409	1.6212	0.4761
log(bilirubin)	-0.6138	0.0606	-0.5789	0.0698	-0.5885	0.0752	-0.6175	0.0669
Edema	-0.6383	0.1930	-0.8781	0.2768	-0.8430	0.2604	-0.7985	0.3179
log(protime)	-2.3208	0.3072	-2.7680	0.9085	-2.3331	0.8543	-2.4095	0.8050

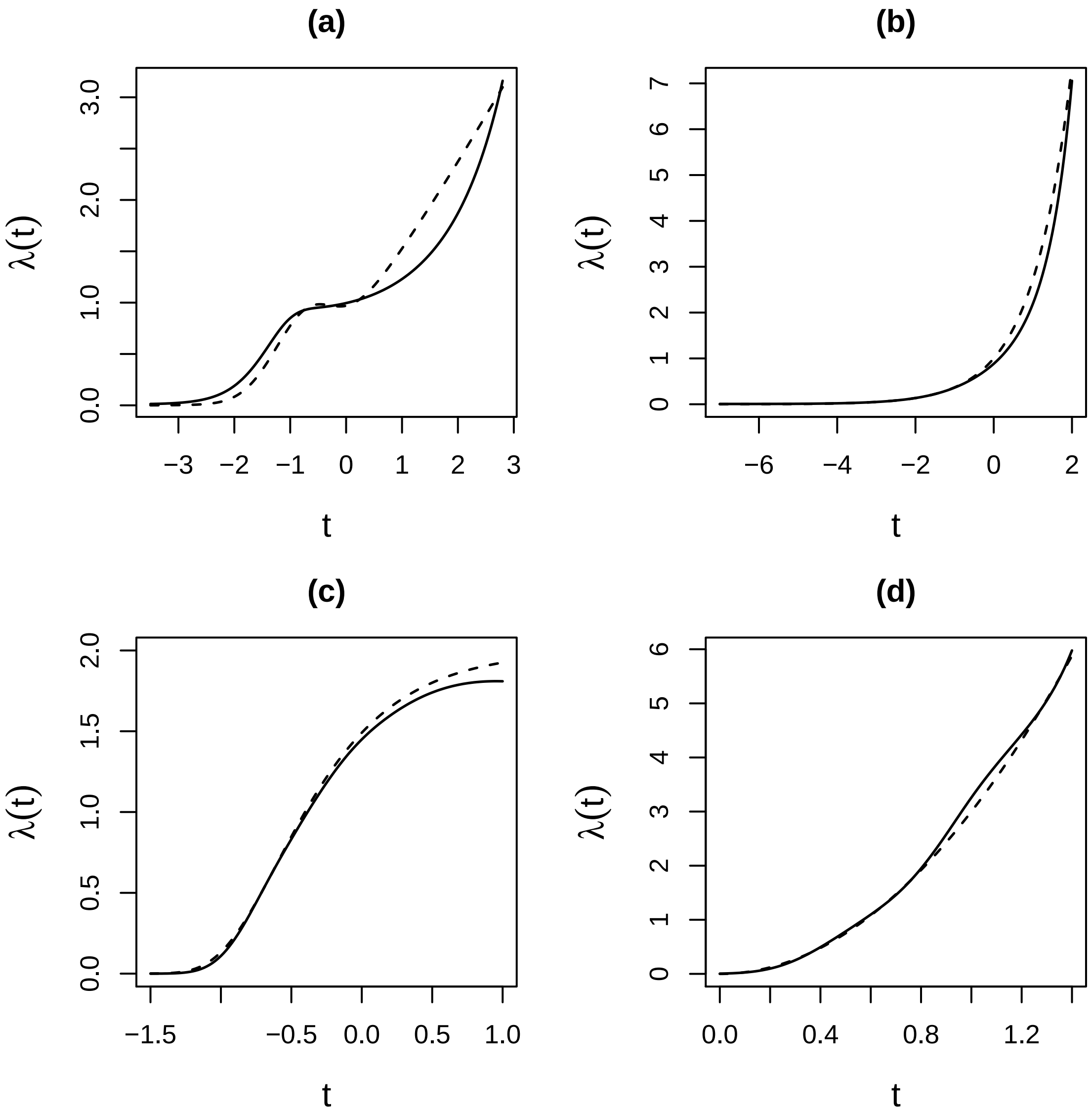

Figure 3.1: The solid lines are the estimated hazard functions ($\hat{\lambda}_n(t)$) by the proposed method for the model $\log T = \beta_0 X + e_0$, where $\beta_0 = 0$ and e_0 follows four distributions with (a): $0.5N(0,1) + 0.5N(-1, 0.5^2)$; (b): standard extreme-value; (c): Gumbel(-0.5γ,0.5) and (d): Weibull(3,1). The dashed lines are the true hazard functions $\lambda_0(t)$.

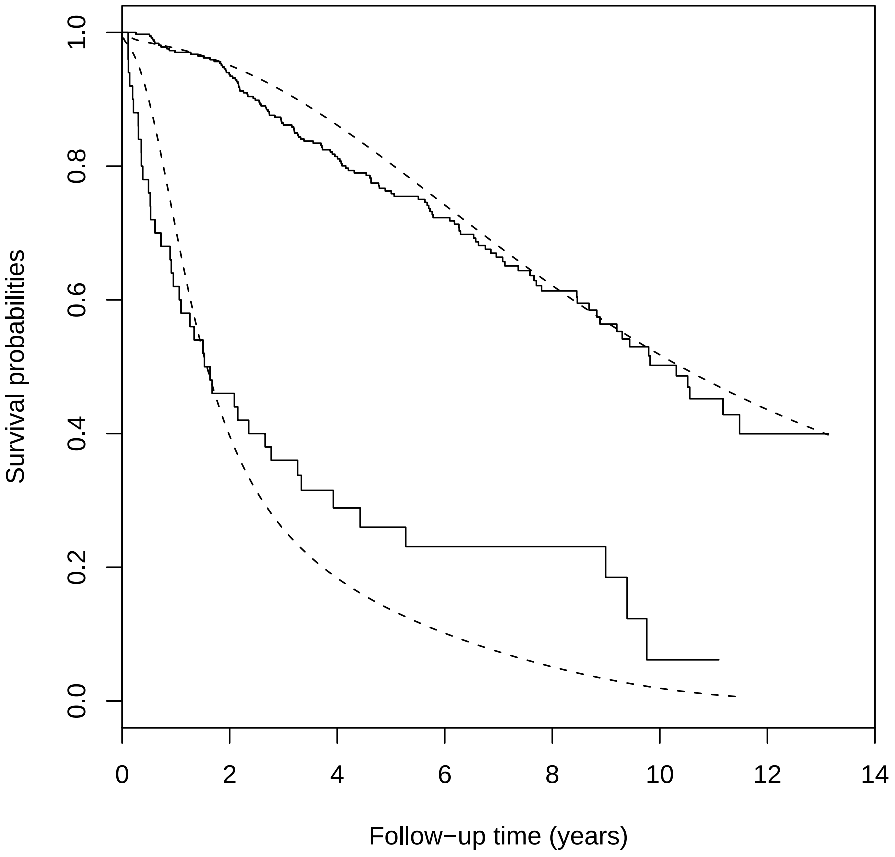

Figure 3.2: Estimated survival functions for the PBC data. The upper and lower solid lines are the Kaplan-Meier estimates for the patients with Edema=0 and Edema=1; the dashed lines are the corresponding model-based estimates.

CHAPTER IV

Future Work

The utility of novel biomarkers and clinical information for predicting future survival status for patients plays an important role in medical decision making. In addition to the survival probabilities, the actual survival times provide more straightforward information and can be crucial for treatment decision making as well. The result in Chapter 2 makes the prediction of survival times possible under a semiparametric linear model when covariate range is wide in practice. Then good statistical measures of model predictive performance are desirable. One type of commonly used prediction accuracy measures in the survival context includes Brier score (Brier 1950), integrated Brier score, time-dependent ROC curves (Heagerty et al. 2000), and etc. All of these measures use the time-dependent survival or event status, i.e., $I(T > t)$, as the quantity of interest. However, if the prediction of survival time is of primary interest, a measure that incorporates the survival time directly would be more attractive. The inverse probability of censoring weighted mean square error loss (Robins and Rotnitsky 1992) is one such measure. But it depends on an accurate estimate of the censoring distribution, which is not always the case. When truncation exists, this measure does not provide a consistent estimate of the mean square error loss. Developing straightforward yet reasonable prediction accuracy measures to evaluate the survival time prediction performance is worth exploration, and investigation in this direction is one of our future research plans.

The extended general theorem on the asymptotic normality of semiparametric M-estimators that has been proposed in Chapter 3 will be useful for the theoretical justification for other semiparametric models where the nuisance parameter is a function of the parameter of interest. One direct application would be the AFT model with other types of censoring such as the interval censoring and current status data. The likelihood functions in these cases are built upon residual survival times, which is a function of the slope parameters. Another application is the proportional hazards models with unknown link function (Wang 2004; Huang and Liu 2006). Instead of assuming the exponential form for the dependence of the hazard function on covariates as in the traditional Cox proportional hazards model, a more flexible relative risk form can be assume:

$$\lambda(t|X) = \lambda_0(t)\exp\{\psi(\beta_0'X)\},$$

where $\psi(\cdot)$, the link function, is an unknown smooth function. Clearly, the nuisance parameter $\psi(\cdot)$ in the partial likelihood is a function of the regression parameter β. Finally, the extended general theorem can be useful for proving the asymptotic normality of the parameter estimates from the single index models (Ichimura 1993) and generalized linear models with unknown links (Weisberg and Welsh 1994; Chiou and Muller 1998, 1999).

BIBLIOGRAPHY

BIBLIOGRAPHY

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review 78*, 1–3.

Buckley, J. and I. James (1979). Linear regression with censored data. *Biometrika 66*, 429–436.

Chiou, J. M. and H. G. Muller (1998). Quasi-likelihood regression with unknown link and variance functions. *Journal of the American Statistical Association 93*, 1376–1387.

Chiou, J. M. and H. G. Muller (1999). Nonparametric quasi-likelihood. *The Annals of Statistics 27*, 36–64.

Cox, D. R. (1972). Regression models and lifetables. *Journal of the Royal Statistical Society, Series B 34*, 187–220.

Fleming, T. R. and D. P. Harrington (1991). *Counting Processes and Survival Analysis.* New York: Wiley.

Fygenson, M. and Y. Ritov (1994). Monotone estimating equations for censored data. *The Annals of Statistics 22*, 732–746.

Geman, A. and C. Hwang (1982). Nonparametric maximum likelihood estimation by the method of seives. *The Annals of Statistics 10*, 401–414.

Heagerty, P. J., T. Lumley, and M. S. Pepe (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics 56*, 337–344.

Heller, G. and J. S. Simonoff (1990). A comparison of estimators for regression with a censored response variable. *Biometrika 77*, 515–520.

Huang, J. (1999). Efficient estimation of the partly linear additive cox model. *The Annals of Statistics 27*, 1536–1563.

Huang, J. Z. and L. Liu (2006). Polynomial spline estimation and inference of proportional hazards regression models with flexible relative risk form. *Biometrics 62*, 793–802.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single index models. *Journal of Econometrics 58*, 71–120.

Jin, Z., D. Y. Lin, L. J. Wei, and Z. Ying (2003). Rank-based inference for the accelerated failure time model. *Biometrika 90*, 341–353.

Jin, Z., D. Y. Lin, and Z. Ying (2006). On least-squares regression with censored data. *Biometrika 93*, 147–161.

Kalbfleisch, J. D. and R. L. Prentice (2002). *The Statistical Analysis of Failure Time Data, 2nd edition.* Hoboken, NJ: Wiley.

Lai, T. L. and Z. Ying (1988). Stochastic integrals of empirical-type processes with applications to censored regression. *Journal of Multivariate Analysis 27*, 334–358.

Lai, T. L. and Z. Ying (1991). Large sample theory of a modified Buckley-James estimator for regression analysis with censored data. *The Annals of Statistics 10*, 1370–1402.

Lin, D. Y. and C. J. Geyer (1992). Computational methods for semiparametric linear regression with censored data. *Journal of Computational and Graphical Statistics 1*, 77–90.

Lin, D. Y., J. M. Robins, and L. J. Wei (1996). Comparing two failure time distributions in the presence of dependent censoring. *Biometrika 83*, 381–393.

Lin, D. Y., L. J. Wei, and Z. Ying (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika 80*, 557–572.

Lin, D. Y., L. J. Wei, and Z. Ying (1998). Accelerated failure time models for counting processes. *Biometrika 85*, 605–618.

Miller, R. G. and J. Halpern (1982). Regression with censored data. *Biometrika 69*, 521–531.

Prentice, R. L. (1978). Linear rank tests with right censored data. *Biometrika 65*, 167–179.

Ritov, Y. (1990). Estimation in a linear regression model with censored data. *The Annals of Statistics 18*, 303–328.

Ritov, Y. and J. A. Wellner (1988). Censoring, martingales and the Cox model. In N.U.Prabhu (Ed.), *Statistical Inference from Stochastic Processes*, pp. 191–219. Providence, RI: American Mathematical Society.

Robins, J. and A. Rotnitsky (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In *Aids Epidemiology: Methodological Issues*, pp. 297–331. Birkhauser.

Schneider, H. and L. Weissfeld (1986). Estimation in linear models with censored data. *Biometrika 73*, 741–745.

Schumaker, L. (1981). *Spline Functions: Basic Theory.* New York: Wiley.

Shen, X. T. and W. H. Wong (1994). Convergence rate of seive estimates. *The Annals of Statistics 22*, 580–615.

Stone, C. (1985). Additive regression and other nonparametric models. *The Annals of Statistics 13*, 689–705.

Stone, C. (1986). The dimensionality reduction principle for generalized additive models. *The Annals of Statistics 14*, 1346–1370.

Stute, W. (1993). Consistent estimation under random censorship when covariables are available. *Journal of Multivariate Analysis 45*, 89–103.

Stute, W. (1996). Distributional convergence under random censorship when covariables are present. *Scandinavian Journal of Statistics 23*, 461–471.

Stute, W. and J. L. Wang (1993). The strong law under random censorship. *The Annals of Statistics 21*, 1591–1607.

Susarla, V., W. Y. Tsai, and J. Van Ryzin (1984). A Buckley-James-type estimator for the mean with censored data. *Biometrika 71*, 624–625.

Susarla, V. and J. Van Ryzin (1980). Large sample theory for an estimator of the mean survival time from censored samples. *The Annals of Statistics 8*, 1002–1016.

Tsiatis, A. A. (1990). Estimating regresion parameters using linear rank tests for censored data. *The Annals of Statistics 18*, 354–372.

van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge University Press.

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

Wang, S., B. Nan, J. Zhu, and D. G. Beer (2008). Doubly penalized Buckley-James method for survival data with high-dimensional covariates. *Biometrics 64*, 132–140.

Wang, W. (2004). Proportional hazards regression models with unknown link function and time-dependent covariates. *Statistica Sinica 14*, 885–905.

Wei, L. J., Z. Ying, and D. Y. Lin (1990). Linear regression analysis of censored survival data based on rank tests. *Biometrika 77*, 845–851.

Weisberg, S. and A. H. Welsh (1994). Adapting for the missing link. *The Annals of Statistics 22*, 1675–1700.

Wellner, J. A. (2007). On an exponential bound for the Kaplan-Meier estimator. *Lifetime Data Anal 13*, 482–496.

Wellner, J. A. and Y. Zhang (2007). Two likelihood-based semiparametric estimation methods for panel count data with covariates. *The Annals of Statistics 35*, 2106–2142.

Ying, Z. (1993). A large sample study of rank estimation for censored regression data. *The Annals of Statistics 21*, 76–99.

Yu, M. G. and B. Nan (2006). A hybrid Newton-type method for censored survival data using double weights in linear models. *Lifetime Data Analysis 12*, 345–364.

Zeng, D. and D. Y. Lin (2007). Efficient estimation for the accelerated failure time model. *Journal of the American Statistical Association 102*, 1387–1396.