ANALYSIS OF INTERVAL CENSORED DATA USING A LONGITUDINAL BIOMARKER

Noorie Hyun

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics in the Gillings School of Global Public Health.

Chapel Hill
2014

Approved by:

Dr. Donglin Zeng

Dr. David J. Couper

Dr. Jianwen Cai

Dr. Michael G. Hudgens

Dr. M. Alan Brookhart

ABSTRACT

NOORIE HYUN: Analysis of Interval Censored Data Using a Longitudinal Biomarker
(Under the direction of Dr. Donglin Zeng and Dr. David J. Couper)

In many medical studies, interest focuses on studying the effects of potential risk factors on some disease events, where the occurrence time of disease events may be defined in terms of the behavior of a biomarker. For example, in diabetic studies, diabetes is defined in terms of fasting plasma glucose being 126 mg/dl or higher. In practice, several issues complicate determining the exact time-to-disease occurrence. First, due to discrete study follow-up times, the exact time when a biomarker crosses a given threshold is unobservable, yielding so-called interval censored events. Second, most biomarker values are subject to measurement error due to imperfect technologies, so the observed biomarker values may not reflect the actual underlying biomarker levels. Third, using a common threshold for defining a disease event may not be appropriate due to patient heterogeneity. Finally, informative diagnosis and subsequent treatment outside of observational studies may alter observations after the diagnosis. It is well known that the complete case analysis excluding the externally diagnosed subjects can be biased when diagnosis does not occur completely at random.

To resolve these four issues, we consider a semiparametric model for analyzing threshold-dependent time-to-event defined by extreme-value-distributed biomarkers. First, we propose a semiparametric marginal model based on a generalized extreme value distribution. By assuming the latent error-free biomarkers to be non-decreasing, the proposed model implies a class of proportional hazards models for the time-to-event

defined for any given threshold value. Second, we extend the marginal likelihood to a pseudo-likelihood by multiplying the likelihoods over all observation times. Finally, to adjust for externally diagnosed cases, we consider a weighted pseudo-likelihood estimator by incorporating inverse probability weights into the pseudo-likelihood by assuming that external diagnosis depends on observed data rather than unobserved data. We estimate the three model parameters using the nonparametric EM, pseudo-EM and weighted-pseudo-EM algorithm, respectively.

Herein, we theoretically investigate the models and estimation methods. We provide a series of simulations, to test each model and estimation method, comparing them against alternatives. Consistency, convergence rates, and asymptotic distributions of estimators are investigated using empirical process techniques. To show a practical implementation, we use each model to investigate data from the ARIC study and the diabetes ancillary study of the ARIC study.

ACKNOWLEDGMENTS

I am extremely grateful to my advisors, Drs. Donglin Zeng and David J. Couper for their thoughtful guidance, insightful advice, and sincere encouragement during my dissertation research period. I know I would not be able to reach the end of this long journey without their help. The hands-on experience under supervision of my advisors have made me more confident as a researcher. I would also like to thank Drs. Donglin Zeng and David J. Couper for their financial support as well.

I would like to express my sincere gratitude to my committee members,Drs. Jianwen Cai, Michael G. Hudgens, and M. Alan Brookhart for their constructive and perceptive comments on my dissertation papers.

I want to extend my sincere thanks to the faculty for their brilliant teaching and staff for kind help of the Department of Biostatistics at UNC-Chapel Hill. I send my gratitude to my former coworkers at the UNC Collaborative Studies Coordinating Center, where I had a research assistantship for the last two and half years. I want to express my special thanks to all my friends and family for their love, support, and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER1: INTRODUCTION

Many longitudinal studies of chronic disease such as cancer, AIDS, and diabetes monitor patients for biomarkers, as an indicator of disease occurrence, in order to investigate potential associations between risk exposures and time to disease occurrence. For disease events determined by some biomarker and threshold, when interval between visits is long or patients miss visits, the exact date of the event that an individual's biomarker value crosses the threshold is unobservable. Instead, what is usually known are the latest and earliest visit dates at which an individual's biomarker value crosses a given threshold. Such data is called interval censored data. Using the interval rather than the exact date of event occurrence may lead to invalid inferences (Lindsey and Ryan 1998).

Most biomarkers measurement has variation and the variation consists of short-term intra-individual variability and measurement error. Assay variability and within-person effects complicate determination of whether an individual's biomarker has actually exceeded the threshold. In clinical practice, *ad hoc* approaches that are used to take into account biomarker variability include taking two or more measurements over a period of time. Regarding measurement error, for example, the National Institute of Standards and Technology maintains the blood sample materials as the gold standard and provides guidelines for instrument manufactures to determine the accuracy of their measurement devices. If measurement error is non-ignorable but ignored in the analysis, the analysis may yield an inaccurate conclusion.

Furthermore, biomedical studies have several limitations to the use of a single

threshold that are usually ignored in practice. The threshold is generally regarded as a fixed constant that is appropriate for everyone; however, this may not be appropriate to all biomarkers. For instance, hypercholesterolemia does not cause symptoms but can significantly increase risk of developing coronary heart disease (CHD). To reduce risk, including that of CHD, people with substantially elevated cholesterol levels are advised to start therapeutic lifestyle changes or drug therapy. The cholesterol level at which to consider therapeutic intervention varies across different risk categories such as smoking, hypertension, age, etc. (the National Cholesterol Education Program Expert Panel 2001)

Finally, informative diagnosis outside of observational studies, which causes alter observations after the diagnosis. It is well known that the complete case analysis excluding the externally diagnosed subjects can be biased when diagnosis does not occur completely at random (Ibrahim et al. 2005).

We are motivated by the Atherosclerosis Risk in Communities (ARIC) study and an ancillary ARIC study, which present the problems described above. The ARIC Study recruited a population-based cohort from four U.S. communities, namely, Forsyth County, NC, Jackson, MS, suburbs of Minneapolis, MN, and Washington County, MD. Participants underwent a baseline examination in 1987-1989 had three follow-up examinations at approximately three-year intervals, and a further examination in 2011-2013. The ARIC Study was designed to investigate the causes of atherosclerosis, and hypercholesterolemia is a crucial risk factor for atherosclerosis. Hence, assessing risk factors associated with time-to-hypercholesterolemia is of interest. An ancillary study of the ARIC study investigated type 2 diabetes mellitus The standard ARIC definition of diabetes is having a fasting plasma glucose (FPG) $\geq 126 mg/dL$, non-fasting glucose $\geq 200 mg/dL$, a self-reported physician diagnosis of diabetes, or use of diabetes medication in the two weeks prior to the study visit. The outcome variables of the two studies are

time-to-disease occurrence, diabetes or hypercholesterolemia, and in the dissertation, we focus on the time until the biomarkers reach the corresponding threshold levels.

To resolve these four issues, we consider a semiparametric model for analyzing threshold-dependent time-to-event defined by extreme-value-distributed and longitudinal biomarkers and break down the problems into the three steps:

(1) *Threshold-Dependent Proportional Hazards Model for Analyzing Time-to-Event Defined by Biomarker with Subject to Measurement Error*: to mitigate the problems, we concentrate on the first follow-up visit after baseline and ignore the informative external diagnosis altogether. We propose a semiparametric model based on a generalized extreme value distribution for the time-to-disease occurrence. By assuming the latent error-free biomarkers to be non-decreasing, the proposed model has a natural class of proportional hazards models for the time-to-event defined for any given threshold value. To account for the additive measurement errors, we estimate the model parameters using the nonparametric maximum likelihood approach.

(2) *Semiparametric Regression Model for Analyzing Time-to-Event Defined by Extreme Longitudinal Biomarkers*: the model proposed in the first step is extended to model the longitudinal biomarkers at follow-ups by constructing a pseudo-likelihood, which is multiplying the marginal likelihoods at follow-ups.

(3) *Weighted Pseudo-Likelihood for Adjusting Informative Diagnosis: an Application to Time-to-Hypercholesterolemia in the ARIC study*: to adjust for cases with external diagnosis, we consider a weighted pseudo-likelihood estimator by incorporating inverse probability weights into the pseudo-likelihood proposed in the second step by assuming that external diagnosis depends on observed data rather

than unobserved data. We employ a marginal structure model based on auxiliary information and subject's status at the previous visits to predict an external diagnosis.

We estimate the three model parameters via the nonparametric Expectation Maximization (EM), pseudo-EM, and weighted-pseudo-EM algorithm, respectively. In this dissertation, we theoretically investigate the models and estimation methods. We provide a series of simulations, to examine each model and estimation method comparing them with the existing methods. Consistency, convergence rates, and asymptotic distribution of estimators are investigated using the empirical process techniques. We illustrate the first marginal model by applying it to data from the diabetes ancillary ARIC study and the other two models by applying those to data from the ARIC study.

In Chapter 2, existing methods to address each problem, interval censored data, measurement error in response, and missing data are reviewed. In Chapter 3 to 5, we elaborate the three methods briefly described in (1) to (3). Conclusion with a discussion on the proposed three methods and future work are contained in Chapter 6.

CHAPTER2: LITERATURE REVIEW

Interval censoring in survival analysis is a generalized scheme of left or right censoring. Observed exact failure times practically correspond to narrow intervals (Turnbull 1976, Kalbfleisch and Prentice 2002). In left or right censoring, the probability that exact failure time is observed is positive; however, we are unable to observe it at all in interval censored data. Therefore, statistical methods and inferences for interval censored data are more complicated than left or right censored data (Huang and Wellner 1997).

Commonly used methods in survival analysis with right censoring such as the Kaplan-Meier estimator for survival functions and the partial likelihood for the Cox proportional hazards (PH) model are inapplicable to interval censored data due to the incomplete event times. Moreover, it is difficult to incorporate counting processes and martingale theory into interval censored data, and this leads to the need for alternatives for investigating asymptotic properties. One alternative is using empirical processes requiring advanced mathematical techniques (Zhang and Sun 2010). Furthermore, unlike most semiparametric models for right censored data as nuisance parameters infinite dimensional parameters are not removed from the inference for regression parameters.

Interval censoring can be classified into four types: current status data, case 2 interval censored data, panel count data, and a mixed case. Along with the definition of each interval censoring type, corresponding survival function estimators and regression models are summarized in the following subsections.

2.1 Interval Censored Data

2.1.1 Current Status Data

When only one observation time is applied and each patient is known to experience the onset of the event either before or after the observation time, the data are called as *case 1* interval censored data or *current status data*. Current status data often occur in cross-sectional studies when the outcome is a mile-stone event such as the onset of chronic disease. Also, current status data are easily found in animal studies such as tumorigenicity experiments on nonlethal tumors (Hoel and Walburg 1991).

For the subject i with a vector of covariates $\boldsymbol{X}_i$, let T_i be the unobservable failure time and V_i be the examination or observation time. Then the observed data of the subject i are $(V_i, \delta_i, \boldsymbol{X}_i)$ denote as $\boldsymbol{W}_{1,i}$, where $\delta_i = I(T_i \leq V_i)$. It is assumed that T is independent of V given $\boldsymbol{X}$. In addition, the joint distribution of $(V, \boldsymbol{X})$ is assumed to be independent on $\boldsymbol{\theta}$-that is a vector of coefficients for the covariate $\boldsymbol{X}$-and any unspecified non-decreasing baseline function of T.

Survival Estimation with Current Status Data

In this section, nonparametric maximum likelihood estimators (NPMLEs) for the survival or distribution function of current status data are reviewed. Denote the ordered observed times by $\{V_{(i)} | i = 1, \ldots, n\}$, that is, $V_{(i)} \leq V_{(i+1)}$ for $i = 1, \ldots, n-1$.

The observed log likelihood function for current status data $\{(V_i, \delta_i) | i = 1, \ldots, n\}$ is

$$l_n(F) = \sum_{i=1}^{n} \{\delta_i \log F(V_i) + (1 - \delta_i) \log(1 - F(V_i))\}, \tag{2.1}$$

where $F(\cdot)$ is the distribution function of T.

Maximizing the log likelihood in (2.1) with respect to $\{F(V_i)\}_{i=1}^{n}$ is equivalent to

minimizing

$$\sum_{i=1}^{n} n_i \left[\frac{\delta_i}{n_i} - F(V_{(i)}) \right]^2, \text{ subject to } F(V_{(1)}) \le \ldots \le F(V_{(n)}), \tag{2.2}$$

where $n_1 = n$ and $n_i = n-i+1$ (Robertson et al. 1988). The NPMLE for F are determined only at the observation times $\{V_i\}$ with $\delta_i = 1$, $1 \le i \le n$. Let $\{s_j\}_{j=1}^m$ be the uniquely ordered observation times at which $\delta_i = 1$ for $1 \le i \le n$. The set of values of $\{F(s_j)\}_{j=1}^m$ that minimizes (2.2) is referred to as the isotonic regression of $\{1/n_1, \ldots, 1/n_m\}$ with weights $\{n_1, \ldots, n_m\}$

We can find a NPMLE for F minimizing (2.2) by various approaches. Using the max-min formula for isotonic regression, Ayer et al. (1955) obtained the explicit forms for $\{\hat{F}(s_j)\}_{j=1}^m$ as

$$\hat{F}(s_j) = \max_{u \le j} \min_{v \ge j} \frac{\sum_{l=u}^{v} \delta_l}{v - u + 1}. \tag{2.3}$$

They also introduced the pool adjacent violators algorithm (PAVA) and recommended this algorithm rather than direct calculation of the formula in (2.3) to facilitate the computation.

Huang and Wellner (1997) proposed an algorithm: after plotting $(i, \sum_{j=1}^{i} \delta_{(j)}), i = 1, \ldots, n$ and forming the Greatest Convex Minorant (GCM), G^* of the points in the plot, then left-derivative of G^* at i is calculated for $\hat{F}_n(V_{(i)})$, $i = 1, \ldots, n$. This algorithm obtains the same NPMLE as the max-min formula in (2.3). The GCM algorithm is faster than the PAVA algorithm from a small to a large sample size except when the left truncation probability is over 0.85 (Zhang and Newton 1997).

The NPMLEs calculated through the max-min formula, the PAVA algorithm, and the GCM algorithm are mutually equivalent and consistent under certain regularity conditions (Ayer et al. 1955, Huang and Wellner 1997).

Regression Model with Current Status Data

Regression analysis of survival data is used to quantify the effect of some covariates on the survival time or to predict the survival probabilities for new individuals. In this section, we review commonly used regression models for current status data such as the Cox proportional hazards (PH) models, proportional odds models, additive hazard models, and accelerated failure time (AFT) models, etc.

The observed log likelihood for current status data is given by

$$l_n(F|\boldsymbol{W}_1) = \sum_{i=1}^{n}\{\delta_i \log F(V_i|\boldsymbol{W}_{1,i}) + (1-\delta_i)\log(1-F(V_i|\boldsymbol{W}_{1,i}))\}, \tag{2.4}$$

where $F(t|\boldsymbol{W}_1)$ is the distribution function of T given the observed data.

Current status data almost allows us to obtain explicit forms of the efficient influence function and semiparametric efficient variance for regression parameters.

Each regression model is the special case of the following transformation model. The transformation model postulates that the conditional distribution $F(t|\boldsymbol{x})$ of T given the covariates $\boldsymbol{X} = \boldsymbol{x}$ satisfies

$$g(F(t|\boldsymbol{x})) = h(t) + \boldsymbol{\theta}^T\boldsymbol{x}, \tag{2.5}$$

where g is a specified function; $h(t)$ is an unknown non-decreasing function; $\boldsymbol{\theta}$ is the unknown finite d-dimensional regression parameter.

First, if we take $g(s) = \log[-\log(1-s)]$, $0 < s < 1$, then (2.5) results in the proportional hazards model by Cox (1972). The Cox proportional hazards model has been the most commonly used for survival analysis due to the availability of efficient inference procedures that are implemented in all statistical software packages. The model postulates

$$\lambda(t|\boldsymbol{X} = \boldsymbol{x}) = \lambda_0(t)e^{\boldsymbol{\theta}^T\boldsymbol{x}} \tag{2.6}$$

for the hazard function of the survival time T given the covariate $\boldsymbol{x}$, where $\lambda_0(t)$ denotes an unknown baseline hazard function. The regression parameter of θ provides the log hazard ratio of x on time to the event. Between two levels of the covariate $\boldsymbol{X}$, the PH model constrains the ratio of the hazards to be constant over time. The model in (2.6) can be cast as a transformed linear model of $\log \int_0^t \lambda(u)du = -\boldsymbol{\theta}^T\boldsymbol{X} + \epsilon$, where ϵ follows the extreme value distribution of $1 - \exp(-e^{\epsilon})$ (Dabrowska and Doksum 1988).

The observed log likelihood under the model in (2.6) is

$$l_n(\boldsymbol{\theta}, \Lambda|\boldsymbol{W}_1) = \sum_{i=1}^{n} \Big\{ \delta_i \log \big[1 - \exp\big(-\Lambda(V_i) \exp(\boldsymbol{\theta}^T\boldsymbol{X}_i) \big) \big] - (1 - \delta_i) \exp(\boldsymbol{\theta}^T\boldsymbol{X}_i) \Lambda(V_i) \Big\},$$

where $\Lambda(t) = \int_0^t \lambda(u)du$.

Related to PH regression models for current status data, Huang (1996) provided very influential and thorough study. He obtained a maximum profile likelihood estimator (profile-MLE) for $(\boldsymbol{\theta}, \Lambda)$ by the iterative convex minorant algorithm (this algorithm will be reviewed in detail in section 2.1.2). The consistency, asymptotic normality, and semiparametric efficiency of the profile-MLE for regression parameters were established under certain regularity conditions. The convergence rate of the estimators $\widehat{\Lambda}$ dominates the convergence rate of $(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda})$ by $n^{1/3}$. Nonetheless, it was shown that the regression parameter estimates asymptotically converges to normal distribution in the rate of $\sqrt{n}$. The profile likelihood method requires intensive computation for data with large covariates.

Second, if we take $g(s) = \text{logit}(s) \equiv \log[s/(1-s)]$ for $0 < s < 1$ in the regression model of (2.5), then we obtain a proportional odds regression model:

$$\text{logit}\big[F(t|\boldsymbol{X} = \boldsymbol{x})\big] = h(t) + \boldsymbol{\theta}^T\boldsymbol{x}. \tag{2.7}$$

Let $h(t) = \text{logit}F(t|\boldsymbol{X} = 0)$, the baseline non-decreasing log odds function. Then θ_k

is the log odds ratio for two samples with unit difference in the kth covariate. In the proportional odds model, the hazard ratio for two samples is not constant over time but converges to unity as time t increases. The model (2.7) can be described as a transformed linear model of $h(t)$=$\boldsymbol{\theta}^T\boldsymbol{X} + \epsilon$, where ϵ has the logistic distribution of $[1 + exp(-\epsilon)]^{-1}$ (Dabrowska and Doksum 1988). For right censored data, Bennett (1983a;b) provided a proportional odds model and a log-logistic regression model. Pettitt (1984) suggested several levels of specification about $h(t)$ in the proportional odds model (2.7).

The observed log likelihood function under the model (2.7) is

$$l_n(\boldsymbol{\theta}, h|\boldsymbol{W}_1) = \sum_{i=1}^{n} \delta_i \left\{h(V_i) + \boldsymbol{\theta}^T\boldsymbol{X}_i\right\} - \log\left[1 + \exp\left\{h(V_i) + \boldsymbol{\theta}^T\boldsymbol{X}_i\right\}\right]. \tag{2.8}$$

Rossini and Tsiatis (1996) suggested a semiparametric proportional odds model for current status data using an approximate maximum likelihood. The approximate likelihood is replacing $h(t)$ in the likelihood (2.8) with a non-decreasing step function. The maximum likelihood estimator (MLE) maximizing the approximate likelihood can be viewed as a sieve MLE based on the sieve of non-decreasing continuous piecewise constant functions. They showed consistency, asymptotic normality, and semiparametric efficiency of the regression parameters estimates under certain regularity conditions and provided the explicit form of the asymptotic variance for the estimates.

Third, we consider an accelerated failure time model:

$$\log(T) = \boldsymbol{\theta}^T\boldsymbol{X} + \epsilon, \tag{2.9}$$

where the distribution function F of ϵ is completely unspecified and ϵ is i.i.d. We transform the failure time by logarithm to avoid restriction on the distribution for ϵ; however, the failure time can be transformed by any appropriate functions. This

model can be written as $F\{\log(T)|\boldsymbol{X}\} = F\{\log(T) - \boldsymbol{\theta}^T\boldsymbol{X}\}$ in terms of a conditional distribution.

The observed log likelihood under the model (2.9) is

$$l_n(F|\boldsymbol{W}_1) = \sum_{i=1}^{n}\left[\delta_i \log F(\log V_i - \boldsymbol{\theta}^T\boldsymbol{X}_i) + (1-\delta_i)\log\left\{1 - F(\log V_i - \boldsymbol{\theta}^T\boldsymbol{X}_i)\right\}\right]. \quad (2.10)$$

Huang and Wellner (1997) proposed a profile-MLE under the AFT model. They showed consistency of the profile-MLE and provided the information bound for $\boldsymbol{\theta}$ under certain regularity conditions; however, left an open problem about the convergence rate of $\widehat{\boldsymbol{\theta}}$. The estimated MLE of $F(\cdot|\boldsymbol{\theta})$ for each fixed $\boldsymbol{\theta}$ is not smooth, and it results in the non-smooth profile likelihood with respect to $\boldsymbol{\theta}$, so the convergence rate is unspecified yet.

Tian and Cai (2006) constructed an estimator under the AFT model by inverting a Wald-type statistics for testing a null proportional hazards. Due to the equivalence between two assumptions: residual in (2.9) is independent of $\boldsymbol{X}$; $S_\epsilon(t|\boldsymbol{X}) = S_0(t)^{\exp(\gamma^T\boldsymbol{X})}$, the regression parameters can be estimated by solving the estimating equation, $\widehat{\boldsymbol{\gamma}}(\boldsymbol{\theta}) = o_p(n^{-1/2})$, where $\widehat{\boldsymbol{\gamma}}$ is the NPMLE of Huang (1996). Using the semiparametric efficient variance of $\boldsymbol{\gamma}$, $\boldsymbol{B}$ calculated by Huang (1996), the asymptotic variance of $\widehat{\boldsymbol{\theta}}$ can be approximated by sandwich variance, $\boldsymbol{D}^{-1}\boldsymbol{B}(\boldsymbol{D}^T)^{-1}$, where $\boldsymbol{D} = d\boldsymbol{\gamma}_0(\boldsymbol{\theta})/d\boldsymbol{\theta}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$. The estimator was proved to be consistent under certain regularity conditions.

Finally, if we take $g(s) = -\log(1-s)$, $0 < s < 1$, in the regression model of (2.5), then we have an additive hazard regression model:

$$\lambda(t|\boldsymbol{X} = \boldsymbol{x}) = \lambda_0(t) + \boldsymbol{\theta}^T\boldsymbol{x}(t), \quad (2.11)$$

where $\lambda_0(t)$ is an unspecified baseline hazard function. This model describes the association between the failure time and covariates in difference between two hazards.

Lin et al. (1998) proposed an additive hazard model based on a counting process and martingale theory for current status data. When the counting process is defined as $N_i(t) = (1 - \delta_i)I(V_i \leq t)$, which jumps by unit whenever the subject i is observed at time t and found still to be failure-free, the probability that the counting process has one is: under the assumption that T and V are independent

$$dH_i(t) = e^{-\boldsymbol{\theta}^T \boldsymbol{X}^*(t)} dH_0(t), \tag{2.12}$$

where $dH_0(t) = e^{-\Lambda_0(t)} d\Lambda_V(t)$ and $\boldsymbol{X}_i^*(t) = \int_0^t \boldsymbol{X}_i(s)ds$. This form is the Cox proportional hazards model and this mediates using the partial likelihood principle to estimate the regression parameters. When the assumption-that is independence of V and T-is changed to the more flexible assumption that V is independent of T given $\boldsymbol{X}$, they formulated the association through the proportional hazards model. The latter model improves efficiency but does not achieve the semiparametric efficiency.

Related to other regression models for current status data, Shen (2000) proposed a linear regression model using a constructed random-sieve likelihood and constraints that combine benefits of a semiparametric likelihood with estimating equations. It was assumed that ϵ in the linear model, $T = \boldsymbol{\theta}^T \boldsymbol{X} + \epsilon$, is independent of $(V, \boldsymbol{X})$; ϵ has zero mean and a finite variance; the true residual ($\epsilon = T - \boldsymbol{\theta}^T \boldsymbol{X}$) and the observed residual ($\epsilon(\boldsymbol{\theta}) = V - \boldsymbol{\theta}^T \boldsymbol{X}$) have the same support. For inference, the asymptotic distribution for the regression parameter estimates and the profile likelihood ratio test statistics were obtained. Graphical tools for model diagnostics were proposed.

Ma and Kosorok (2005) extended Huang (1996)'s model for current status data by adding a smooth nonparametric covariate effect: $\lambda(t|\boldsymbol{X}) = \lambda_0(t)e^{\boldsymbol{\theta}^T \boldsymbol{X} + a(u)}$, where $a(u)$ is an unknown smooth function of a continuous variable u. To resolve issues arising in carrying out this extension, they used a nonparametric maximum penalized

log likelihood,

$$l_n^p(\boldsymbol{\theta}, a, H) \equiv \sum_{i=1}^{n} l(\boldsymbol{\theta}, a, H|\boldsymbol{W}_{1,i}) - \lambda_n^2 J^2(a),$$

where the log likelihood $l(\boldsymbol{\theta}, a, H|\boldsymbol{W}_{1,i})$ is the log likelihood in (2.4) with the conditional distribution function $F(V_i|\boldsymbol{W}_{1,i})$=$F\{\boldsymbol{\theta}^T\boldsymbol{X}_i + a(u) + H(V_i)\}$ and H is an unknown non-decreasing transformation. They suggested a sieve approximation for the nonparametric covariate effect $a(u)$ and showed that the cumulative sum diagram approach as discussed by Groeneboom and Wellner (1992) works for general transformation models. For the convergence rate, the estimator for the nonparametric transformation H achieves the optimal rate of $n^{1/3}$, but it slows down the convergence of $\widehat{a}$ in ordinary spline settings. The penalized MLE for $\widehat{\boldsymbol{\theta}}$ is asymptotically normal in the convergence rate of $\sqrt{n}$ and is efficient. The semiparametric efficient variance for $\widehat{\boldsymbol{\theta}}$ was obtained, and the block jackknife method was suggested for the asymptotic variance estimation.

Ma (2009) applied Ma and Kosorok (2005)'s approach to current status data from heterogeneous mixture population such as the mixture population of a cured subgroup and a disease susceptibility subgroup. A generalized linear model for the cure probability is applied. For subjects not cured, both of the linear Cox model and the partly linear Cox model were considered to model the survival risk. Under the assumptions and the partly linear model, the conditional survival function is $S(t|\boldsymbol{X}, \boldsymbol{Z}, u) = p(\boldsymbol{Z}) + \{1 - p(\boldsymbol{Z})\}\exp\{-\Lambda_0(t)e^{\boldsymbol{\theta}^T\boldsymbol{X}+a(u)}\}$, where the cure probability $p(\boldsymbol{Z}) = g^{-1}(\boldsymbol{\alpha}^T\boldsymbol{Z})$; $g(\cdot)$ is a known link function; $\boldsymbol{\alpha}$ and $\boldsymbol{Z}$ are the unknown regression parameter and covariates in the generalized linear model, respectively. It was shown that the regression parameters estimators are consistent, asymptotically normal, and efficient under certain regularity conditions. The nonparametric baseline function and covariate effect can be estimated with the convergence rate of $n^{1/3}$. The weighted bootstrap was proposed for the asymptotic variance estimation of $\widehat{\boldsymbol{\theta}}$.

2.1.2 Case 2 Interval Censored Data

When more than one observation time is applied and each patient is known to experience the onset of the event of interest either before the first observed time, between the two observation times, or after the last observation time, such data are called *case 2* interval-censored data. Longitudinal studies with periodic follow-up often produce case 2 interval censored data. For the subject i with a vector of covariates $\boldsymbol{X}_i$, let two observation times be given by V_{Li} and V_{Ui}, where $V_{Li} < V_{Ui}$. The observed data of the subject i are $(\delta_{1i}, \delta_{2i}, V_{Li}, V_{Ui}, \boldsymbol{X}_i)$ denote as $\boldsymbol{W}_{2,i}$, where $\delta_{1i} = I(T_i \leq V_{Li})$ and $\delta_{2i} = I(V_{Li} < T_i \leq V_{Ui})$.

We assume that T is independent of (V_L, V_U) given $\boldsymbol{X}$ and that (V_L, V_U) are random variables from a distribution with support $\big\{(v_L, v_U)|0 < \tau_L \leq v_L,\ v_U \leq \tau_U < \infty,\ v_U \geq v_L + c\big\}$, where c is a positive constant. In addition, the joint distribution of $(V, \boldsymbol{X})$ is assumed to be independent of $\boldsymbol{\theta}$ and any unspecified non-decreasing baseline function of T.

Survival Estimation with Case 2 Interval Censored Data

Unlike current status data, the NPMLE for the distribution function of case 2 or general interval censored data has no explicit form available, so using of iterative algorithm is inevitable.

Turnbull (1976) used the expectation maximization (EM) algorithm for incomplete data due to grouping, general censoring and/or truncation, and this corresponds to the self-consistency introduced by Efron (1967). For case 2 interval censored data, $\{[V_{Li}, V_{Ui}]|i = 1, \ldots, n\}$, when T_i is truncated by $B_i \subseteq R$ and $[V_{Li}, V_{Ui}] \subset B_i$ for the subject i, the likelihood is proportional to

$$l_n(F) = \prod_{i=1}^{n} \big[F(V_{Ui}) - F(V_{Li})\big]/P_F(B_i). \tag{2.13}$$

If T_i is not truncated, then $B_i{=}R$ with $P(B_i) = 1$.

Let $\{l_j\}_{j=1}^m$ and $\{r_j\}_{j=1}^m$ denote the unique ordered elements of $\{V_{Ui}|i = 1,\ldots,n\}$ and $\{V_{Li}|i = 1,\ldots,n\}$ respectively and satisfy $l_1 \le r_1 < l_2 \le r_2 < \ldots < l_m \le r_m$. That is, for each $j, 1 \le j \le m$, $l_j{=}L_i$ for some i, $1 \le i \le n$ and $r_j{=}R_k$ for some k, $1 \le k \le n$. Hence, two or more intervals of $\{[V_{Li}, V_{Ui}]\}_{i=1}^n$ include $[l_j, r_j]$. Define $p_j = F(r_j) - F(l_j) \ge 0$ for $1 \le j \le m$ and $\sum_{j=1}^m p_j = 1$.

Using the fact that any distribution function increasing outside $\bigcup_{j=1}^m [l_j, r_j]$ cannot be a maximum likelihood estimate of F except in the trivial case when $[V_{Li}, V_{Ui}] \bigcap \bigcup_{j=1}^m [l_j, r_j] = B_i \bigcap \bigcup_{j=1}^m [l_j, r_j]$ for all i, the problem maximizing (2.13) reduces to maximizing the following likelihood with respect to the vector of $\mathbf{p} = (p_1,\ldots,p_m)$

$$L_n(p_1,\ldots,p_m) = \prod_{i=1}^n \frac{\sum_{j=1}^m \alpha_{ij} p_j}{\sum_{j=1}^m \beta_{ij} p_j}, \text{ subject to } \sum_{j=1}^m p_j = 1, p_j \ge 0, \tag{2.14}$$

where $\alpha_{ij} = I([l_j, r_j] \subset [V_{Li}, V_{Ui}])$ and $\beta_{ij} = I([l_j, r_j] \subset B_i)$ for $1 \le i \le n$ and $1 \le j \le m$.

The proportion of observations in interval $[l_j, r_j]$ to be used in the EM algorithm is given by

$$\frac{\sum_{i=1}^n \left(\mu_{ij} + \nu_{ij}\right)}{\sum_{i=1}^n \sum_{j=1}^m \left(\mu_{ij} + \nu_{ij}\right)} = \pi_j(\mathbf{p}), \tag{2.15}$$

where $\mu_{ij}(\mathbf{p})$ and $\nu_{ij}(\mathbf{p})$ are the probabilities that the exact time is in $[l_j, r_j]$ and that the interval $[l_j, r_j]$ is truncated respectively. Hence, $\mu_{ij}(\mathbf{p}) = \alpha_{ij} p_j / \sum_{k=1}^m \alpha_{ik} p_k$ and $\nu_{ij}(\mathbf{p}) = (1 - \beta_{ij}) p_j / \sum_{k=1}^m \beta_{ik} p_k$.

The vector of probabilities $\mathbf{p}$ is called *self-consistent* if $p_j = \pi_j(\mathbf{p})$, $1 \le j \le m$. The self-consistent algorithm is an example of the EM algorithm. It was shown that the self-consistent algorithm converges monotonically.

The candidates obtained from the self-consistency algorithm is not guaranteed to be the NPMLE. Gentleman and Geyer (1994) provided easily verifiable conditions, the Lagrange multiplier criterion for the self-consistent estimator to be the NPMLE. They

also provided sufficient conditions for the uniqueness of the NPMLE. This result is similar to the approach of Peto (1973) for interval censoring with no truncation using the constrained Newton-Raphson (NR) method.

Groeneboom (1991) characterized the NPMLE for case 2 interval censored data even though the NPMLE has no closed form. Groeneboom and Wellner (1992) introduced the iterative convex minorant (ICM) algorithm for NPMLEs. Jongbloed (1998) described the ICM algorithm in its general form and showed that it does not converge under mild regularity conditions and proposed a modified version by adding a line search into the algorithm so that it achieves global convergence.

Compared with the ICM algorithm, the EM algorithm converges rather slowly to the solution of the optimization problem; however, global convergence of the EM algorithm under certain regularity conditions was proved by Dempster et al. (1977) and Wu (1983). For censoring problems, a combination of the EM and ICM algorithm was proposed in Zhan and Wellner (1995). Simulation results indicate this hybrid algorithm to behave very well for the double censoring model.

Hudgens et al. (2001) extended Turnbull (1976)'s NPMLE to the general setting of competing risks allowing for any number of failure types and for each failure time to be subject to interval censoring and truncation. The cumulative incidence function NPMLE gives rise to an estimate of the survival function that can be undefined over a potentially larger set of regions than the NPMLE of the marginal survival function. Alternatively, a pseudo-likelihood estimator was considered. Without truncation, the pseudo-likelihood estimate of the cumulative incidence function has fewer undefined regions than the NPMLE of the cumulative incidence function. However, when truncation is included, the result has trade-off. Consistency of the NPMLEs of cumulative incidence functions was proved by Hudgens et al. (2007).

Hudgens (2005) adapted the graph theories to characterize the support set of the

NPMLE of a survival function and derived conditions for the existence of the NPMLE when data are interval censored and left-truncated. These results help to explain the NPMLEs' underestimation of survival functions in practice.

Regression Model with Case 2 Interval Censored Data

The observed log likelihood function for case 2 interval censored data is

$$\begin{aligned} l_n(F|\boldsymbol{W}_2) &= \sum_{i=1}^{n}\Big[\delta_{1i}\log\{F(V_{Li}|\boldsymbol{W}_{2,i})\}+\delta_2\log\{F(V_{Ui}|\boldsymbol{W}_{2,i})-F(V_L|\boldsymbol{W}_{2,i})\} \\ &\quad +(1-\delta_{1i}-\delta_{2i})\log\{1-F(V_{Ui}|\boldsymbol{W}_{2,i})\}\Big], \end{aligned} \tag{2.16}$$

where $F(t|\boldsymbol{W}_2)$ is a conditional distribution function of T given the data.

First, for the Cox PH model with case 2 interval censored data, Finkelstein (1986) proposed a MLE based on the approach of Turnbull (1976) discarding truncation. By replacing p_j in (2.14) with $\exp(-\exp(\boldsymbol{\theta}^T\boldsymbol{X}+\gamma_{lj}))-\exp(-\exp(\boldsymbol{\theta}^T\boldsymbol{X}+\gamma_{rj}))$, where $\gamma_{lj}=\log\{\Lambda(l_j)\}$ and $\gamma_{rj}=\log\{\Lambda(r_j)\}$, the log likelihood (2.14) is re-expressed in terms of the Cox PH model. Then the MLE is calculated by treating the log likelihood function as the one arising from a parametric model, that is, considering the observation time as a discrete random variable. The score function and the information bound for the MLE are easily obtained; however, the relevant asymptotic property was not figured out.

Huang and Wellner (1997) proposed a MLE under the Cox PH model using the log likelihood in (2.16), where $F(t|\boldsymbol{X})=1-\exp\{-\Lambda(t)\exp(\boldsymbol{\theta}^T\boldsymbol{X})\}$. The consistency, asymptotic normality, and efficiency of the MLE were established under certain regularity conditions. They suggested the use of the observed Fisher information or the curvature of the profile likelihood to estimate the asymptotic variance of $\widehat{\boldsymbol{\theta}}$ (Murphy and van der Vaart 2000).

Second, in the proportional odds model for case 2 interval censored data, Huang and Rossini (1997) presented a sieve maximum likelihood estimator for the regression parameter based on the observed log likelihood in (2.16) with $F(t|\boldsymbol{W}_2) = \exp\{h(t) + \boldsymbol{\theta}^T\boldsymbol{X}\}/\{1+\exp(h(t)+\boldsymbol{\theta}^T\boldsymbol{X})\}$. The sieve used by Huang and Rossini (1997) is the collection of non-decreasing continuous piecewise linear functions. Asymptotic properties the estimator were thoroughly established. The form of semiparametric efficient variance of $\widehat{\boldsymbol{\theta}}$ is intractable since the efficient score function has no closed form. Instead, they proposed an alternative to estimate the variance matrix of $\widehat{\boldsymbol{\theta}}$ by the inverse of the curvature of the profile likelihood at the estimate of the regression parameter.

Third, for the AFT model with case 2 interval censored data, Huang and Wellner (1997) proposed a profile MLE, which is similar to the approach for current status data by Huang and Rossini (1997). The MLE is based on the log likelihood (2.16) by letting $F(t|\boldsymbol{X})$ be $F(t-\boldsymbol{\theta}^T\boldsymbol{X})$. In contrast to current status data, the information bound for the regression parameter has no explicit expression. Moreover, since the information calculation includes an integral equation with a singular kernel, the Fredholm theory of integral equations cannot be directly applied. Instead, this equation is similar to the one encountered in calculating the information for smooth functionals of the distribution function in the NPMLE setting, and this is solved by Geksus and Groeneboom (1996a;b; 1999).

Tian and Cai (2006) extended their approach explained in section 2.1.1 to case 2 interval censored data using Huang and Wellner (1997)'s MLE based on the PH model for case 2 interval censored data.

Finally, Zeng et al. (2006) provided an additive model using the log likelihood in (2.16) in which $F(t|\boldsymbol{X})$ is replaced with $1-\exp\{-\Lambda(t)-\boldsymbol{\theta}^T\boldsymbol{X}(t)\}$, where $\boldsymbol{X}(t) = \int_0^t \boldsymbol{X}(u)du$. The MLE can be derived by maximizing the log likelihood under the constraint that $\exp(-\Lambda(t))$ holds monotonicity in the uniquely ordered observation

times. Since the information bound for the regression parameter has no explicit form, alternatively, Zeng et al. (2006) proposed to use the curvature of the profile likelihood for the information estimation. The consistency, asymptotic normality, and semiparametric efficiency of the estimator were shown.

2.1.3 Panel Count Data

For recurrent event data such as tumor or disease symptoms, if the occurrence process is observed only at discrete time points, what is only known are the numbers of the event occurrences between observation times. Such data are referred to *panel count data* (Sun 2006). The observation data for the subject i consist of $\boldsymbol{W}_{3,i} = (K_i, \boldsymbol{V}_{i,K_i}, \mathbb{N}_{i,K_i}, \boldsymbol{X}_i)$, where K_i is a random number of random times $0 \equiv V_{i,K_i,0} < V_{i,K_i,1} < \ldots < V_{i,K_i,K_i}$, $\boldsymbol{V}_{i,K_i} = (V_{i,K_i,1}, \ldots, V_{i,K_i,K_i})$, and $\mathbb{N}_{i,K_i} = (\mathbb{N}(V_{i,K_i,1}), \ldots, \mathbb{N}(V_{i,K_i,K_i}))$ for a univariate counting process $\mathbb{N}(t), t > 0$. It is assume that K and V_K is conditionally independent of the counting process $\mathbb{N}$ given a covariate vector $\boldsymbol{X}$.

Survival Estimation with Panel Count Data

For panel count data, a non-homogeneous Poisson process for the counting process is often assumed. The marginal distributions of $\mathbb{N}$ is $P(\mathbb{N}(t) = k) = \exp\{-\Lambda_0(t)\}\Lambda_0(t)^k/k!$, where $\Lambda_0(t) = E\{\mathbb{N}(t)\}$ the mean function of the counting process of $\mathbb{N}$. Then the log pseudo-likelihood function ignoring the dependence is

$$l_n^{ps}(\Lambda) = \sum_{i=1}^{n} \sum_{j=1}^{K_j} \big\{\mathbb{N}(V_{i,K_i,j}) \log \Lambda(V_{i,K_i,j}) - \Lambda(V_{i,K_i,j})\big\}. \tag{2.17}$$

According to the definition of a non-homogeneous Poisson process, the increments of the counting process are independent. The marginal distribution of $\triangle\mathbb{N}$ is $P\{\triangle\mathbb{N}(s,t) =$

$k\} = \exp\{- \triangle \Lambda_0(s,t)\}\{\triangle\Lambda_0(s,t)\}^k/k!$ and the log likelihood function under the assumption is followed by

$$l_n(\Lambda) = \sum_{i=1}^{n}\sum_{j=1}^{K_j}\Big\{ \triangle\, \mathbb{N}(V_{i,K_i,j}) \log \triangle\Lambda(V_{i,K_i,j}) - \triangle\Lambda(V_{i,K_i,j})\Big\}. \tag{2.18}$$

Wellner and Zhang (2000) studied both a nonparametric maximum pseudo-likelihood estimator based on (2.17) and a nonparametric maximum likelihood estimator based on (2.18) using the assumption that the counting process is a non-homogeneous Poisson process. They showed that the maximum pseudo-likelihood estimator is exactly the one proposed by Sun and Kalbfleisch (1995). The two estimators were established to be consistent, and both estimators at a fixed time point have the asymptotic distribution of a two-sided Brownian motion process starting from zero. The NPMLE is more efficient than the maximum pseudo-likelihood estimator, but its computation is more difficult.

Sen and Banerjee (2007) constructed a pseudo-likelihood ratio statistic from (2.17) for testing the value of the distribution function at a fixed time point and showed that this converges to a known limit distribution-that can be expressed as a function of different convex minorants of a two-sided Brownian motion process with parabolic drift-under the null hypothesis and certain regularity conditions. Unlike the Wald-based approach, the likelihood-ratio-based method excludes nuisance parameter estimation and provides an extremely clear-cut way of constructing confidence intervals for the survival rate at a fixed time point. Simulation result comparing confidence intervals showed the estimation based on the pseudo-likelihood ratio is superior to the estimates based on the limit distribution of the maximum pseudo-likelihood estimator with kernel-based estimation of nuisance parameters and sub-sampling with appropriate block-size in terms of precision.

Lu et al. (2007) proposed two NPMLEs based on the log likelihood in (2.18) and the log pseudo-likelihood in (2.17) using monotone polynomial splines to ease intensive computation required in Wellner and Zhang (2000)'s approach. $I-$spline basis functions were used to linearly span the class of polynomial splines, and the non-negativity and monotonicity of the $I-$splines are guaranteed by the non-negativity of coefficients (Ramsay1988). The generalized Rosen algorithm proposed by Zhang and Jamshidian (2004) was used to compute the estimators. The proposed spline likelihood/pseudo-likelihood-based estimators are consistent and have faster convergence rate than $n^{1/3}$ when the true baseline hazard function is sufficiently smooth. Simulation study showed that the two estimators have smaller variance and mean square error than their alternatives proposed by Wellner and Zhang (2000).

Regression Model with Panel Count Data

Wellner and Zhang (2007) established two likelihood-based semiparametric estimators with the Cox model that is the mean function of a counting process. The pseudo-likelihood and likelihood from which the two models are derived are

$$\begin{aligned} l_n^{ps}(\boldsymbol{\theta}, \Lambda) &= \sum_{i=1}^{n} \sum_{j=1}^{K_j} \big\{ \mathbb{N}(V_{i,K_i,j}) \log \Lambda(V_{i,K_i,j}) + \mathbb{N}_i(V_{i,K_i,j}) \boldsymbol{\theta}^T \boldsymbol{X}_i \\ &\quad - e^{\boldsymbol{\theta}^T \boldsymbol{X}_i} \Lambda(V_{i,K_i,j}) \big\}, \end{aligned} \tag{2.19}$$

$$\begin{aligned} l_n(\boldsymbol{\theta}, \Lambda) &= \sum_{i=1}^{n} \sum_{j=1}^{K_j} \big\{ \triangle\, \mathbb{N}(V_{i,K_i,j}) \log \triangle \Lambda(V_{i,K_i,j}) + \triangle \mathbb{N}(V_{i,k_i,j}) \boldsymbol{\theta}^T \boldsymbol{X}_i \\ &\quad - e^{\boldsymbol{\theta}^T \boldsymbol{X}_i} \triangle\, \Lambda(V_{i,K_i,j}) \big\}, \end{aligned} \tag{2.20}$$

under the assumptions that the counting process and the increment of the counting process are a non-homogeneous Poisson process. For the asymptotic variance estimation,

bootstrap resampling was utilized because the asymptotic variance is too complicated to be consistently estimated. The asymptotic properties of consistency, convergence rate, and asymptotic normality of the both models were established under certain regularity conditions. The proposed semiparametric estimation methods are robust against the underlying conditional Poisson process assumption. Simulation studies provided that the maximum likelihood method based on the Poisson process assumption is more efficient than the pseudo-likelihood method both on and off the Poisson model.

Lu et al. (2009) was motivated by the advantage that the spline likelihood estimators of Lu et al. (2007) outperform the semiparametric estimators proposed by Wellner and Zhang (2007) in view of the convergence rate and performance at finite samples. They developed semiparametric likelihood-based methods for panel count data using $B-$spline approximation for the cumulative hazard function in the models (2.19) and (2.20) in order to ease the intensive computation in the bootstrap semiparametric inference procedure utilized by Wellner and Zhang (2007). The monotonicity of the resulting spline function is guaranteed by imposing non-decreasing constraints on the coefficients. It was shown that the proposed spline-based likelihood estimator of the cumulative hazard function is consistent and asymptotic normal under certain regularity conditions. The ease of computing spline estimators make the statistical inference based on the bootstrap procedure feasible. Moreover, the spline estimation is insensitive to selection of the number and placement of the knots.

Although the independent random-censorship model is often reasonable, in many situations the censoring process is linked to the failure time process. For example, the termination date for a medical trial is not fixed before the study commences but is chosen later, with the choice influenced by the results of the study up to that time. Sun and Wei (2000) proposed a semiparametric regression model for analyzing panel count data when both observation and censoring times may depend on covariates. One

limitation of this approach is that both the observation time process and censoring time depend on the event time process, so if we stop following up the subject immediately after the occurrence of a certain number of events, the proposed method is inapplicable.

2.1.4 Mixed Case of Interval Censored Data

In the mixed-case interval censoring model each individual is followed up for a number of times, where the number and the times of observation can vary from person to person (Schick and Yu 2000). It is determined between which two consecutive observation times that the event of interest occurs. Current status data or case 2 interval censored data are special cases of the mixed-case interval censored data.

Hudgens et al. (2007) compared three nonparametric estimators of the joint distribution function for a survival time and a continuous mark variable in view of the uniqueness and consistency of NPMLE when the survival time is interval censored and the mark variable may be missing for the interval-censored observations. The three estimators compared are the NPMLE, estimators based on midpoint imputation, and estimators based on discretizing the mark variable. The estimator obtained by discretizing the mark variable results in interval-censored competing risks survival data for which the NPMLE characterized by Hudgens et al. (2001). Regardless of whether the mark variable is missing, the estimators based on discretizing the mark variable is consistent, whereas the NPMLE and the estimators based on midpoint imputation are inconsistent under certain regularity conditions.

Ma (2010) extends Ma (2009)'s Cox PH linear model for current status data to the one for mixed case of interval censored data with a cured subgroup. Identifiability and the asymptotic properties of consistency and weak convergence were established under certain regularity conditions, and the inference based on the weighted bootstrap was investigated because information matrix has no explicit form.

Random Effect Model with Interval Censored Data

In longitudinal data or clustered data, correlation among failure times is of interest. Frailty models have been proposed to accommodate the correlation. Frailty models specify the intra-subject correlation explicitly through an unobservable random variable (frailty). For a commonly used frailty model, it is assumed that the failure times given the frailty are independent and the conditional hazard given the frailty U_i is $\lambda_{ik}(t|U_i) = U_i\lambda_0(t)\exp(\boldsymbol{\theta}^T\boldsymbol{X}_{ik})$ for the ith cluster and kth observation, where $\{U_i\}_{i=1}^n$ are i.i.d.

While frailty models with right censored data have been studied by many researchers, frailty models for interval censored data have been less developed. Almost all regression models for correlated data with interval censoring use parametric approaches to describe the covariate effects although semiparametric models can be more flexible.

Li and Ma (2010) developed two-part models, which consist of the cure process and event process. The cure rate is described in a generalized linear model, and the survival rate is expressed in a location-scale parametric model including normal, logistic and Gumbel distributions. Each model includes one random effect to account for correlations between measurements. The cure rate depends on a random effect, as a consequence, the cure rate may change over time. Semiparametric models to address both the cure and event processes simultaneously need to be considered.

Interval censored and clustered data often occur in dental studies. Exact dates of tooth eruption and caries occurrence are practically unobservable. Moreover, when the response variable is time from tooth eruption to caries occurrence, doubly interval censoring occurs. Also, teeth of a subject are correlated. Komárek and Lesaffre (2007) was motivated by such dental data and proposed a Bayesian approach for an accelerated failure time model with interval censored data. The likelihood contribution of the ith

cluster is given by

$$L_i = \int_{\mathbb{R}^q} \Big\{ \prod_{k=1}^{n_i} \int_{V_{L,i,k}}^{V_{U,i,k}} f(v - \boldsymbol{\theta}^T \boldsymbol{X}_{i,k} - \boldsymbol{b}^T \boldsymbol{Z}_{i,k}) dv \Big\} g(\boldsymbol{b}) d\boldsymbol{b}, \tag{2.21}$$

where $\boldsymbol{X}$ is a vector of covariates for fixed effects; $\boldsymbol{\theta}$ is the unknown vector of regression coefficients; $\boldsymbol{b}_i$ is a vector of random effects with the density $g(\boldsymbol{b})$; $\boldsymbol{b}_i$s are i.i.d for $1 \leq i \leq n$; $\boldsymbol{Z}$ is a vector of covariates for random effects. The density of the error f is assumed to follow a penalized normal mixture distribution with unspecified components and the density of random effect g is assumed to follow multivariate normal distribution. The prior distributions for mean, variance, and the covariance matrix are assumed to be normal, inverse-gamma, and inverse-Wishart, respectively. Simulation results showed that the estimators nearly correct estimate the shape of the survival curves, and the regression parameter estimates have acceptable bias and precision.

Komárek and Lesaffre (2008) suggested an accelerated failure time model with random effects taking account of correlated observations and doubly interval censoring in the failure time from tooth emergence to caries occurrence. The assumed model is given by

$$log(E_{i,k}) = d_i + \boldsymbol{\delta}^T \boldsymbol{Z}_{i,k} + \zeta_{i,k}, \tag{2.22}$$

$$log(T_{i,k}) = b_i + \boldsymbol{\beta}^T \boldsymbol{X}_{i,k} + \epsilon_{i,k}, \tag{2.23}$$

where E is the chronological emerging time; T is the time to caries occurrence; the two times of $E_{i,k}$ and $T_{i,k}$ are independent for each i and k; $\boldsymbol{\delta}$ and $\boldsymbol{\beta}$ are the unknown regression parameter; $\zeta_{i,k}$, $\epsilon_{i,k}$, b_i, and d_i are mutually independent for all i and k. The likelihood contribution of the ith cluster is given by

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Bigg[\prod_{k=1}^{n_i} \int_{V_{L,i,k}^{(E)}}^{V_{U,i,k}^{(E)}} \Bigg\{ \int_{V_{L,i,k}^{(T)} - e_{i,k}}^{V_{U,i,k}^{(T)} - e_{i,k}} p(t_{i,k}|b_i) dt_{i,k} \Bigg\} p(e_{i,k}|d_i) de_{i,k} \Bigg] p(d_i, b_i) db_i dd_i,$$

where the density p is assumed to be a penalized normal mixture distribution with an overspecified number of components and non-informative priors for hyperparmeters are used. For sensitivity analysis, a model assuming that (b_i, d_i) follows bivariate normal distribution was also considered, and the proposed estimate is robust against the underlying correlation of (b_i, d_i) assumption. Simulation results showed that the regression parameter are estimated with only small bias and reasonable precision. The shape of the survivor curves is correctly estimated. However, the both approaches provided by Komárek and Lesaffre (2007; 2008) can not handle time-varying covariates.

2.2 Measurement Error in Data

In regression analysis, measurement error in response variables is mingled with the error residual, so it is generally ignored or less focused than error in predictor variables (Abrevaya and Hausman 2004). In this section, we primarily concentrate on statistical modeling to account for error-prone dependent variables. Let $\boldsymbol{Y}$ indicate the response variable without error and $\boldsymbol{S}$ be the observed response variable with measurement error. Let $\boldsymbol{X}$ be observed covariates without measurement error. The measurement error process is specified by modeling the relationship between $\boldsymbol{Y}$ and $\boldsymbol{S}$, possibly depending on $\boldsymbol{X}$. This is called measurement error model. The classical error model is an additive model, $\boldsymbol{S} = \boldsymbol{Y} + \boldsymbol{U}$, where $\boldsymbol{U}$ has mean zero and finite variance, and is independent of $\boldsymbol{Y}$ such that $E(\boldsymbol{S} \mid \boldsymbol{Y}) = \boldsymbol{Y}$. An alternative model is the Berkson error model: the model connect $\boldsymbol{Y}$ and $\boldsymbol{S}$ as $\boldsymbol{Y} = \boldsymbol{S} + \boldsymbol{U}$, where $\boldsymbol{U}$ has mean zero and finite variance and is independent of $\boldsymbol{S}$. In the Berkson model $E(\boldsymbol{Y} \mid \boldsymbol{S}) = \boldsymbol{S}$, and $\boldsymbol{S}$ is said to be an unbiased predictor of $\boldsymbol{Y}$ (Guolo 2008). Models for the unobserved variable $\boldsymbol{Y}$ can be interpreted in two ways: it is a functional method if $\boldsymbol{Y}$ is modeled as parameters, whereas it is a structural method if $\boldsymbol{Y}$ is regarded as random variables. If the density or mass function for $\boldsymbol{S}$ given $(\boldsymbol{Y}, \boldsymbol{X})$, $f_{\boldsymbol{S}|\boldsymbol{Y},\boldsymbol{X}}(s \mid y, x, \gamma)$ depends only on

the true response, that is, $f_{\boldsymbol{S}|\boldsymbol{Y},\boldsymbol{X}}(s \mid y, x, \gamma) = f_{\boldsymbol{S}|\boldsymbol{Y}}(s \mid y, \gamma)$, then $\boldsymbol{S}$ is called surrogate response, and the error is called non-differential.

Generally, the likelihood function for the observed response is

$$f_{\boldsymbol{S}|\boldsymbol{X}}(s \mid x, \boldsymbol{B}, \boldsymbol{\gamma}) = \int f_{\boldsymbol{Y}|\boldsymbol{X}}(s \mid y, x, \boldsymbol{B}) f_{\boldsymbol{S}|\boldsymbol{Y},\boldsymbol{X}}(s \mid y, x, \boldsymbol{\gamma}) dy, \tag{2.24}$$

where $\boldsymbol{B}$ and $\boldsymbol{\gamma}$ are unknown parameters. If $\boldsymbol{S}$ is a surrogate, the second density function in (2.24) is replaced by $f_{\boldsymbol{S}|\boldsymbol{Y}}(s \mid y, \boldsymbol{\gamma})$. Hence, we can use naive methods to test whether there is association between the predictors and the true response, if $\boldsymbol{S}$ is a surrogate. However, note that we lose power in contrast to tests derived from true response (Prentice 1989). The likelihood in (2.24) shows that we need to model the distribution of response error model. Usually, additional information is needed for the identifiability of the parameters for the error model. It is called validation data.

Suppose that there is validation subsample data obtained by measuring the true response in the primary sample selected with probability $\pi(\boldsymbol{S}, \boldsymbol{X})$. Let an indicator variable, $\triangle_i = 1$ if subject i's true response is measured in the validation data, $\triangle_i = 0$ otherwise. As taking account of the validation data, the observed likelihood for a general $\boldsymbol{S}$ is

$$\prod_{i=1}^{n} \left[\{ f(\boldsymbol{S}_i \mid \boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{\gamma}) f(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \boldsymbol{B}) \}^{\triangle_i} \{ f(\boldsymbol{S}_i \mid \boldsymbol{X}_i, \boldsymbol{B}, \boldsymbol{\gamma}) \}^{1-\triangle_i} \right]. \tag{2.25}$$

Here the distribution of $\boldsymbol{S} \mid (\boldsymbol{Y}, \boldsymbol{X})$ is a crucial component. This likelihood approach requires a correctly specified model for the measurement error. Calculation of the likelihood could be challenging in implementation of maximizing the likelihood.

In this review, we restrict our interest to the surrogate and observed response variables.

2.2.1 Linear Regression with Response Error

Suppose $\boldsymbol{Y} \mid \boldsymbol{X}$ follows a normal linear model with mean $\beta_0 + \boldsymbol{\beta}_X^T \boldsymbol{X}$ and variance σ_ϵ^2, while $\boldsymbol{S} \mid (\boldsymbol{Y}, \boldsymbol{X})$ follows a normal linear with mean $\gamma_0 + \gamma_1 \boldsymbol{Y}$ and variance σ_U^2. Then $\boldsymbol{S}$ is biased, and the observed data follow a normal linear model with mean $\gamma_0 + \beta_0\gamma_1 + \gamma_1 \boldsymbol{\beta}_X^T \boldsymbol{X}$ and variance $\sigma_U^2 + \gamma_1^2 \sigma_\epsilon^2$. Thus naive regression ignoring measurement error in $\boldsymbol{S}$ estimates $\gamma_1 \boldsymbol{\beta}_X$ rather than $\boldsymbol{\beta}_X$.

To make $\boldsymbol{S}$ unbiased variables, we can transform $\boldsymbol{S}$ to $(\boldsymbol{S} - \gamma_0)/\gamma_1$. If we obtain information about (γ_0, γ_1), we can apply any existing analysis method to an estimated unbiased response as $(\boldsymbol{S} - \widehat{\gamma}_0)/\widehat{\gamma}_1$. Suppose that there exists available validation data. Buonaccorsi and Tosteson (1993) and Buonaccorsi (1996) proposed the following procedure. We use the validation subsample data to obtain the estimates $\boldsymbol{B}$, the parameters relating $\boldsymbol{Y}$ and $\boldsymbol{X}$, and (γ_0, γ_1). Denote the estimator obtained from the validation data as $\widehat{\boldsymbol{B}}_1$ and the estimator obtained from the analysis based on the original data with the transformed estimator $(\boldsymbol{S} - \widehat{\gamma}_0)/\widehat{\gamma}_1$ as $\widehat{\boldsymbol{B}}_2^T$. Then we estimate the joint covariance matrix of these estimates, $(\widehat{\boldsymbol{B}}_1^T, \widehat{\boldsymbol{B}}_2^T)$ using the bootstrap, and it is called Σ. We form the best weighted combination of the two estimates, namely $\widehat{\boldsymbol{B}} = (\boldsymbol{J}^T \Sigma^{-1} \boldsymbol{J})^{-1} \boldsymbol{J}^T \Sigma^{-1} (\widehat{\boldsymbol{B}}_1^T, \widehat{\boldsymbol{B}}_2^T)^T$, where $\boldsymbol{J} = (\boldsymbol{I}, \boldsymbol{I})$ and $\boldsymbol{I}$ is the $r \times r$ identity matrix, r is sum of the dimensions of $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$. An estimated covariance matrix for the combined estimates $\widehat{\boldsymbol{B}}$ is $(\boldsymbol{J}^T \widehat{\Sigma}^{-1} \boldsymbol{J})^{-1}$.

If there is no validation data, instead one might have two independent replicate unbiased measurements of $\boldsymbol{Y}$ denoted by $(\boldsymbol{S}_{1*}, \boldsymbol{S}_{2*})$. These unbiased replicates are in addition to the biased surrogate $\boldsymbol{S}$ measured on the main study sample. In this case, we use the same algorithm as for validation data, with the following changes (Carroll 2006): we use the unbiased response the average of $\boldsymbol{S}_{1*}$ and $\boldsymbol{S}_{2*}$ to get $\widehat{\boldsymbol{B}}_1$. In fact, the replication data is modeled: $\boldsymbol{S} = \gamma_0 + \gamma_1 \boldsymbol{Y} + V$ and $\boldsymbol{S}_{j*} = \boldsymbol{Y} + \boldsymbol{U}_{j*}$ for $j = 1$ and 2, where U_{1*} and U_{2*} are independent with mean zero, and V has mean zero and finitie variance.

We have considered homoscedastic regression models so far, but when the data are heteroscedastic, in the additive unbiased response measurement error model the variance function for $\boldsymbol{Y}$ has general form of $\sigma_\epsilon^2 g^2(\boldsymbol{X}, \boldsymbol{B})$. However, we can keep applying the same procedure used for homoscedastic data with the changed variance form.

2.2.2 Logistic Regression with Response Error

Response error in binary dependent variables is called misclassification. Assuming misclassification is independent of $\boldsymbol{X}$, we classify observed responses with probabilities $pr(\boldsymbol{S} = 1 \mid \boldsymbol{Y} = 1, \boldsymbol{X}) = \pi_1$ and $pr(\boldsymbol{S} = 0 \mid \boldsymbol{Y} = 0, \boldsymbol{X}) = \pi_0$. Then

$$pr(\boldsymbol{S} = 1 \mid \boldsymbol{X}) = (1 - \pi_0) + (\pi_1 + \pi_0 - 1)H(\beta_0 + \boldsymbol{\beta}_X^T \boldsymbol{X}), \tag{2.26}$$

where $pr(\boldsymbol{Y} = 1 \mid \boldsymbol{X}) = H(\beta_0 + \boldsymbol{\beta}_X^T \boldsymbol{X})$, and $H(x) = \exp(x)/\{1 + \exp(x)\}$.

If the misclassification probabilities are unknown and independent of the covariates, then the parameters $(\pi_1, \pi_0, \beta_0, \boldsymbol{\beta}_X)$ can be estimated by using the following log-likelihood function: let the probability in (2.26) be $\Psi(\boldsymbol{S} = 1, \boldsymbol{X}, \pi_1, \pi_0, \beta_0, \boldsymbol{\beta}_X^T)$,

$$\begin{aligned} &\sum_{i=1}^{n} \big[\boldsymbol{S}_i \log\{\Psi(\boldsymbol{S} = 1, \boldsymbol{X}, \pi_1, \pi_0, \beta_0, \boldsymbol{\beta}_X^T)\} \\ &+ (1 - \boldsymbol{S}_i) \log\{1 - \Psi(\boldsymbol{S} = 1, \boldsymbol{X}, \pi_1, \pi_0, \beta_0, \boldsymbol{\beta}_X^T)\} \big]. \end{aligned} \tag{2.27}$$

There are many existing algorithms to maximize the log-likelihood in (2.27): scoring, iteratively reweighted least squares, and the EM-algorithm (Carroll 2006). In practice, it is difficult to identify the classification probabilities. Paulino et al. (2003) resolved this identifiability problem by using informative prior distribution under Bayesian framework.

When one has validation data, the classification probabilities can be estimated as the proportion of correct classification among each group with $\boldsymbol{Y} = 1$ or 0. The estimates

$\widehat{\pi}_1$ and $\widehat{\pi}_0$ are incorporated into the log-likelihood in (2.27) as if it is known parameters. This is called a pseudo-likelihood approach (Carroll et al. 1984, Schafer 1987). If selection into the validation study is independent on the observed values of $\boldsymbol{S}$ and $\boldsymbol{X}$, the pseudo-likelihood approach is valid, but not guaranteed to be efficient. Prescott and Garthwaite (2002) presented a two-stage Bayesian method for analyzing case-control studies when binary outcomes are subject to measurement error. In the first stage, analysis of the data from the validation study yields in prior information for the second stage.

We need to consider what if there is no validation study. Previous studies can provide the information about the misclassification with standard error.On the other hand, replication of the observed variables can be used for the misclassification probabilities. For example, if the misclassification probability is the same for both values of $\boldsymbol{Y}$, then two independent replicates of $\boldsymbol{S}$ a subject suffice to identify the probability.

2.2.3 Semiparametric Methods for Validation Data

Semiparametric analysis by allowing a nonparametric specification of the error model is an alternative to the likelihood method with a drawback, which is sensitive to the assumption about the distribution for the error-prone response (Carroll et al. 1984, Pepe et al. 1989).

Similar to the approaches for the mismeasured covariate with validation data by Carroll and Wand (1991) and Pepe and Fleming (1991), Pepe (1992) proposed a pseudo-likelihood method by assuming that the selection into the second stage validation study is by simple random sampling. For the likelihood in (2.25), the validation data is used to nonparametrically estimate $f(\boldsymbol{S} \mid \boldsymbol{Y}, \boldsymbol{X})$ by kernel regression methods. Then the estimator $\widehat{f}(\boldsymbol{S} \mid \boldsymbol{Y}, \boldsymbol{X})$ is substituted into the likelihood in (2.25). Eventually, the

pseudo-likelihood to maximize is

$$\prod_{i=1}^{n} f(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \boldsymbol{B})^{\triangle_i} \widehat{f}(\boldsymbol{S}_i \mid \boldsymbol{X}_i, \boldsymbol{B}, \boldsymbol{\gamma})^{1-\triangle_i}, \tag{2.28}$$

where $\widehat{f}(\boldsymbol{S}_i \mid \boldsymbol{X}_i, \boldsymbol{B}, \boldsymbol{\gamma}) = \int f(\boldsymbol{Y}_i = y \mid \boldsymbol{X}_i, \boldsymbol{B}) \widehat{f}(\boldsymbol{S}_i \mid \boldsymbol{Y}_i = y, \boldsymbol{X}_i, \boldsymbol{B}, \boldsymbol{\gamma}) dy$. In this approach, estimating $f(\boldsymbol{S}_i \mid \boldsymbol{Y}_i, \boldsymbol{X}_i, \boldsymbol{B}, \boldsymbol{\gamma})$ is challenging because the number of conditional distribution is proportional to the number of all possible combinations of both levels of $\boldsymbol{Y}$ and $\boldsymbol{X}$. Moreover when $\boldsymbol{S}$ is continuous, it is more complicated. In practice, numerical performance of this approach in finite sample sizes needs to be studied further.

2.3 Weighted Estimating Equations Accounting for MAR Data

Missing data is a crucial problem arising in longitudinal and observational studies. A simple approach to missing data is a complete case analysis, that is, analyzing only subjects with complete observations. However, it is well know that the complete case analysis can be biased when the data are not missing complete at random (MCAR). Another ad hoc method for missing covariate data is to exclude the corresponding covariates from the analysis. However, this can result in model misspecification (Ibrahim et al. 2005). There are several approaches of dealing with missing data problem: maximum likelihood (ML), multiple imputation (MI), fully Bayesian (FB), and weighted estimating equations (WEEs). In contrast to ML, MI, and FB methods for missing data, WEEs-based estimates are robust because WEEs require no distributional assumption. In this section, we review mainly weighted estimating equations in a regression setting. Without loss of generality, we assume that response variables are always observed. However, the four methods can be extended to the case that there is missing data in both responses and covariates by minor adjustments.

Horvitz and Thompson (1952) proposed a method for survey data analysis accounting for different proportions of observations within strata by using inverse probability weights (IPW), which are the inverse of the inclusion probability in sampling data analysis, and then the method can be applied to the missing data problem. Motivated from the Horvitz-Thomson estimator, Rotnitzky and Robins (1995), Robins and Rotnitzky (1995), and Robins et al. (1994; 1995) developed a class of estimating equations based on inverse probability weights in a regression setting when data are missing at random (MAR), namely, missingness depends on only observed data rather unobserved data.

Following Ibrahim et al. (2005), denote the mean model by $\mu_i = \mu(\boldsymbol{X}_i, \boldsymbol{\beta}) = E(y_i \mid \boldsymbol{X}_i, \boldsymbol{\beta})$, where $\mu_i(\boldsymbol{X}_i; \boldsymbol{\beta})$ is a known twice-differentiable function of $\boldsymbol{\beta}$. For missing data, we define an indicator variable $R_i = 1$ if covariates are fully observed, $R_i = 0$ otherwise. The distribution of $R_i \mid (Y_i, \boldsymbol{X}_i)$ is Bernoulli with probability $\pi_i(\boldsymbol{\alpha}) = Pr(R_i = 1 \mid Y_i, \boldsymbol{X}_i, \boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ denotes unknown parameters.

For now, let us assume that π_i is known. Robins et al. (1994) proposed weighted quasi-likelihood estimating equations in the complete case:

$$\boldsymbol{u}_{WEE}(\boldsymbol{\beta}) = \sum_{i=1}^{n} R_i \pi_i^{-1} \boldsymbol{d}_i v_i^{-1} (y_i - \mu_i), \tag{2.29}$$

where $\boldsymbol{d}_i = \partial \mu_i / \partial \boldsymbol{\beta}$ and $v_i = v_i(\boldsymbol{\beta}) = var(y_i \mid \boldsymbol{X}_i)$. Although only subjects with complete data contribute to the equation in (2.29), the weighting equations in (2.29) provide a consistent estimate of $\boldsymbol{\beta}$ because

$$\begin{aligned} & E\left\{R_i pi_i^{-1} \boldsymbol{d}_i v_i^{-1} (y_i - \mu_i)\right\} \\ = \; & E_{\boldsymbol{X}_i}\left[E_{y_i|\boldsymbol{X}_i}\{\boldsymbol{d}_i v_i^{-1}(y_i - \mu_i)\}\{E_{R_i|y_i,\boldsymbol{X}_i}(R_i \pi_i^{-1})\}\right] \\ = \; & E_{\boldsymbol{X}_i}\left[E_{y_i|\boldsymbol{X}_i}\{\boldsymbol{d}_i v_i^{-1}(y_i - \mu_i)\}\right] = E_{\boldsymbol{X}_i}(\boldsymbol{0}) = \boldsymbol{0}. \end{aligned} \tag{2.30}$$

The key point of the derivation in (2.30) is $E(R_i \mid y_i, \boldsymbol{X}_i) = \pi_i$. As a consequence, if π_i in

the weighted equations in (2.29) is known or replaced by a consistent estimate, then the weighted estimating equation leads to an unbiased estimator for $\boldsymbol{\beta}$. Zhao et al. (1996) presented a consistent and robust sandwich variance estimator for $\widehat{\boldsymbol{\beta}}$. To estimate the probability of being observed, we can apply a logistic regression model under the MAR assumption. In practice, to apply an ordinary logistic model, we need to assume that π_i depends only on y_i and $\boldsymbol{X}_{all,i}$, where $\boldsymbol{X}_{all,i}$ are the variables observed on all subjects. However, this is a stronger assumption than MAR. Ideally, when covariates are fully observed, the weighted quasi-likelihood estimating equations are most useful.

Robins et al. (1995) proposed semiparametric regression models for longitudinal outcomes with MAR data. Their approach can be regarded as an extension of generalized estimating equations. Under the assumption of monotone missing pattern, a marginal structure model with past history is used to estimate the probability for being observed. Robins et al. (1995) also considered extension to arbitrary missing data patterns.

In fact, the estimates from the weighted estimating equations described above are inefficient because it uses only the information in the complete cases. To improve efficiency, Robins and Rotnitzky (1995) extended (2.30) by using the incomplete cases to estimate the weights.

$$\boldsymbol{u}_{WEE2}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ R_i \pi_i^{-1} \boldsymbol{d}_i v_i^{-1} (y_i - \mu_i) + (1 - R_i \pi_i^{-1}) \boldsymbol{q}(y_i, \boldsymbol{X}_{obs,i}; \boldsymbol{\beta}, \boldsymbol{\alpha}) \right\}, \tag{2.31}$$

where $\boldsymbol{q}(y_i, \boldsymbol{X}_{obs,i}; \boldsymbol{\beta}, \boldsymbol{\alpha})$ is a specified function of the observed data $(y_i, \boldsymbol{X}_{obs,i})$, $\boldsymbol{\beta}$ is the parameters of interest, and $\boldsymbol{\alpha}$ is the parameter related to π_i. IF π_i is correctly specified, then $E_{R_i|y_i,\boldsymbol{X}_i}(1 - R_i\pi_i^1) = 0$. Expectation of the second term in (2.31) will have 0 regardless of the function $\boldsymbol{q}(y_i, \boldsymbol{X}_{obs,i}; \boldsymbol{\beta}, \boldsymbol{\alpha})$. Moreover, if π_i is correctly specified, the first term in (2.31) will have zero expectation. Hence, the estimates of $\widehat{\boldsymbol{\beta}}_{WEE2}$ obtained from the equations in (2.31) will be consistent.

Rotnitzky and Robins (1995) showed the optimal function for $\boldsymbol{q}$ to minimize the asymptotic variance of $\widehat{\boldsymbol{\beta}}_{WEE2}$ by

$$\boldsymbol{q}(y_i, \boldsymbol{X}_{obs,i}; \boldsymbol{\beta}, \boldsymbol{\alpha}) = E(\boldsymbol{u}_i(\boldsymbol{\beta}) \mid y_i, \boldsymbol{X}_{obs}; \boldsymbol{\beta}, \boldsymbol{\alpha}), \tag{2.32}$$

where $\boldsymbol{u}_i(\boldsymbol{\beta}) = \boldsymbol{d}_i v_i^{-1}(y_i - \mu_i)$. To calculate the optimal function in (2.32), we need to know the distribution of the covariates, $p(\boldsymbol{X}_{missing,i} \mid \boldsymbol{X}_{obs,i}; \boldsymbol{\beta}, \boldsymbol{\alpha})$. Thus we need another set of estimating equations to estimate $\widehat{\boldsymbol{\alpha}}$.

Lipsitz et al. (1999) proposed a WEEs with $\boldsymbol{\gamma} = (\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\phi})$ as following:

$$\boldsymbol{S}(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \begin{bmatrix} R_i \pi_i^{-1} \boldsymbol{u}_i(\boldsymbol{\beta}) + (1 - R_i \pi_i^{-1}) E_{\boldsymbol{X}_{missing} | y_i, \boldsymbol{X}_{obs}} \{\boldsymbol{u}_i(\boldsymbol{\beta})\} \\ R_i \pi_i^{-1} \boldsymbol{s}_i(\boldsymbol{\alpha}) + (1 - R_i \pi_i^{-1}) E_{\boldsymbol{X}_{missing} | y_i, \boldsymbol{X}_{obs}} \{\boldsymbol{s}_i(\boldsymbol{\alpha})\} \\ (y_i, \boldsymbol{X}_i^T)^T (R_i - \pi_i) \end{bmatrix}, \tag{2.33}$$

where $\boldsymbol{u}_i(\boldsymbol{\beta}) = \boldsymbol{u}(\boldsymbol{\beta}; y_i, \boldsymbol{X}_i)$ and $\boldsymbol{s}_i(\boldsymbol{\alpha}) = \boldsymbol{s}(\boldsymbol{\alpha}; \boldsymbol{X}_i)$. To solve $\boldsymbol{S}(\widehat{\boldsymbol{\gamma}}) = 0$, the weighted EM-algorithm or Monte Carlo EM algorithm is used when missing covariate is categorical or continuous, respectively (Lipsitz et al. 1999). Robbins and Ritov (1997) showed the estimator from (2.33) is doubly robust, that is, it remains consistent when either a model for the missingness mechanism or the score vector for the missing data given the observed data is correctly specified. However, these are asymptotic properties, and the estimating equations in (2.33) does not extend easily when missing data pattern is non-monotone.

CHAPTER3: THRESHOLD-DEPENDENT PROPORTIONAL HAZARDS MODEL FOR ANALYZING TIME-TO-EVENT DEFINED BY BIOMARKER WITH SUBJECT TO MEASUREMENT ERROR

3.1 Introduction

Type 2 diabetes mellitus (hereafter referred to simply as diabetes) is an adult-onset metabolic disorder and is one of the leading causes of morbidity and mortality (Kumar et al. 2005, pp. 1194–1195). The incidence of diabetes has been increasing over several decades, and many studies have been conducted in various communities to investigate the pathogenesis of diabesity, with the eventual goal being to control or prevent diabetes (Duncan et al. 2003; UK Prospective Diabetes Study Group 1998; Isomaa et al. 2001). During 1987-1989, the Atherosclerosis Risk in Communities (ARIC) Study recruited a population-based cohort from four U.S. communities, Forsyth County, NC, Jackson, MS, suburbs of Minneapolis, MN, and Washington County, MD. Participants underwent a baseline examination in 1987-1989, three follow-up examinations at approximately three-year intervals, and a further examination in 2011-2013. The ARIC Study was designed to investigate the causes of atherosclerosis, but various ancillary studies have investigated several other diseases, including diabetes. The standard ARIC definition of diabetes is a fasting plasma glucose (FPG) $\geq 126 mg/dL$, non-fasting glucose $\geq 200 mg/dL$, a self-reported physician diagnosis of diabetes, or use of diabetes medication in the two weeks preceding the study visit.

The diabetes data from the ARIC Study pose three challenges for analysis. For incident diabetes determined by the FPG value, because of the relatively long intervals

between visits, the exact date of crossing the specified threshold, and hence the exact incident date was unobservable. What is known is the date of the visits at which an individual's FPG values were below or above the threshold. This can be regarded as measurement error of the event time or as interval censoring. Using as event time the visit time at which a value above the threshold is first recorded may lead to invalid inferences (Lindsey and Ryan 1998).

Secondly, the threshold of FPG (126 mg/dL) used as the diagnostic criterion for diabetes is based on the World Health Organization (WHO) guidelines updated in 2005. According to the guidelines, two sets of information have influenced determination of diagnostic cutpoints for diabetes: plasma glucose levels associated with micro-vascular (particularly retinopathy) and cardiovascular complications, and the population distribution of plasma glucose. The Expert Committee on the Diagnosis & Classification of Diabetes Mellitus (2003) reported history of changing the diagnostic threshold for the FPG over time and some countries. Miyazaki et al. (2004) suggested that the threshold for diagnostic fasting plasma glucose level based on prevalence of retinopathy in a Japanese population is lower than that of the current diagnostic criteria. This implies the criterion proposed by WHO has limitations because of the data from which the diagnostic criterion for diabetes was derived. Not only FPG but other biomarkers have different distributions across populations (Vasan 2006; Rule et al. 2004). In analyzing the data from the ARIC Study, we found that factors significantly associated with time to diabetes onset varied with the criteria used to define diabetes. This motivated us to investigate methods for relaxing the requirement of using a specified, fixed threshold.

Finally, there is marked variability (both pre-analytical and analytical) involved in glucose testing, and the pre-analytical variation results from intra-individual and inter-individual variation, whereas analytical variation results from methodology used in measurement of glucose (Schwartz, Reichberg, and Gambino 2005; Schrot, Patel, and

Foulis 2007;Tonyushkina and Nichols 2009; Hellman 2012). Generally, we are unable to distinguish measurement error and individual variability. The pre-analytical variation can be inferred by blood glucose values measured repeatedly over time. The National Institute of Standards and Technology maintains the glucose sample materials that are the gold standard by which instrument manufacturers determine the accuracy of their glucose measurement devices. Variation due to measurement error or individual variability complicates defining time to diabetes occurrence. Moreover, if measurement error is non-ignorable but ignored in the analysis, the analysis may yield an inaccurate conclusion.

In the literature, there exist different methods to address each of the three issues described above. For interval censoring in data with only one follow-up after baseline, Huang (1996) provided a thorough study based on a proportional hazards model and Rossini and Tsiatis (1996), Shen (2000), Ma and Kosorok (2005), and Xue, Lam, and Li (2004) developed several semi-parametric models for interval censored data based on events determined by a fixed threshold. Measurement error in categorical response is called misclassification, and various approaches to account for misclassification have been developed (Hausmana, Abrevayab, and Scott-Mortonb 1998;Neuhaus 2002;Paulino, Soares, and Neuhaus 2003). For measurement error in a continuous response, some authors have proposed methods adjusting for measurement error (Carroll 2006). As a likelihood method, Pepe (1992) developed a nonparametric estimator for the conditional probability of a surrogate response given covariates. Huang and Wang (2000) and Tsiatis and Davidian (2001) handled covariates subject to measurement error in the context of time-to-event subject to right censoring. However, none of these approaches can simultaneously handle the challenges as seen in the ARIC Study, including imprecise event time, non-fixed threshold and measurement error.

In this paper, we propose a novel semiparametric regression model for modelling

the FPG values. Our model is based on an extension of the generalized extreme value distribution. Interestingly, the proposed model is equivalent to modelling threshold-dependent time to diabetes via a Cox proportional hazards model, where the event time is defined as the FPG value crossing the given threshold. To account for measurement error, we develop nonparametric maximum likelihood estimation for inference. The paper is structured as follows. We describe the ARIC Study in Section 3.2. We then propose our method and inference procedure in Section 3.3. Asymptotic Results and the technical details are summarized in Section 3.4 and Application A, respectively. Simulation study and application to the ARIC Study are in Sections 3.5 and $sec : application$, respectively. We give some conclusions in Section $sec : discussion$.

3.2 The ARIC Study

The ARIC Study recruited a population-based cohort of 15,792 (Duncan et al. 2003). The study participants were predominantly white or African-American, and they underwent a baseline examination in 1987-1989, three follow-up examinations at approximately three-year intervals, and a further examination in 2011-2013. We excluded 2,018 participants with prevalent diabetes, 95 participants who were neither white nor African-American and African-Americans in the Minnesota and Washington County cohorts because of small numbers, 853 not returning to any follow-up visit, 26 having no valid diabetes determination at follow-ups, 7 with restrictions on stored plasma use, 12 with missing baseline anthropometrics, and 1,011 with missing FPG values or baseline characteristics. To study the association between baseline risk factors and time to diabetes, although we have the FPG values from multiple follow-up visits, we use only the value at the first follow-up visit because once diagnosed as diabetic by non-study physicians, participants may have started taking medication for diabetes or have adjusted their life style and their FPG levels at the subsequent visits may have

been influenced by these changes. For the same reason, we excluded 128 subjects (1%) diagnosed with diabetes by their own physicians between baseline and the first follow-up visit. We summarize the demographics and baseline characteristics of the 11,642 participants to be included in the analysis in Table 1.

If we ignore the complicated issues arising in the data and simplify the problem by focusing on the binary outcome of presence or absence of diabetes, we can apply a logistic regression model. Defining presence of diabetes as an observed FPG value $\geq 126 mg/dL$, the logistic regression model for the probability of diabetes (refer to the supplemental material) leads to a result that differs from the well-known facts: African-Americans and people with a parental history of diabetes have significantly lower risk of diabetes than whites and people without a parental history of diabetes, respectively. This incorrect conclusion drove us to consider another approach.

We examined the distributions of the FPG values at the follow-up visits as shown in Figure 3.1. Clearly, the distribution of FPG values is very skewed and has a long right tail. Even after logarithm transformation, the distribution remains skewed. We fitted a parametric generalized extreme value distribution (the dashed curve in Figure 3.1), and this suggests that it is a good approximation to the distribution of the FPG values. There appears to be some mismatch at the mode of the FPG distribution, but we believe this is primarily due to measurement errors in the FPG values. This empirical observation motivated us to propose a semiparametric regression model based on the generalized extreme value distribution as described in the methods section.

3.3 Method

3.3.1 Model

For subject i, let $\boldsymbol{X}_i$ be time-invariant covariates such as demographic characteristics and risk factors at baseline such as race, sex, hypertension, parents diabetes history,

age, body mass index, fasting plasma glucose value, high-density lipoprotein, and total cholesterol, and $Y_i^*(t)$ and $Y_i(t)$ be the true FPG value and observed FPG value at time t, respectively. Observed visit time is denoted by V, which can be fixed or random and is assumed to be independent of $Y_i^*(t)$ given $\boldsymbol{X}_i$. Thus, the observed data from n independently and identically distributed subjects are $\{Y_i(V_i), V_i, \boldsymbol{X}_i \mid i = 1, \ldots, n\}$, where V_i is the visit time for subject i. The observed data is denoted by $\{\boldsymbol{W}_i \mid i = 1, \ldots, n\}$ hereafter.

Motivated by the empirical observation in Section 2, we assume that the true FPG values follow one type of generalized extreme value distribution: $\exp\{-\alpha \exp(-\mu y^* + \gamma)\}$, for parameters $\alpha > 0, \mu > 0$, and $-\infty < \gamma, y^* < \infty$. The underlying trend of the true biomarker values is never observable because of intra-individual variability and measurement error; however, it is reasonable to assume that in a population of middle-aged and older adults the underlying trend of $Y_i^*(t)$ is non-decreasing over time t within the follow-up period because chronic diseases such as diabetes, hypertension, and asthma are irreversible without medication or lifestyle changes. To incorporate baseline covariates and account for the time-dependent nature of the FPG values, our proposed semiparametric regression model is

$$P(Y_i^*(t) \le y^* \mid \boldsymbol{X}_i) = \exp\{-\Lambda_0(t) \exp(-\mu y^* + \boldsymbol{\beta}^T \boldsymbol{X}_i)\}, \tag{3.1}$$

where $\Lambda_0(t)$ is non-decreasing over time and positive when $t > 0$, and both μ and $\boldsymbol{\beta}$ are unknown parameters. In the absence of covariates, this model can be regarded as a stochastic process with the mean function $\mu^{-1}(\log \Lambda_0(t) + \gamma_0)$, where γ_0 is the Euler-Mascheroni constant.

Interestingly, the above model is equivalent to modeling the threshold-dependent time-to-diabetes events. Specifically, for any given threshold value ξ, we define $T_{i\xi}$ to be the first time that $Y_i(t)$ crosses the threshold ξ. Then under the assumption that

$Y_i^*(t)$ is non-decreasing, we have $P(T_{i\xi} > t \mid \boldsymbol{X}_i) = P(Y_i^*(t) \leq \xi \mid \boldsymbol{X}_i)$. The equation in (3.1) is equivalent to

$$P(T_{i\xi} > t \mid \boldsymbol{X}_i) = \exp\{-\Lambda_0(t)\exp(-\mu\xi + \boldsymbol{\beta}^T\boldsymbol{X}_i)\}.$$

That is, we obtain a proportional hazard model with a threshold-dependent baseline hazard function for $T_{i\xi}$ as

$$\lambda_i(t) = \exp(-\mu\xi)\lambda_0(t)\exp(\boldsymbol{\beta}^T\boldsymbol{X}_i), \tag{3.2}$$

where $\lambda_0(t) = d\Lambda_0(t)/dt$. Equivalently, $\log\Lambda_0(T(\xi))=-\boldsymbol{\beta}^T\boldsymbol{X}_i + \mu\xi + \epsilon$, where ϵ is independent of $\boldsymbol{X}_i$ and has the extreme value distribution. This new expression gives a nice interpretation of the parameters μ and $\boldsymbol{\beta}$: $\mu > 0$ is essentially the effect of using different thresholds for the threshold-dependent time to diabetes. Clearly, the larger the threshold, the longer the time to diabetes. The regression parameter $\boldsymbol{\beta}$ in the model (3.2) gives the log-hazard ratio of $\boldsymbol{X}$ on time to diabetes occurrence after controlling for any given threshold value. Therefore, $\boldsymbol{\beta}$ being positive implies that greater risk of developing diabetes is associated with larger values of $\boldsymbol{X}$.

Our second model considers the effect of measurement error using the classical additive measurement error model (Carroll 2006; Fuller 1987; Tsiatis, DeGruttola, and Wulfsohn 1995)

$$Y_i(t) = Y_i^*(t) + \epsilon_i(t),\ i = 1, \ldots, n. \tag{3.3}$$

We assume the measurement error $\epsilon_i(t)$ has a normal distribution with mean zero and variance σ^2 for any time t and is independent of $Y_i^*(t)$, $\boldsymbol{X}_i$, and ξ. The measurement error variance of σ^2 may be estimated in practice by taking repeated measurements (Schwartz et al. 2005;Tonyushkina and Nichols 2009). Since information about the measurement error variation can be obtained from outside the dataset being analyzed,

we consider σ^2 to be known. The measurement error model can be regarded as a mixture model to give flexibility to the distribution for the observed FPG values.

Under the above two models, we construct the likelihood for the observed biomarker $Y_i(V_i)$ given $(V_i, \boldsymbol{X}_i), i = 1, .., n$:

$$\prod_{i=1}^{n} \int_{-\infty}^{\infty} \exp\{-\Lambda_0(V_i) e^{\boldsymbol{\beta}^T \boldsymbol{X}_i - \mu\xi}\} \Lambda_0(V_i) \mu \exp(\boldsymbol{\beta}^T \boldsymbol{X}_i - \mu\xi) \frac{1}{\sigma} \phi \left\{ \frac{Y_i(V_i) - \xi}{\sigma} \right\} d\xi, \qquad (3.4)$$

where $\phi(\cdot)$ is the standard normal density function.

3.3.2 Inference Procedure

We maximize (3.4) to estimate all the parameters, including $\theta = (\mu, \beta^T)^T$ and Λ_0. Specifically, we estimate Λ_0 as a step function, with jumps at the observed V_i's. Let $v_{(1)} < \ldots < v_{(K)}$ be ordered observed times of $\{v_i \mid i = 1, \ldots, n\}$ and $\Lambda_k = \Lambda_0(v_{(k)})$ and $v_{(0)} = 0$. Then we maximize (3.4) over θ and Λ_k's subject to constraints $0 \leq \Lambda_1 \leq \ldots \leq \Lambda_K$.

To facilitate the maximization, we introduce a latent threshold variable ξ_i for each subject. Furthermore, the random variables of $(Y_i(V_i), V_i, \boldsymbol{X}_i, \xi_i)$ follow a joint distribution given by

$$\exp\{-\Lambda_0(V_i) e^{\boldsymbol{\beta}^T \boldsymbol{X}_i - \mu\xi_i}\} \Lambda_0(V_i) \mu \exp(\boldsymbol{\beta}^T \boldsymbol{X}_i - \mu\xi_i) \frac{1}{\sigma} \phi \left\{ \frac{Y_i(V_i) - \xi_i}{\sigma} \right\}.$$

Then the likelihood (3.4) is the observed likelihood function with $\xi_i, i = 1, ..., n$ being missing. Therefore, we adopt the expectation-maximization (EM) algorithm. Then the

complete log-likelihood function is

$$l_c(\theta) = \sum_{k=1}^{K}\sum_{i=1}^{n} I(V_i = V_{(k)})\Big[-\Lambda_k \exp(\boldsymbol{\beta}^T \boldsymbol{X}_i - \mu\xi_i) + \log\Lambda_k + \log\mu + \boldsymbol{\beta}^T \boldsymbol{X}_i - \mu\xi_i - \frac{1}{2}\log\sigma^2 - \frac{\{Y_i(V_i) - \xi_i\}^2}{2\sigma^2}\Big].$$

In the M-step at the lth iteration of the EM algorithm, we first maximize the conditional expectation of the complete log-likelihood function given observed data over Λ_k's. We then update θ via the Newton-Raphson algorithm. Specifically, we maximize $Q(\Lambda)$ defined by

$$Q(\Lambda) = \sum_{k=1}^{K}\sum_{i=1}^{n} I(V_i = V_{(k)}) E\{-\Lambda_k \exp(\boldsymbol{\beta}^T \boldsymbol{X}_i - \mu\xi_i) + \log\Lambda_k \mid \boldsymbol{W}_i, \theta^{(l)}\}. \tag{3.5}$$

Since $Q(\Lambda)$ is a concave function over a convex cone satisfying $\Lambda_1 \le \ldots \le \Lambda_K$, this maximization can be carried out using one of the many existing algorithms for convex optimization. To update θ, we apply the following one-step Newton-Raphson algorithm,

$$\theta^{(l+1)} = \theta^{(l)} + E(-\partial^2 l_c/(\partial\theta)^2 \mid \boldsymbol{W}, \theta^{(l)})^{-1}_{\theta=\theta^{(l)}} E(\partial l_c/\partial\theta \mid \boldsymbol{W}, \theta^{(l)})_{\theta=\theta^{(l)}}. \tag{3.6}$$

The conditional expectations in (3.6) are calculated in the E-step of the EM algorithm based on the following expression,

$$E(g(\xi_i) \mid W_i, \theta^{(l)}) = \frac{I(V_i = V_{(k)}) \int_{-\infty}^{\infty} g(\xi_i) \exp(-\Lambda_k e^{\boldsymbol{\beta}^T \boldsymbol{X}_i - \mu\xi_i}) e^{-\mu\xi_i} \phi\big(\frac{Y_i(V_i) - \xi_i}{\sigma}\big) d\xi_i}{\int_{-\infty}^{\infty} \exp(-\Lambda_k e^{\boldsymbol{\beta}^T \boldsymbol{X}_i - \mu\xi_i}) e^{-\mu\xi_i} \phi\big(\frac{Y_i(V_i) - \xi_i}{\sigma}\big) d\xi_i},$$

where the $g(\xi)$'s to be calculated are ξ, ξ^2, $e^{-\mu\xi}$, $e^{-\mu\xi}\xi$, and $e^{-\mu\xi}\xi^2$. This integration can be approximated by the Gauss-Hermite quadrature (Davis 1984, pp. 190), so it can be

approximated by

$$\sum_{k=1}^{N}\left(\sqrt{2}\sigma w_k g\{\sqrt{2}\sigma w_k + Y_i(V_i)\}\exp\left[-\Lambda_0(V_{ij})e^{\boldsymbol{\beta}^T X_i - \mu\{\sqrt{2}\sigma z_k + Y_i(V_i)\}}\right]e^{-\mu\{\sqrt{2}\sigma z_k + Y_i(V_i)\}}\right), \tag{3.7}$$

where N is the number of the quadratures and ω_k and z_k are weights and abscissae for the Gauss-Hermite quadrature, respectively. This loop of the E-step and the M-step is repeated until $|\theta^{(l+1)} - \theta^{(l)}|$ is smaller than a pre-specified criterion. We denote the final estimators as $\widehat{\theta}^T = (\widehat{\mu}, \widehat{\boldsymbol{\beta}}^T)$ and $\widehat{\Lambda}$.

3.3.3 Variance Estimation

In the asymptotic results given in the Appendix and the supplemental material, we show that the proposed estimator for $\boldsymbol{\theta}_0$ is semiparametrically efficient. Moreover, the efficient score function for θ at $\theta = \theta_0$ is

$$l_\theta^*(\theta_0, \Lambda_0, \boldsymbol{W}) = \begin{pmatrix} \mu_0^{-1} - E(\kappa\xi \mid \boldsymbol{W}) - E(\kappa \mid \boldsymbol{W})R_1(V) \\ E(\kappa \mid \boldsymbol{W})\{\boldsymbol{X} - R_2(V)\} \end{pmatrix}, \tag{3.8}$$

where $\kappa{=}1 - \Lambda_0(V)\exp(\beta_0^T\boldsymbol{X} - \mu_0\xi)$, and

$$R_1(V) = E\big[E(\kappa \mid \boldsymbol{W})\left\{\mu_0^{-1} - E(\kappa\xi \mid \boldsymbol{W})\right\} \mid V\big] \big/ E\big\{E(\kappa \mid \boldsymbol{W})^2 \mid V\big\},$$
$$R_2(V) = E\left\{\boldsymbol{X}E(\kappa \mid \boldsymbol{W})^2 \mid V\right\} \big/ E\left\{E(\kappa \mid \boldsymbol{W})^2 \mid V\right\}.$$

Therefore, the asymptotic variance of $n^{1/2}\widehat{\theta}$ is the inverse of the information for θ_0, that is, $I(\theta_0) = E(l_\theta^{*\otimes 2})$, where $a^{\otimes 2} = aa^T$ for any vector a.

For the asymptotic variance of $n^{1/2}\widehat{\theta}$, we estimate $I(\theta_0)$ by $n^{-1}\sum_{i=1}^n \widehat{l}_{\theta i}^{*\otimes 2}$, where

$$\widehat{l}_{\theta i}^* = \begin{pmatrix} \widehat{\mu}^{-1} - \widehat{E}(\kappa\xi \mid \boldsymbol{W}_i) - \widehat{E}(\kappa \mid \boldsymbol{W}_i)\widehat{R}_1(V_i) \\ \widehat{E}(\kappa \mid \boldsymbol{W}_i)\big\{\boldsymbol{X}_i - \widehat{R}_2(V_i)\big\} \end{pmatrix}, \tag{3.9}$$

and $\widehat{E}(\kappa \mid \boldsymbol{W}), \widehat{E}(\kappa\xi \mid \boldsymbol{W}), \widehat{R}_1(V_i)$, and $\widehat{R}_2(V_i)$ are some consistent estimators for $E(\kappa \mid \boldsymbol{W})$, $E(\kappa\xi \mid \boldsymbol{W})$, $R_1(V_i)$, and $R_2(V_i)$, respectively. Specifically, $\widehat{E}(\kappa \mid \boldsymbol{W})$ and $\widehat{E}(\kappa\xi \mid \boldsymbol{W})$ are

$$\widehat{E}(\kappa \mid \boldsymbol{W}) = 1 - \widehat{\Lambda}(V)\exp(\widehat{\beta}^T\boldsymbol{X})\widehat{E}(\exp(-\widehat{\mu}\xi) \mid \boldsymbol{W}),$$
$$\widehat{E}(\kappa\xi \mid \boldsymbol{W}) = \widehat{E}(\xi \mid \boldsymbol{W}) - \widehat{\Lambda}(V)\exp(\widehat{\beta}^T\boldsymbol{X})\widehat{E}(\exp(-\widehat{\mu}\xi)\xi \mid \boldsymbol{W}),$$

and the other two estimators are some type of kernel estimators with bandwidth h_n:

$$\widehat{R}_1(v) = \frac{\sum_{j=1}^n K_{h_n}(V_j - v)\widehat{E}(\kappa \mid W_j)\{\widehat{\mu}^{-1} - \widehat{E}(\kappa\xi \mid W_j)\}}{\sum_{j=1}^n K_{h_n}(V_j - v)\widehat{E}(\kappa \mid W_j)^2},$$
$$\widehat{R}_2(v) = \frac{\sum_{j=1}^n \boldsymbol{X}_j K_{h_n}(V_j - v)\widehat{E}(\kappa \mid W_j)^2}{\sum_{j=1}^n K_{h_n}(V_j - v)\widehat{E}(\kappa \mid W_j)^2},$$

where $K_{h_n}(x) = h_n^{-1}\exp(-x^2/h_n)$. In the Appendix, we establish the consistency of this variance estimator assuming that $h_n \to 0$ and $nh_n \to \infty$ as $n \to \infty$. We choose $(n/2)^{-1/2}$ for h_n.

When the number of observations is large, as in the ARIC Study, an alternative approach to estimating the variance is via the profile likelihood. Specifically, for each parameter in θ, we fix it in the proposed EM algorithm and at convergence, we compute the log-likelihood function as its profile likelihood function. Then based on the profile likelihood theory, the inverse of the negative curvature of the profile likelihood function should give a consistent estimator for the variance.

3.4 Asymptotic Results

We establish asymptotic properties for the proposed estimators under the following conditions and the proofs are summarized in Appendix A. Let θ_0 and Λ_0 denote the true regression parameter and cumulative hazard function, respectively.

(A1) The finite-dimensional parameter space $\boldsymbol{\Theta}$ is a compact subset of the domain of $\boldsymbol{\theta}$.

(A2) The covariate $\boldsymbol{X}$ has bounded support with probability 1. If $\boldsymbol{\beta}^T\boldsymbol{X} + \alpha = 0$ almost surely (a.s.), then $\boldsymbol{\beta} = 0$ and $\alpha = 0$.

(A3) The support of the observation time, V, is an interval $\mathscr{S}[V] = [l_V, u_V]$, with $0 < l_V \leq u_V < \infty$.

(A4) The cumulative hazard function Λ_0 has strictly positive derivative on $\mathscr{S}[V]$.

The assumptions that parameter, covariate, and observation time are bounded are standard. Condition (A2) ensures the identifiability of $\boldsymbol{\theta}$ and Λ. These conditions hold naturally in most applications.

For convergence of the estimates to the true parameters, we need to define a topology. Let the bounded regression parameter space $\Theta(\subset \mathscr{R}^d)$ be equipped with the Euclidean topology. Regarding infinite dimensional nonparametric space, let $\mathscr{F}$ be the set of all Borel subprobability measures on $\mathscr{S}[V]$. Then $\mathscr{F}$ can be equipped with the vague topology by defining that, for any sequence $F_n \in \mathscr{F}$ and $F \in \mathscr{F}$, F_n converges vaguely to F if and only if

$$\int f dF_n \to \int f dF \quad \text{for every} \quad f \in C_0(\mathscr{S}[V]),$$

where $C_0(\mathscr{S}[V])$ is the set of all continuous functions that vanish outside a compact subset of $\mathscr{S}[V]$. Then the product space $\Theta \times \mathscr{F}$ can be equipped with the product topology of the Euclidean topology and the vague topology. In the product topology, it is said that $(\widehat{\boldsymbol{\theta}}, \widehat{F})$ converges to $(\boldsymbol{\theta}, F)$ when $\widehat{\boldsymbol{\theta}}$ and $\widehat{F}$ converge to $\boldsymbol{\theta}$ and F, respectively.

Theorem 3.4.1. *(Consistency of the MLE) Under conditions (A1)–(A3),* $\widehat{\boldsymbol{\theta}} \to \theta_0$ *almost surely, and if* $v \in \mathscr{S}[V]$ *is a continuity point of* Λ_0, $\widehat{\Lambda}(v) \to \Lambda_0(v)$ *almost surely.*

Moreover, if Λ_0 *is continuous, then* $\sup_{v\in\mathscr{S}[V]} | \widehat{\Lambda}(v) - \Lambda_0(v) | \to 0$ *almost surely.*

Before discussing the overall convergence rate, we define the distance d on $\mathscr{R}^d \times \Phi$ as follows:

$$d\{(\theta_1, \Lambda_1), (\theta_2, \Lambda_2)\} = |\theta_1 - \theta_2| + \|\Lambda_1 - \Lambda_2\|_{2,P_V},$$

where $|\theta_1 - \theta_2|$ is the Euclidean distance in $\mathscr{R}^d$,

$$\|\Lambda_1 - \Lambda_2\|_{2,P_V} = \left[\int \{\Lambda_1(v) - \Lambda_2(v)\}^2 dP_V\right]^{1/2},$$

and P_V is the marginal probability measure of the measurement time variable V.

Our next theorem gives the convergence rates of the estimators in terms of this distance.

Theorem 3.4.2. *(Rate of convergence) Under Conditions (A1)–(A3),*

$$d\{(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda}), (\theta_0, \Lambda_0)\} = O_p(n^{-1/3}).$$

The overall rate of convergence is dominated by $\widehat{\Lambda}$. However, it is shown in the next theorem that the convergence rate of $\widehat{\boldsymbol{\theta}}$ can be refined to achieve a rate of $n^{1/2}$.

Theorem 3.4.3. *(Asymptotic normality and efficiency) Suppose that* θ_0 *is an interior point of* Θ *and that conditions (A1)–(A4) are satisfied. Then*

$$n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = n^{1/2}(\mathbb{P}_n - P) I(\boldsymbol{\theta}_0)^{-1} l^*_{\boldsymbol{\theta}_0}(\boldsymbol{W}) + o_p(1) \to N(0, I(\boldsymbol{\theta}_0)^{-1}) \quad \textit{in distribution,}$$

where $\mathbb{P}_n$ *is the empirical measure of* $\boldsymbol{W}_i$, $i = 1, \ldots, n$, *that is,* $\mathbb{P}_n l^*_{\theta_0}(\boldsymbol{W}) = n^{-1} \sum_{i=1}^n l^*_{\theta_0}(\boldsymbol{W}_i)$, P *is the probability measure, that is,* $P l^*_{\theta_0}(\boldsymbol{W}) = \int l^*_{\theta_0}(\boldsymbol{W}) dP$, $l^*_{\theta_0}(\boldsymbol{W})$ *is the efficient score defined in* (3.8)*, and* $I(\theta_0)$ *is the information.*

Since $\widehat{\boldsymbol{\theta}}$ is asymptotically linear with efficient influence function, and the model (the likelihood function) is sufficiently smooth (Hellinger differentiable) with respect to $(\boldsymbol{\theta}, \Lambda)$, it is asymptotically efficient in the sense that any regular estimator has asymptotic variance matrix no less than that of $\widehat{\boldsymbol{\theta}}$.

Theorem 3.4.4. *(Consistency of information estimator) When the bandwidth h_n satisfies that h_n and $\log n/(nh_n)$ converge to 0 as $n \to \infty$, $\mathbb{P}_n \widehat{l}^{*\otimes 2}_{\boldsymbol{\theta} i}$ converges to $Pl^{*\otimes 2}_{\theta_0}$.*

3.5 Simulation Study

Simulation studies were conducted to assess the performance of the estimators proposed in Section 3. We considered two sets of simulations in which the observation times were either discrete or continuous random variables. For discrete measurement times, the time point for each subject was chosen randomly from $\{0.1, 0.2, 0.4, 0.8\}$, while continuous observation times were generated from the uniform distribution over $[0, 1]$. In each simulation, two covariates were included in the model: one generated from the Bernoulli distribution with probability 0.5, and the other from the normal distribution with mean 0 and variance 0.1. The true values for (β_1, β_2) were set as (0.3, 0.3), and the true cumulative baseline hazard assumed to be $2t^{1/5}$. Consequently, the true FPG value was generated as:

$$Y_i^*(t_i) \quad = \quad \mu^{-1}\Big[\beta^T X_i - \log\big\{-\log(p_i)/\Lambda_0(t_i)\big\}\Big], \tag{3.10}$$

where t_i and p_i were from the uniform distribution over $[0, 1]$. For the observed biomarker values, Y_i's were obtained by adding Y_i^* and ϵ_i, where ϵ_i was independently generated from a normal distribution with mean 0 and variance σ^2=0.25. In the simulations, we used μ=0.5 or 1.0, where the corresponding ratios of measurement error variance to true biomarker variance were 0.04 and 0.16, respectively. We varied sample

sizes from 400 to 800 and conducted 1,000 replicates for each simulation.

For each simulated dataset, we applied the proposed the EM algorithm to estimate the parameters. The initial values for β and $\Lambda_0(t)$ were 0 and observed times, respectively. In the M-step, the spectral projected gradient method was used for constrained optimization in (3.5). The convergence criterion for the EM algorithm was 10^{-6}. In the simulations, we noticed that the threshold effect of μ was sensitive to the initial values. Therefore, we first calculated the profile likelihood μ using the same algorithm except that μ was held at some fixed value; we then carried out a grid search to find the maximizer for μ. The variance estimation was based on the formula in Section3.3.3. For comparison, we also calculated the maximum likelihood estimates assuming that the threshold value was fixed for all the subjects, and there was no measurement error. Every subject may have a different threshold in the simulation scenario; however, we need a fixed threshold value for the ICM method. We set fixed thresholds to be at the 90%, 80%, or 70% quantiles of the true biomarker.

In both scenarios, for the discrete and continuous time points, we observed similar results. We present the results of simulations for continuous time points. Table 3.2 shows that bias of the proposed estimators is small, and it decreases as the sample size increases or the variance ratio decreases; the estimated variances agree well with the empirical variance, and the coverage probability is reasonable. Under the setting that the threshold value varies from person to person, and the FPG value includes measurement error, the empirical standard deviation of our estimators are much smaller than the ICM method, and the bias is smaller than the ICM method when sample size increases.

Table 3.2 shows that for the ICM method bias increases and efficiency improves as the fixed threshold increases; the standard deviation and bias of the ICM method decrease when the sample size increases; and accuracy and efficiency of the ICM method

are influenced more by the choice of the fixed threshold than by the measurement error.

In additional simulation studies, we examined the robustness of our method against misspecification of the measurement error distribution in the case of the random continuous time points. Specifically, we let the true distribution for the measurement error be the log-gamma distribution with mean 0 and variance 0.33 or with mean 0.85 and variance 0.28; but we misspecified it in the model (3.4) as the normal distribution with the same mean and variance as the true distribution. The resulting bias and variance (not shown) were similar to those in Table 3.2. The coverage rates of the estimated regression parameters and the estimated threshold parameters were still around 95%. This shows that the proposed method is not sensitive to the distribution of the measurement errors. Finally, to investigate which of measurement error and varying threshold contributes more to the difference in numerical performance between our estimators and the ICM method, we applied the ICM method to the simulated data excluding the measurement error. Bias and empirical standard deviation are similar in the simulations with and without measurement error. Hence, accuracy and efficiency of the ICM method is influenced more by having a fixed threshold than by measurement error.

3.6 Analysis of the ARIC Study Data

We analyzed the ARIC Study data using the proposed model. As potential risk factors, we considered baseline characteristics of participants such as race, gender, hypertension, parents diabetes history, age, BMI, FPG, HDL cholesterol, and total cholesterol. These variables are regarded as major factors associated with diabetes (Duncan et al. 2003).

Both the College of American Pathologists (CAP) and the Clinical Laboratory Improvement Amendments of 1988 determined the total allowable error, 10% for glucose, and generally laboratories are well within the total allowable error (Schwartz et al.

2005). Hence we chose $\sigma^2 = 0.3^2$, corresponding to 0.09 ratio of measurement error variance to the standardized FPG variance.

In order to facilitate calculations, we standardized FPG values as well as baseline continuous covariates to have mean zero and unit variance. Also, the observation time was scaled to (0,1]. FPG values below 75 mg/dL were winsorized to reduce the influence of outliers in the lower tail of the distribution because our interest is in crossing a threshold towards the upper end of the distribution. The number of subjects having FPG values below 75 mg/dL is 204 (1.7%). In winsorization of FPG at 70 or 65 mg/dL, estimates of parameters for continuous covariates are practically unchanged, whereas estimates for those of discrete covariates changed slightly. However, the statistical significance of the risk factors for diabetes remained unchanged.

Although we proposed the variance estimator in Section3.3.3, we adopted the profile likelihood function as described in Section3.3.3 to estimate the asymptotic variance of the estimators. For each regression parameter, we computed the log-likelihood function as its profile likelihood function in the way described in section 3.3.3. The estimated profile likelihood functions appeared to be unimodal, and we obtained the MLE from the profile likelihood for age, which has the maximum value among all the profile likelihood functions (Table 3.3). Then we numerically calculated the inverse of the negative curvature of the profile likelihood function for each parameter at the MLE.

For comparison, we considered two semiparametric models using the fixed threshold of $126mg/dL$, a naive method and Pan (1999)'s method, both ignoring the measurement error in the observed FPG value; for the naive method, we treated the interval-censored data as right-censored data and applied the Cox proportional hazard model. When the FPG value is above the threshold at the next visit after baseline, the event time is set to be the mid-point between baseline and the next visit. Otherwise, data are regarded as right-censored at the visit after baseline. Pan (1999)'s method modified

the iterative convex minorant algorithm as a generalized gradient projection method, and the algorithm can be implemented using the R-package, INTCOX, developed by Henschel et al. (2007)(referred to as the ICM method hereafter). In the ICM method, we treat the visit times as current status data for time to diabetes. These two methods ignore measurement error in FPG values and use the fixed threshold of 126 mg/dL for the counting processes. Moreover, the naive method improperly accounts for interval censoring. The ICM method does not provide standard errors, so we used the simple bootstrap sampling method with 200 replications to estimate the standard errors of the regression parameters.

The ICM method yielded similar results to the naive method for most risk factors, but the effect size and significance of gender and age are different; from the ICM method, it is found that African-Americans, people with parental history of diabetes, older age, higher BMI and FPG, and lower HDL cholesterol have significantly higher risk of diabetes than people with the opposite characteristics. Our method found additional significant factors for diabetes such as hypertension and higher total cholesterol. The factors associated with diabetes found from the proposed model agree with the generally known factors (Mokdad et al. 2003). Total cholesterol consists of HDL cholesterol, LDL cholesterol, and triglycerides. It is known that higher LDL cholesterol and triglycerides and lower HDL cholesterol increase diabetes risk, and high total cholesterol plays a more critical role as a diabetes risk indicator than low HDL cholesterol. We gain less biased and more precise risk estimates from the proposed model.

In the four US communities, African-Americans, males, people with hypertension, and people with parents diabetes history have 1.20, 1.45, 1.31, and 1.42 times greater hazard of diabetes than whites, females, people with hypotension or normal blood pressure, and people without parental diabetes history, respectively. When baseline BMI, FPG, and total cholesterol increase by 1 unit, and HDL cholesterol decreases by

1 unit, where the unit is on the original measurement scale, then the hazard of diabetes increases by a factor of 0.024, 0.052, 0.001, and 0.002, respectively. To investigate the goodness-of-fit of our model, we used the log-likelihood ratio test to compare the model in (3.4) with the model with no measurement error, and the test based on the mixture chi-square distribution shows that there is significant measurement error ($p < 0.001$). In addition, we generated the predicted glucose values and the empirical marginal distribution for the predicted values using the density formula in (3.10) and the formula in (3.4) based on the estimates, respectively. We compared the histogram of the observed FPG values with the empirical marginal distribution for the model-fit (left in Figure 3.2). The marginal distribution in Figure 3.2 shows better fit to the observed distribution of FPG values than the generalized extreme value distribution in the mode of the distribution. Using the predicted values, we suggest another graphical method for model diagnosis, a residual plot, subtracting the predicted means from the real observed glucose values (right in Figure 3.2). The residual plot of Figure 3.2 shows a fairly good fit and the residuals are randomly scattered around 0.

However, in the residual plot, we observe that there are 52 observations with relatively large residuals. The observations with large residuals are above the 99.7% quantiles of FPG values (inter-quartile ranges: 179–246 mg/dL) and their observation times are relatively early. As a sensitivity analysis, we reanalyzed the data excluding those observations to investigate the influence of these observations on the result. After excluding the observations, the estimated hazard ratios barely changed and significance of the factors remains unchanged.

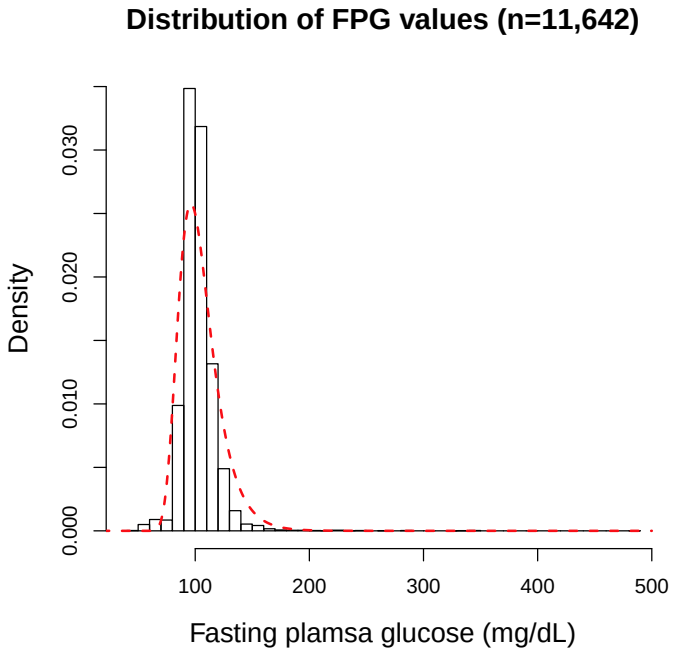
3.7 Concluding Remarks

Motivated by the ARIC Study to find associations between potential risk factors and time to diabetes occurrence determined by FPG and a threshold, we propose a

semiparametric regression model based on the generalized extreme value distribution, which turns out to be equivalent to a class of proportional hazard models for threshold-dependent time to diabetes. We account for measurement error in observed FPG by incorporating the additive measurement error model into the observed likelihood. The application to the ARIC Study reveals significant risk factors which are consistent with clinical findings from this study. Compared to the existing methods, the proposed model yields risk effect estimates in the correct direction and with improved efficiency.

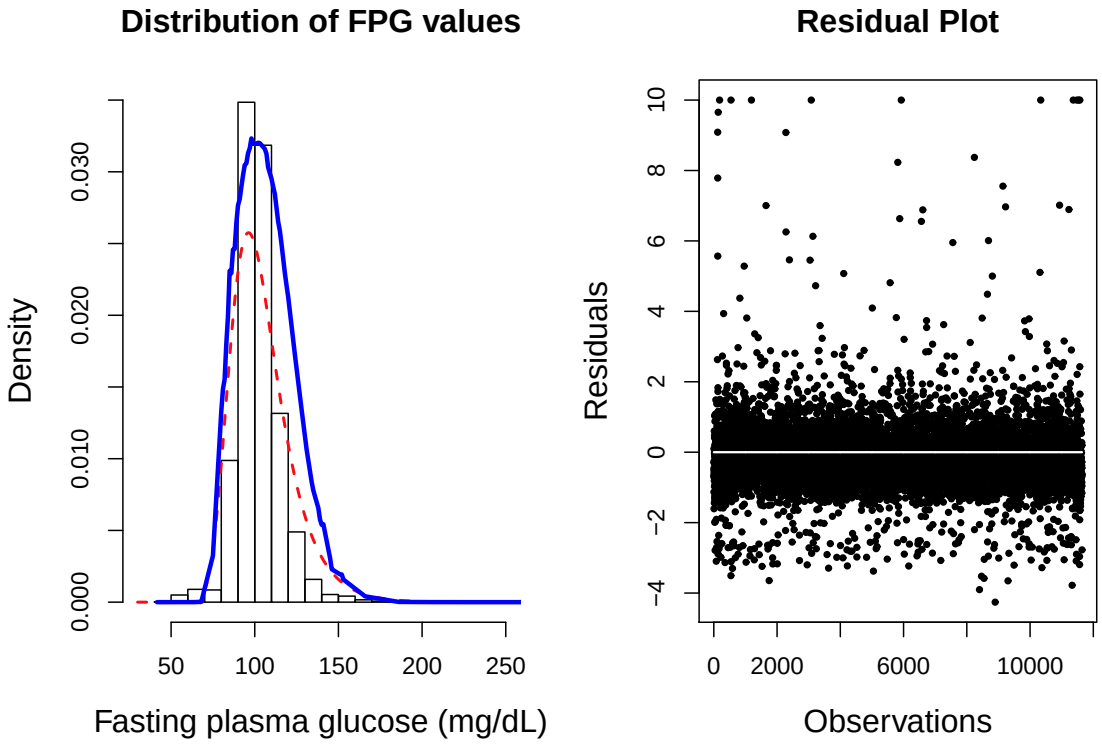
Although we have focused on only one follow-up time per subject, the proposed model can be generalized to repeated observations using pseudo-likelihood ignoring dependence between biomarker values within the same subject. On the other hand, when a covariance structure for the true biomarker values is postulated, semiparametric maximum likelihood methods can be constructed. Furthermore, we can generalize the linear model with respect to threshold to a nonparametric model.

Figure 3.1: Distribution of Fasting Blood Glucose Values



The dashed probability density curve is for the generalized extreme value distribution: $\exp\left(-14\exp(-0.07y+4.1)\right) I(-\infty < y < \infty)$.

Figure 3.2: Quantile-Quantile and Residual Plots



In the left plot, the dashed curve is the generalized extreme value distribution: $\exp\left(-14\exp(-0.07y+4.1)\right) I(-\infty < y < \infty)$. The solid curve is the distribution for the predicted values based on the estimates using our method.

Table 3.1: Baseline Characteristics of the Study Participants

Baseline Factor	n=11,642 N (%) or mean (±SE)
Center	
Forsyth	3,132 (26.9%)
Jackson	2,224 (19.1%)
Minneapolis	3,318 (28.5%)
Washington	2,968 (25.5%)
Race	
White	9,120 (78.3%)
Gender	
Female	6,499 (55.8%)
Parental history of diabetes	
Yes	2,935 (25.2%)
Hypertension	
Yes	3,471 (29.8%)
Age (years)	53.9(±5.7)
BMI (kg/m^2)	27.2 (±5.0)
FPG (mg/dL)	97.7 (±10.4)
HDL (mg/dL)	52.7 (±17.1)
Total cholesterol (mg/dL)	214.1 (±40.9)

SE: standard error

Table 3.2: Simulation Result in the Scenario with Continuous Random Time Points

Sample Size	Variance Ratio	Parameter	True Value	Our method				ICM method					
								90%		80%		70%	
				Bias	SE	SEE	CP	Bias	SE	Bias	SE	Bias	SE
400	0.04	μ	0.5	0.013	0.024	0.023	0.930	-	-	-	-	-	-
		β_1	0.3	0.010	0.104	0.122	0.977	-0.004	0.189	-0.002	0.160	0.000	0.148
		β_2	0.3	0.000	0.174	0.195	0.981	0.004	0.281	0.010	0.260	0.009	0.240
	0.16	μ	1.0	0.028	0.054	0.053	0.924	-	-	-	-	-	-
		β_1	0.3	0.009	0.118	0.135	0.974	-0.005	0.352	-0.004	0.247	-0.015	0.193
		β_2	0.3	0.002	0.213	0.214	0.951	-0.003	0.540	-0.011	0.387	-0.010	0.300
800	0.04	μ	0.5	0.007	0.016	0.016	0.940	-	-	-	-	-	-
		β_1	0.3	0.001	0.081	0.084	0.955	0.002	0.124	-0.002	0.112	-0.004	0.102
		β_2	0.3	0.002	0.121	0.132	0.966	-0.006	0.201	-0.009	0.183	-0.008	0.167
	0.16	μ	1	0.014	0.040	0.036	0.920	-	-	-	-	-	-
		β_1	0.3	0.001	0.086	0.092	0.960	-0.001	0.235	-0.005	0.168	-0.009	0.132
		β_2	0.3	0.000	0.148	0.146	0.941	-0.031	0.377	-0.016	0.274	-0.015	0.212

Variance ratio=true biomarker value variance vs. measurement error variance,
SE=empirical standard error,
SEE=standard error estimate,
CP=coverage probability rate,
β_1 is for the discrete variable,
β_2 is for the continuous variable.

Table 3.3: Analysis of Time to Diabetes Occurrence from the ARIC Study Data

	Naive method			ICM method			Our method		
	Estimate	SE.	p-value	Estimate	SE.	p-value	Estimate	SE.	p-value
Threshold Effect	–	–	–	–	–	–	1.633	0.010	<0.0001
Race									
African-Americans	0.707	0.095	<0.0001	0.993	0.132	<0.0001	0.182	0.022	<0.0001
Gender									
Male	−0.190	0.095	0.0454	−0.156	0.104	0.1344	0.372	0.002	<0.0001
Hypertension									
Yes	0.088	0.092	0.3395	−0.061	0.098	0.5316	0.269	0.019	<0.0001
Parents diabetes history									
Yes	0.417	0.088	<0.0001	0.321	0.100	0.0013	0.348	0.064	<0.0001
Age (10-year)	0.082	0.078	0.2935	0.445	0.087	<0.0001	0.075	0.018	<0.0001
BMI ($5kg/m^2$)	0.256	0.036	<0.0001	0.242	0.041	<0.0001	0.120	0.010	<0.0001
Baseline FPG ($10.4mg/dL$)	1.282	0.050	<0.0001	1.240	0.093	<0.0001	0.532	0.009	<0.0001
HDL ($17.1mg/dL$)	−0.315	0.059	<0.0001	−0.377	0.068	<0.0001	−0.027	0.010	0.0092
Total Cholesterol ($40.9mg/dL$)	−0.007	0.042	0.8634	−0.032	0.043	0.4517	0.041	0.010	<0.0001

SE.; standard error

CHAPTER4: SEMIPARAMETRIC REGRESSION MODEL FOR ANALYZING TIME-TO-EVENT DEFINED BY EXTREME LONGITUDINAL BIOMARKERS

4.1 Introduction

In many medical studies, interest focuses on studying the effects of potential risk factors on some disease events, where the occurrence time of disease events is often defined in terms of the behavior of a biomarker. For example, in diabetic studies, diabetes is defined in terms of fasting plasma glucose (FPG) being 126 mg/dl or higher. In practice, due to discrete study follow-up times, the exact time when a biomarker crosses a given threshold is unobservable, yielding so-called interval censored events (Schick and Yu 2000). In addition, most biomarker values are subject to measurement error due to imperfect technologies, so the observed biomarker values may not reflect the actual trend of the underlying biomarker. Finally, using a common threshold for defining a disease event may not be appropriate due to patient heterogeneity, which could lead to potential over-treating or under-treating some patients. Hypercholesterolemia does not cause symptoms but can significantly increase risk of developing coronary heart disease (CHD). To reduce risk, including that of CHD, people with substantially elevated cholesterol levels are advised to start therapeutic lifestyle changes or drug therapy. The cholesterol level at which to consider therapeutic intervention varies across different risk categories such as cigarette smoking, hypertension, family history of premature CHD, age, etc. (the National Cholesterol Education Program Expert Panel 2001).

Our work is motivated by the Atherosclerosis Risk in Communities (ARIC) study. In

1987-1989, the ARIC Study recruited a population-based cohort from four U.S. communities, Forsyth County, NC, Jackson, MS, suburbs of Minneapolis, MN, and Washington County, MD. Participants underwent a baseline examination in 1987-1989, three follow-up examinations at approximately three-year intervals, and a further examination in 2011-2013. The ARIC Study was designed to investigate the causes of atherosclerosis, and hypercholesterolemia is a crucial risk factor for cardiovascular disease. Hence, assessing risk factors associated with time-to- hypercholesterolemia appears to be of interest. The time-to-hypercholesterolemia data from the ARIC study pose the difficulties in analysis that we described above. Hypercholesterolemia is determined by the total cholesterol value, and this is observed at the study visits. Hence the exact incidence date for hypercholesterolemia was unobservable, and instead what is known is only the dates of the visits at which a subject's cholesterol values were below or above the specified threshold defining hypercholesterolemia. Also, total cholesterol is subject to measurement error, which is undifferentiated from intra-individual variability (Oppenheim et al. 1994). Based on total cholesterol for determining disease, the corresponding threshold levels can vary across sub-populations (the National Cholesterol Education Program Expert Panel 2001).

Despite the fact that many statistical methods have been developed for analyzing interval censored data (Ma 2010; Wen 2012; Pan 1999;Komárek and Lesaffre 2007) and measurement errors (Hausmana, Abrevayab, and Scott-Mortonb 1998; Neuhaus 2002;Paulino, Soares, and Neuhaus 2003;Pepe 1992), there is no work to address the challenges discussed above, including biomarkers measured at discrete times (interval-censored data), imperfect measurements (measurement errors), and heterogeneous threshold values for disease events. In this paper, we propose a novel semiparametric regression model for modeling biomarker values at each observed time. Our model is based on an extension of the generalized extreme value distribution, and it is equivalent to

modeling threshold-defined time-to-event via a class of Cox proportional hazards models. The paper is structured as follows. Inference procedures using the expectation-maximization (EM) algorithm for parameter estimation are presented and variance estimation are provided in Section 4.2. Asymptotic results for the proposed estimators are established in Section 4.3 and the technical detail is described in Appendix B. A simulation study and an application to the data of the ARIC study are given in Section 4.4 and Section 4.5, respectively. Some discussion is given in Section 4.6.

4.2 Method

4.2.1 Model

For subject i, let $\boldsymbol{X}_i$ be time-invariant covariates and $Y_i^*(t)$ and $Y_i(t)$ be the true biomarker value and observed biomarker value at time t, respectively. We assume $Y_i^*(t)$ to be non-decreasing over the study period, which is plausible for many chronic diseases or conditions such as diabetes, hypertension, and hypercholesterolemia as they are usually irreversible without medication or other intervention. The observed data from n i.i.d subjects are $\{\boldsymbol{X}_i, Y_i(v_{ij}), v_{ij} \mid i = 1, \ldots, n, j = 1, ..., n_i\}$, abbreviated as $\{\aleph_i \mid i = 1, \ldots, n\}$ hereafter, where $v_{i1}, ..., v_{in_i}$ are the observed times and these are assumed to be independent of $Y_i(t)$ given $\boldsymbol{X}_i$.

Since the disease event is determined by the tail behavior of the true biomarker $Y^*(t)$, our model for the distribution of $Y^*(t)$ is based on the generalized extreme value distribution, which has the form $\exp\{-\alpha \exp(-\mu y^* + \gamma)\}$ with parameters $\alpha > 0, \mu > 0$, and $-\infty < \gamma, y^* < \infty$. To further incorporate baseline covariates and account for the time-dependent nature of the biomarker values, our proposed semiparametric regression model is

$$P(Y_i^*(t) \leq y^* \mid \boldsymbol{X}_i) = \exp\{-\Lambda_0(t) \exp(-\mu y^* + \boldsymbol{\beta}^T \boldsymbol{X}_i)\}, \tag{4.1}$$

where $\Lambda_0(t)$ is non-decreasing over time and positive when $t > 0$, and both μ and $\boldsymbol{\beta}$ are unknown parameters. Interestingly, the above model in (4.1) is equivalent to modeling the threshold-defined time-to-disease events. Specifically, for any given threshold value ξ, we define $T_{i\xi}$ to be the first time that $Y_i^*(t)$ crosses the threshold ξ. Then we have $P(T_{i\xi} > t \mid \boldsymbol{X}_i) = P(Y_i^*(t) \leq \xi \mid \boldsymbol{X}_i)$. Therefore, the model in (4.1) is equivalent to assuming $P(T_{i\xi} > t \mid \boldsymbol{X}_i) = \exp\{-\Lambda_0(t)\exp(-\mu\xi + \boldsymbol{\beta}^T\boldsymbol{X}_i)\}$. That is, we obtain a proportional hazard model with a threshold-dependent baseline hazard function for $T_{i\xi}$ as

$$\lambda_i(t) = \exp(-\mu\xi)\lambda_0(t)\exp(\boldsymbol{\beta}^T\boldsymbol{X}_i), \tag{4.2}$$

where $\lambda_0(t) = d\Lambda_0(t)/dt$. This new expression gives a nice interpretation of the parameters μ and $\boldsymbol{\beta}$: $\mu > 0$ is essentially the effect of using different thresholds for the threshold-defined time to disease occurrence. Clearly, the larger the threshold, the longer the time to disease. The regression parameter $\boldsymbol{\beta}$ in model (4.2) gives the log-hazard ratio of $\boldsymbol{X}$ on time to disease occurrence defined based on any arbitrary threshold value. Therefore, $\boldsymbol{\beta}$ being positive implies that greater risk of developing disease is associated with larger values of $\boldsymbol{X}$.

Since the observed biomarker values contain measurement error, our second model considers the effect of measurement error using the classical additive measurement error model (Carroll 2006; Fuller 1987; Tsiatis, DeGruttola, and Wulfsohn 1995): $Y_i(t) = Y_i^*(t) + \epsilon_i(t),\ i = 1, \ldots, n$. We assume the measurement error $\epsilon_i(t)$ has a normal distribution with mean zero and variance σ^2 for any time t and is independent of $Y_i^*(t)$, $\boldsymbol{X}_i$, and ξ. The measurement error variance of σ^2 may be estimated in practice by taking repeated measurements. As an example, the National Cholesterol Education Program and Laboratory Standardization Panel established the goal that a single serum total cholesterol measurement should be accurate within 8.9 percent. Since information about the measurement error variation can be obtained from outside

the dataset being analyzed, we consider σ^2 to be known.

Under the above two models, we obtain the marginal likelihood for the observed biomarker $\{Y_i(v_{ij})\}$ given $\boldsymbol{X}_i, i = 1, \ldots, n,\ j = 1, \ldots, n_i$ as

$$l_n^{ps} = \sum_{i=1}^{n} \sum_{v: dN_i(v)=1} \log \int_{-\infty}^{\infty} \exp\left\{-\Lambda_0(v) e^{\boldsymbol{\beta}^T \boldsymbol{X}_i - \mu\xi_{iv}}\right\} \Lambda_0(v) \mu \exp(\boldsymbol{\beta}^T \boldsymbol{X}_i - \mu\xi_{iv}) \times \frac{1}{\sigma} \phi\left\{\frac{Y_i(v) - \xi_{iv}}{\sigma}\right\} d\xi_{iv}, \tag{4.3}$$

where $\phi(\cdot)$ is the standard normal density function.

4.2.2 Inference Procedure

We maximize (4.3) to estimate all the parameters, including $\boldsymbol{\theta} = (\mu, \boldsymbol{\beta}^T)^T$ and $\Lambda(t)$. Specifically, we estimate $\Lambda(t)$ as a step function, with jumps at the observed times. Let $v_{(1)} < \ldots < v_{(K)}$ be ordered observed times of $\{v_{ij} \mid i = 1, \ldots, n, j = 1, \ldots, n_i\}$ and $\Lambda_k = \Lambda_0(v_{(k)})$ and $v_{(0)} = 0$. Then we maximize (4.3) over $\boldsymbol{\theta}$ and the Λ_k's, subject to constraints $0 \le \Lambda_1 \le \cdots \le \Lambda_K$.

To facilitate the maximization, we adopt the EM algorithm by treating the threshold values ξ as missing data. Then the complete log marginal likelihood function is

$$\begin{aligned} l_c^{ps}\{\theta, \Lambda, \aleph\} \quad = \quad & \sum_{k=1}^{K} \sum_{i=1}^{n} \sum_{j=1}^{n_i} I(v_{ij} = v_{(k)}) \Bigg[-\Lambda_k \exp(\boldsymbol{\beta}^T \boldsymbol{X}_i - \mu\xi_{ij}) + \log \Lambda_k + \log \mu \\ & + \boldsymbol{\beta}^T \boldsymbol{X}_i - \mu\xi_{ij} - \frac{1}{2} \log \sigma^2 - \frac{\{Y_i(v_{ij}) - \xi_{ij}\}^2}{2\sigma^2} \Bigg]. \end{aligned} \tag{4.4}$$

In the maximization step at the lth iteration of the EM algorithm, we first maximize the conditional expectation of the complete log marginal likelihood function given observed data over Λ_k's. We then update $\boldsymbol{\theta}$ via the Newton-Raphson algorithm. Specifically, we

maximize $Q(\Lambda)$ defined by

$$Q(\Lambda) = \sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{j=1}^{n_i} I(v_{ij} = v_{(k)}) E\left\{-\Lambda_k \exp(\boldsymbol{\beta}^T \boldsymbol{X}_i - \mu\xi_{ij}) + \log\Lambda_k \mid \aleph_i, \boldsymbol{\theta}^{(l)}\right\}. \tag{4.5}$$

Since $Q(\Lambda)$ is a concave function over a convex cone satisfying $\Lambda_1 \leq \ldots \leq \Lambda_K$, this maximization can be carried out using one of the many existing algorithms for convex optimization. To update $\boldsymbol{\theta}$, we apply the following one-step Newton-Raphson algorithm, $\boldsymbol{\theta}^{(l+1)} = \boldsymbol{\theta}^{(l)} + E\left(-\partial^2 l_c^{ps}/(\partial\theta)^2 \mid \aleph, \boldsymbol{\theta}^{(l)}\right)^{-1}_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(l)}} E\left(\partial l_c^{ps}/\partial\theta \mid \aleph, \boldsymbol{\theta}^{(l)}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(l)}}$. The conditional expectations are calculated in the expectation step of the EM algorithm based on the following expression

$$E\{g(\xi)|\aleph_i, \boldsymbol{\theta}^{(l)}\} = \frac{I(v_{ij} = v_{(k)})\int_{-\infty}^{\infty} g(\xi)\exp(-\Lambda_k e^{\boldsymbol{\beta}^T \boldsymbol{X}_i - \mu\xi}) e^{-\mu\xi}\phi\left\{\frac{Y_i(v_{ij})-\xi}{\sigma}\right\} d\xi}{\int_{-\infty}^{\infty}\exp(-\Lambda_k e^{\boldsymbol{\beta}^T \boldsymbol{X}_i - \mu\xi}) e^{-\mu\xi}\phi\left\{\frac{Y_i(v_{ij})-\xi}{\sigma}\right\} d\xi}, \tag{4.6}$$

where the $g(\xi)$'s to be calculated are ξ, ξ^2, $e^{-\mu\xi}$, $e^{-\mu\xi}\xi$, and $e^{-\mu\xi}\xi^2$. This integration can be approximated by the Gauss-Hermite quadrature (Davis 1984), so it can be approximated by

$$\sum_{k=1}^{N}\left(\sqrt{2}\sigma w_k g\{\sqrt{2}\sigma w_k + Y_i(v_{ij})\}\exp\left[-\Lambda_0(v_{ij}) e^{\boldsymbol{\beta}^T X_i - \mu\{\sqrt{2}\sigma z_k + Y_i(v_{ij})\}}\right] e^{-\mu\{\sqrt{2}\sigma z_k + Y_i(v_{ij})\}}\right), \tag{4.7}$$

where N is the number of the quadratures and ω_k and z_k are weights and abscissae for the Gauss-Hermite quadrature, respectively. This loop of the E-step and the M-steps is repeated until $\left|\boldsymbol{\theta}^{(l+1)} - \boldsymbol{\theta}^{(l)}\right|$ is smaller than a pre-specified criterion. We denote the final estimators as $\widehat{\boldsymbol{\theta}} = (\widehat{\mu}, \widehat{\boldsymbol{\beta}}^T)^T$ and $\widehat{\Lambda}$.

4.2.3 Variance Estimation

To derive the marginal influence functions leading to asymptotic variance estimation for the regression parameter, we need the complete log marginal likelihood function and

the marginal score function with respect to $\boldsymbol{\theta}$. These are given by

$$\begin{aligned} l_c^{ps}\{\theta,\Lambda,\aleph\} &= \int_0^\infty \Big[-\Lambda(v)\exp(\boldsymbol{\beta}^T\boldsymbol{X}-\mu\xi)+\log\Lambda(v)+\log\mu+\boldsymbol{\beta}^T\boldsymbol{X} \\ &\quad -\mu\xi-\frac{1}{2}\log\sigma^2-\frac{\{Y(v)-\xi\}^2}{2\sigma^2}\Big]dN(v), \qquad (4.8) \\ \dot{l}_\mu^{ps}\{\boldsymbol{\theta},\Lambda,\aleph\} &= E_\xi\{\partial l_c^{ps}/\partial\mu \mid \aleph\} = \int_0^\infty \left[\mu^{-1}-E_\xi\left\{\kappa(v)\xi \mid \aleph\right\}\right]dN(v), \qquad (4.9) \\ \dot{l}_\beta^{ps}\{\boldsymbol{\theta},\Lambda,\aleph\} &= E_\xi\{\partial l_c^{ps}/\partial\beta \mid \aleph\} = \boldsymbol{X}\int_0^\infty E_\xi\left\{\kappa(v) \mid \aleph\right\}dN(v), \qquad (4.10) \end{aligned}$$

where $N(v)$ denotes the counting process associated with measurement times and $\kappa(v)=1-\Lambda(v)\exp(\boldsymbol{\beta}^T\boldsymbol{X}-\mu\xi)$.

Let $\{P_{\boldsymbol{\theta},\Lambda_\eta}\}$ be a regular parametric subfamily of models, $\{P_{\boldsymbol{\theta},\Lambda} \mid P_{\boldsymbol{\theta},\Lambda} \ll m, m$:Lebesque measure$\}$ and set $\partial/\partial\eta|_{\eta=0}\Lambda_\eta(v)=h(v)$ for $v>0$ and $h(v)\in L_2(P_V)$. Then we have a score operator for Λ:

$$\dot{l}_\Lambda^{ps}\{\boldsymbol{\theta},\Lambda,\aleph\}[h(v)] = \int_0^\infty h(v)E(\kappa(v) \mid \aleph)/\Lambda(v)dN(v). \qquad (4.11)$$

Furthermore, we define

$$\begin{aligned} h_1(v) &= E\Big\{e^{\boldsymbol{\beta}^T\boldsymbol{X}}\Big(E\{\xi \mid \aleph\}E\{e^{-\mu\xi} \mid \aleph\}-2E\{\xi e^{-\mu\xi} \mid \aleph\} \\ &\quad -e^{\boldsymbol{\beta}^T\boldsymbol{X}}\Lambda(v)\left[E\{\xi e^{-\mu\xi} \mid \aleph\}E\{e^{-\mu\xi} \mid \aleph\}-E\{\xi e^{-2\mu\xi} \mid \aleph\}\right]\Big) \mid V=v\Big\}, \\ h_2(v) &= E\Big\{\boldsymbol{X}e^{\boldsymbol{\beta}^T\boldsymbol{X}}\Big(\Lambda(v)e^{\boldsymbol{\beta}^T\boldsymbol{X}}\left[E\{e^{-\mu\xi} \mid \aleph\}^2-E\{e^{-2\mu\xi} \mid \aleph\}\right]+E\{e^{-\mu\xi} \mid \aleph\}\Big) \mid V=v\Big\}, \\ h_3(v) &= E\Big(\Lambda(v)^{-2}+e^{2\boldsymbol{\beta}^T\boldsymbol{X}}\left[E\{e^{-\mu\xi} \mid \aleph\}^2-E\{e^{-2\mu\xi} \mid \aleph\}\right] \mid V=v\Big), \\ h_\mu^*(v) &= h_1(v)/h_3(v), \qquad (4.12) \\ h_\beta^*(v) &= h_2(v)/h_3(v). \qquad (4.13) \end{aligned}$$

In the asymptotic proofs given later, we show that the asymptotic covariance of $\widehat{\boldsymbol{\theta}}$

takes a sandwich variance form:

$$I(\boldsymbol{\theta}_0) = \boldsymbol{D}^{-1} P[\mathbf{A}\mathbf{A}^T](\boldsymbol{D}^{-1})^T, \tag{4.14}$$

where $\mathbf{A} = \dot{l}^{ps}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \Lambda_0) - \dot{l}^{ps}_{\Lambda}(\boldsymbol{\theta}_0, \Lambda_0)[h^*_{\boldsymbol{\theta}}(v)]$ and

$$\boldsymbol{D} = P\left(\ddot{l}^{ps}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \Lambda_0) - \ddot{l}^{ps}_{\boldsymbol{\theta}\Lambda}(\boldsymbol{\theta}_0, \Lambda_0)[h^*_{\boldsymbol{\theta}}(v)]\right).$$

Therefore, we can use the empirical data to estimate this covariance matrix. Specifically, the matrix $\boldsymbol{D}$ can be estimated as

$$\widehat{\boldsymbol{D}} = n^{-1}\sum_{i=1}^{n}\begin{pmatrix} \hat{\ddot{l}}^{ps(i)}_{\mu\mu} & \hat{\ddot{l}}^{ps(i)}_{\mu\boldsymbol{\beta}} \\ \hat{\ddot{l}}^{ps(i)}_{\boldsymbol{\beta}\mu} & \hat{\ddot{l}}^{ps(i)}_{\boldsymbol{\beta}\boldsymbol{\beta}} \end{pmatrix} - n^{-1}\sum_{i=1}^{n}\begin{pmatrix} \hat{\ddot{l}}^{ps}_{\boldsymbol{\theta}\Lambda}[\hat{h}^*_{\mu}(v)] \\ \hat{\ddot{l}}^{ps}_{\boldsymbol{\theta}\Lambda}[\hat{h}^*_{\boldsymbol{\beta}}(v)] \end{pmatrix},$$

where

$$\begin{aligned}
\hat{\ddot{l}}^{ps(i)}_{\mu\mu} &= \int_0^{\infty}\Big(-\widehat{\mu}^{-2} - \widehat{\Lambda}(v)e^{\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}_i}\left[3\widehat{E}\{\xi^2 e^{-\mu\xi} \mid \boldsymbol{\aleph}_i\} - 2\widehat{E}\{\xi \mid \boldsymbol{\aleph}_i\}\widehat{E}\{\xi e^{-\mu\xi} \mid \boldsymbol{\aleph}_i\}\right] \\
&\quad + \widehat{\Lambda}(v)^2 e^{2\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}_i}\left[\widehat{E}\{\xi^2 e^{-2\mu\xi} \mid \boldsymbol{\aleph}_i\} - \widehat{E}\{\xi e^{-\mu\xi} \mid \boldsymbol{\aleph}_i\}^2\right] - \widehat{E}\{\xi \mid \boldsymbol{\aleph}_i\}^2 \\
&\quad + \widehat{E}\{\xi^2 | \boldsymbol{\aleph}_i\}\Big)dN_i(v), \\
\hat{\ddot{l}}^{ps(i)}_{\mu\boldsymbol{\beta}} &= \int_0^{\infty}\boldsymbol{X}_i\widehat{\Lambda}(v)e^{\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}_i}\Big(2\widehat{E}(\xi e^{-\mu\xi} \mid \boldsymbol{\aleph}_i) - \widehat{E}\{e^{-\mu\xi} \mid \boldsymbol{\aleph}_i\}\widehat{E}\{\xi \mid \boldsymbol{\aleph}_i\} \\
&\quad + \widehat{\Lambda}(v)e^{\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}_i}\left[\widehat{E}\{e^{-\mu\xi} \mid \boldsymbol{\aleph}_i\}\widehat{E}\{\xi e^{-\mu\xi} \mid \boldsymbol{\aleph}_i\} - \widehat{E}\{\xi e^{-2\mu\xi} \mid \boldsymbol{\aleph}_i\}\right]\Big)dN_i(v), \\
\hat{\ddot{l}}^{ps(i)}_{\boldsymbol{\beta}\boldsymbol{\beta}} &= -\boldsymbol{X}_i\boldsymbol{X}_i^T\int_0^{\infty}\widehat{\Lambda}(v)e^{\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}_i}\Big(\widehat{\Lambda}(v)e^{\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}_i}\left[\widehat{E}\{e^{-\mu\xi} \mid \boldsymbol{\aleph}_i\}^2 - \widehat{E}\{e^{-2\mu\xi} \mid \boldsymbol{\aleph}_i\}\right] \\
&\quad + \widehat{E}\{e^{-\mu\xi} \mid \boldsymbol{\aleph}_i\}\Big)dN_i(v), \\
\widehat{\ddot{l}}^{ps(i)}_{\mu\Lambda}[\widehat{h}^*_{\mu}(v)] &= \int_0^{\infty}\widehat{h}^*_{\mu}(v)e^{\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}_i}\Big(2\widehat{E}\{\xi e^{-\mu\xi} \mid \boldsymbol{\aleph}_i\} - \widehat{E}\{\xi \mid \boldsymbol{\aleph}_i\}\widehat{E}\{e^{-\mu\xi} \mid \boldsymbol{\aleph}_i\} \\
&\quad + \widehat{\Lambda}(v)e^{\widehat{\boldsymbol{\beta}}^T\boldsymbol{X}_i}\left[\widehat{E}\{\xi e^{-\mu\xi} \mid \boldsymbol{\aleph}\}\widehat{E}\{e^{-\mu\xi} \mid \boldsymbol{\aleph}_i\} - \widehat{E}\{\xi e^{-2\mu\xi} \mid \boldsymbol{\aleph}_i\}\right]\Big)dN_i(v),
\end{aligned}$$

$$\widehat{\dot{l}}^{ps(i)}_{\boldsymbol{\beta}\Lambda}[\widehat{h}^*_\beta(v)] = -\boldsymbol{X}_i \int_0^\infty \widehat{h}^*_\beta(v) e^{\widehat{\boldsymbol{\beta}}^T \boldsymbol{X}_i} \Big(\widehat{\Lambda}(v) e^{\widehat{\boldsymbol{\beta}}^T \boldsymbol{X}_i} \big[\widehat{E}\{e^{-\mu\xi} \mid \boldsymbol{\aleph}_i\}^2 - \widehat{E}\{e^{-2\mu\xi} \mid \boldsymbol{\aleph}\} \big] + \widehat{E}\{e^{-\mu\xi} \mid \boldsymbol{\aleph}_i\} \Big) dN_i(v).$$

Here, $\widehat{h}^*_\theta(v) = (\widehat{h}^*_\mu(v), \widehat{h}^*_\beta(v))^T$ is estimated as follows:

$$\begin{aligned}
\widehat{h}^*_\mu(v) &= \sum_{i=1}^n \sum_{j=1}^{n_i} K_{h_n}(v_{ij} - v) e^{\widehat{\boldsymbol{\beta}}^T \boldsymbol{X}_i} \Big(\widehat{E}\{\xi \mid \boldsymbol{\aleph}\} \widehat{E}\{e^{-\mu\xi} \mid \boldsymbol{\aleph}\} - 2\widehat{E}\{\xi e^{-\mu\xi} \mid \boldsymbol{\aleph}\} \\
&\quad - e^{\widehat{\boldsymbol{\beta}}^T \boldsymbol{X}_i} \widehat{\Lambda}(v_{ij}) \big[\widehat{E}(\xi e^{-\mu\xi} \mid \boldsymbol{\aleph}) \widehat{E}\{e^{-\mu\xi} \mid \boldsymbol{\aleph}\} - \widehat{E}\{\xi e^{-2\mu\xi} \mid \boldsymbol{\aleph}\} \big] \Big) \\
&\quad \Big/ \sum_{i=1}^n \sum_{j=1}^{n_i} K_{h_n}(v_{ij} - v) g\{\boldsymbol{\aleph}\}, \qquad (4.15) \\
\widehat{h}^*_\beta(v) &= \sum_{i=1}^n \sum_{j=1}^{n_i} K_{h_n}(v_{ij} - v) \boldsymbol{X}_i e^{\widehat{\boldsymbol{\beta}}^T \boldsymbol{X}_i} \Big(\widehat{E}\{e^{-\mu\xi} \mid \boldsymbol{\aleph}\} \\
&\quad + \widehat{\Lambda}(v_{ij}) e^{\widehat{\boldsymbol{\beta}}^T \boldsymbol{X}_i} \big[\widehat{E}\{e^{-\mu\xi} \mid \boldsymbol{\aleph}\}^2 - \widehat{E}\{e^{-2\mu\xi} \mid \boldsymbol{\aleph}\} \big] \Big) \\
&\quad \Big/ \sum_{i=1}^n \sum_{j=1}^{n_i} K_{h_n}(v_{ij} - v) g\{\boldsymbol{\aleph}\}, \qquad (4.16)
\end{aligned}$$

where $g\{\boldsymbol{\aleph}\} = \widehat{\Lambda}(v_{ij})^{-2} + e^{2\widehat{\boldsymbol{\beta}}^T \boldsymbol{X}_i} \big[\widehat{E}\{e^{-\mu\xi} \mid \boldsymbol{\aleph}\}^2 - \widehat{E}\{e^{-2\mu\xi} \mid \boldsymbol{\aleph}\} \big]$ and $K_{h_n}(x) = h_n^{-1} \exp(-x^2/h_n)$ with h_n being a kernel bandwidth. In particular, we choose h_n as $(cn)^{-1/2}$ for some constant $c > 0$. We estimate the middle term in the sandwich variance form as follows:

$$\widehat{P}\left(\mathbf{A}\mathbf{A}^T\right) = n^{-1} \sum_{i=1}^n \left\{ \left(\hat{l}^{ps(i)}_{\boldsymbol{\theta}} - \hat{l}^{ps(i)}_{\Lambda}[\widehat{h}^*_{\boldsymbol{\theta}}(v)] \right) \left(\hat{l}^{ps(i)}_{\boldsymbol{\theta}} - \hat{l}^{ps(i)}_{\Lambda}[\widehat{h}^*_{\boldsymbol{\theta}}(v)] \right)^T \right\},$$

where $\hat{l}^{ps(i)}_\theta = (\hat{l}^{ps(i)}_\mu, \hat{l}^{ps(i)T}_{\boldsymbol{\beta}})^T$,

$$\begin{aligned}
\hat{l}^{ps(i)}_\mu &= \sum_{j=1}^{n_i} \left[\widehat{\mu}^{-1} - E\left\{ \widehat{\kappa}(v_{ij}) \xi \right) \mid \boldsymbol{\aleph}_i \right\} \Big], \\
\hat{l}^{ps(i)}_{\boldsymbol{\beta}} &= \boldsymbol{X}_i \sum_{j=1}^{n_i} E_\xi \left\{ \widehat{\kappa}(v_{ij}) \mid \boldsymbol{\aleph}_i \right\},
\end{aligned}$$

$$\hat{l}^{ps(i)}_\Lambda[\widehat{h}^*_{\boldsymbol{\theta}}(v)] = \sum_{j=1}^{n_i} E_\xi \left\{ \widehat{\kappa}(v_{ij}) \mid \boldsymbol{\aleph}_i \right\} \widehat{h}^*_{\boldsymbol{\theta}}(v_{ij}) / \widehat{\Lambda}(v_{ij}),$$

and $\widehat{\kappa}(v) = 1 - \widehat{\Lambda}(v)\exp(\widehat{\boldsymbol{\beta}}^T\boldsymbol{X} - \widehat{\mu}\xi)$. Consequently, an estimator for the asymptotic covariance of $\widehat{\boldsymbol{\theta}}$ is $\widehat{I} = \widehat{\boldsymbol{D}}\widehat{P}(\mathbf{A}\mathbf{A}^T)\widehat{\boldsymbol{D}}^{-1}$. In the next section, we show that this variance estimator is consistent if assuming that h_n and $\log\{n/(nh_n)\}$ go to 0 as $n \to \infty$.

4.3 Asymptotic Results

In this section, we provide asymptotic results for the proposed estimators and the technical detail is summrized in Appendix II Let $\boldsymbol{\theta}_0$ and Λ_0 denote the true regression parameter and cumulative hazard function, respectively. We need the following conditions:

(A1) The finite-dimensional parameter space $\boldsymbol{\Theta}$ is a compact subset of the domain of $\boldsymbol{\theta}$.

(A2) The covariate $\boldsymbol{X}$ has bounded support with probability 1. If $\boldsymbol{\beta}^T\boldsymbol{X} + \alpha = 0$ almost surely (a.s.), then $\boldsymbol{\beta} = 0$ and $\alpha = 0$.

(A3) The support of the observation time, V, is an interval $\mathscr{S}[V] = [l_V, u_V]$, with $0 < l_V \leq u_V < \infty$.

(A4) The number of the observation times, $\int_0^{u_V} dN(V)$ is P_V-almost surely finite.

(A5) The cumulative hazard function Λ_0 has strictly positive derivative on $\mathscr{S}[V]$.

The assumptions that parameter, covariate, and observation time are bounded are standard. Condition (A2) ensures the identifiability of $\boldsymbol{\theta}$ and Λ_0. These conditions hold naturally in most applications.

For convergence of the estimators to the true parameters, we need to define a topology. Let the bounded regression parameter space $\boldsymbol{\Theta}(\subset R^d)$ be equipped with the Euclidean topology. Regarding infinite dimensional nonparametric space, let $\mathscr{F}$ be the set of all Borel subprobability measures on $\mathscr{S}[V]$. Then $\mathscr{F}$ can be equipped with the

vague topology by defining that, for any sequence $F_n \in \mathscr{F}$ and $F \in \mathscr{F}$, F_n converges vaguely to F if and only if

$$\int f dF_n \to \int f dF \quad \text{for every} \quad f \in C_0(\mathscr{S}[V]),$$

where $C_0(\mathscr{S}[V])$ is the set of all continuous functions that vanish outside a compact subset of $\mathscr{S}[V]$. Then the product space $\boldsymbol{\Theta} \times \mathscr{F}$ can be equipped with the product topology of the Euclidean topology and the vague topology. In the product topology, it is said that $(\widehat{\boldsymbol{\theta}}, \widehat{F})$ converges to $(\boldsymbol{\theta}, F)$, when $\widehat{\boldsymbol{\theta}}$ and $\widehat{F}$ converge to $\boldsymbol{\theta}$ and F, respectively.

Theorem 4.3.1. *(Consistency of the MLE) Suppose that conditions, (A1), (A2), (A3), and (A4) are satisfied, then $\widehat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}_0$ a.s., and if $v \in \mathscr{S}[V]$ is a continuity point of Λ_0, $\widehat{\Lambda}(v)$ converges to $\Lambda_0(v)$ a.s. Moreover, if Λ_0 is continuous, then $\sup_{v \in \mathscr{S}[V]} |\widehat{\Lambda}(v) - \Lambda_0(v)|$ converges to 0 a.s.*

Before discussing the overall convergence rate, we define the distance d on $R^d \times \Phi$ as follows: $d\{(\boldsymbol{\theta}_1, \Lambda_1), (\boldsymbol{\theta}_2, \Lambda_2)\} = |\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2| + \|\Lambda_1 - \Lambda_2\|_{2,P}$, where $|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2|$ is the Euclidean distance in R^d, $\|\Lambda_1 - \Lambda_2\|_{2,P_V} = \left(\int (\Lambda_1(v) - \Lambda_2(v))^2 dP_V\right)^{1/2}$ and P_V is the marginal probability measure of the measurement time variable V.

We apply Theorem 3.4.1 of van der Vaart and Wellner (1996) and Lemma $B.2.1$ in Appendix B to obtain the rate of convergence.

Theorem 4.3.2. *(Rate of convergence) Suppose that conditions (A1), (A2), (A3), and (A4) are satisfied. Then $d\left\{(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda}), (\boldsymbol{\theta}_0, \Lambda_0)\right\} = O_p(n^{-1/3})$.*

The overall rate of convergence is dominated by $\widehat{\Lambda}$. However, it is shown in the next theorem that the convergence rate of $\widehat{\boldsymbol{\theta}}$ can be refined to achieve a rate of $\sqrt{n}$. The convergence rate we found is applied to prove the asymptotic normality of the regression parameter MLE.

Theorem 4.3.3. *(Asymptotic normality) Suppose that θ_0 is an interior point of Θ and that conditions $(A1)$–$(A5)$ are satisfied. Then*

$$n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = -n^{1/2}(\mathbb{P}_n - P)\widetilde{\psi}^{ps}(\boldsymbol{\aleph}) + o_p(1) \to N(0, I(\theta_0)) \quad \textit{in distribution,}$$

where P is the probability measure, that is, $P\widetilde{\psi}^{ps}(\boldsymbol{\aleph}) = \int \widetilde{\psi}^{ps}(\boldsymbol{\aleph})dP$, $\mathbb{P}_n$ is the empirical measure of $\boldsymbol{\aleph}_i$, $i = 1, \ldots, n$, that is, $\mathbb{P}_n\widetilde{\psi}^{ps}(\boldsymbol{\aleph}) = n^{-1}\sum_{i=1}^{n}\widetilde{\psi}^{ps}(\boldsymbol{\aleph})$, $\widetilde{\psi}^{ps}(\boldsymbol{\aleph})$ is the marginal influence function defined as $\widetilde{\psi}^{ps} = \boldsymbol{D}^{-1}\left\{\dot{l}^{ps}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \Lambda_0) - \dot{l}^{ps}_{\Lambda}(\boldsymbol{\theta}_0, \Lambda_0)[h^{*}_{\boldsymbol{\theta}}(v)]\right\}$, and $I(\theta_0)$ is the information in (4.14).

Theorem 4.3.4. *(Consistency of the asymptotic variance estimator) When the bandwidth h_n satisfies that h_n and $\log\{n/(nh_n)\}$ converge to 0 as $n \to \infty$, then $\widehat{I}$ converges to $I(\boldsymbol{\theta}_0)$ in probability.*

4.4 Simulation Study

We consider scenarios of longitudinal and random measurement time points. The number of the measurement times per subject is three, and the measurement times, $v_{ij}, j = 1, 2, 3$ are independently generated from the normal distributions with means of 0.5, 1.2, and 1.9, respectively and with the common standard deviation of 0.1, because study visit windows are usually fixed, but each visit time varies across subjects. In the simulation, two covariates are included in the model: one is generated from the Bernoulli distribution with probability 0.5, and the other is from the normal distribution with mean 0 and variance 0.1. The true values for (β_1, β_2) were set as (0.3, 0.3), and the true cumulative baseline hazard assumed to be $2t^{1/5}$.

Consequently, the true biomarker value is generated as following:

$$Y_i^*(v_{ij}) \quad = \quad \mu^{-1}\left\{\boldsymbol{\beta}^T\boldsymbol{X}_i + \log\Lambda_0(v_{ij}) - \log\left(-\log p_i\right)\right\} \text{ for } 1 \le i \le n, 1 \le j \le 3, \quad (4.17)$$

where p_i independently follows the uniform distribution (0,1) for all i. The observed biomarker value is generated by $Y_i(v_{ij}) = Y_i^*(v_{ij}) + \epsilon(v_{ij})$, where $\epsilon(v_{ij})$ is independently generated from the normal distribution with zero mean and some finite variance for all i and j and is independent of $Y_i^*(v_{ij})$. We consider two measurement error variances of 0.25 and 1.0. The measurement error variance determines a ratio of measurement error variance to true biomarker value variance and correlations of observed biomarker values within a subject; using measurement error variances of 0.25 and 1.0, the variance ratios are 0.15 and 0.60, respectively, and the correlations are 0.87 and 0.62, respectively. We varied sample sizes from 200 (600 observations) to 400 (1,200 observations) and conducted 1,000 replications.

For each simulated dataset, we applied the proposed EM algorithm to estimate the parameters. The initial values used for β and $\Lambda_0(t)$ in the algorithm were 0's and observed times, respectively. In the maximization-step, the spectral projected gradient method was used for constrained optimization in (4.5). The convergence criterion for the expectation-maximization algorithm was set as 10^{-10}. In the simulations, we noticed that the threshold effect μ was sensitive to the initial values. Therefore, we first calculated the profile likelihood of μ using the same algorithm except that μ was held at some fixed value; we then carried out a grid search to find the maximum likelihood estimate for μ. The variance estimation was based on the formula in Section 4.2.3. We need to decide bandwidth for the variance estimation, and it depends on data. Relatively sparse data in measurement time causes unstable kernel estimate for $h_\theta^*(v)$ in (4.12) and (4.13). The bandwidth is set to be $(n/20)^{-1/2}$. For comparison, we also calculated the maximum likelihood estimates assuming that the threshold value was the same for all the subjects, and there was no measurement error. These estimates can be calculated using the algorithm suggested by Pan (1999) and implemented in the R package, "'intcox"' version 0.9.3 developed by Henschel et al. (2007). Pan (1999)'s

method (referred to as the ICM method hereafter) does not provide a variance estimate. Every subject may have a different threshold in the simulation scenario; however, we need a fixed threshold value for Pan (1999)'s estimate. Then we set fixed thresholds to be 90%, 80%, or 70% quantiles of observed biomarker values.

Table 4.1 shows that bias of the proposed estimators is small, and decreases as the sample size increases or the variance ratio decreases; the estimated variance estimates (summarized by their median value) agree with the empirical variance. Out of 1,000 sets of the simulated data, the asymptotic variance estimate is likely to be overestimated in 10-15 % datasets because of unstable cumulative hazard function estimates at the last observation times. Hence, when the asymptotic variance is calculated using the MLEs, we excluded 2% observations from the simulated data, which corresponds to the last observation times. Consequently, the coverage probability is larger than 95 % when the measurement error variance is 1.0; however it seems to be acceptable as the sample size increases. In the setting in which the threshold value varies from person to person, and the biomarker value includes measurement error, the empirical standard errors of our estimates are much smaller than the estimates of the ICM method, and the bias of our estimate is smaller than that of the ICM method's estimate when measurement error variance is large. From the simulation study, it is also seen that the ICM method is less biased and less efficient the higher the fixed threshold; and the standard error and bias of the ICM method decrease when the sample size increases; accuracy and efficiency of the ICM method are more influenced by measurement error than by the choice of fixed threshold. Additionally, we examined the numerical performance of our method when measurement errors dominate true biomarker values, particularly, the variance ratio of measurement error to observed values is 1.5. In the case, bias increases but is still acceptable; however, the asymptotic variance estimate based on the band-width of $(n/100)^{-1/2}$ is very likely to be overestimated.

4.5 Application

Hypercholesterolemia, that is, high blood cholesterol, is not a disease but a metabolic derangement that can be secondary to many diseases and can contribute to many forms of disease, most notably cardiovascular disease. We are interested in assessing risk factors associated with time-to-hypercholesterolemia, so we apply our proposed models to the data of 10,236 subjects from the four U.S. communities in the ARIC study, Jackson, MS; Forsyth County, NC; suburbs of Minneapolis, MN; and Washington County, MD. The participants were predominantly white or African-American, and the few participants of other races are excluded from the analysis, as is usually done in analyses of ARIC Study data.

In the models, we consider the baseline covariates including race, gender, hypertension, parents' coronary heart disease (CHD) history, categorized age (<50,50-60, ≥60), total cholesterol, and interaction effect between age and sex. These variables are generally regarded as major factors associated with time-to-hypercholesterolemia. All subjects with complete data for the baseline (visit 1) covariates are included in the analysis. Demographic characteristics of the participants in the dataset used here include average age of 53.7 years (range 44-66 years), white race 7,832 (76.5%), and females 5,748 (56.2%). The average total cholesterol at visit 1 was 208.6 (±38.0) mg/dL, and the number of the participants with hypertension and parental history of CHD were 2,920 (28.5 %) and 4,003(39.1%), respectively.

To facilitate calculation, we standardized total cholesterol by the sample mean of $203.4mg/dl$ and standard deviation of $35.9mg/dl$ so that it has zero mean and the unit variance. The observation time is scaled down to (0,1]. The standardized value and the rescaled observation time better facilitate the estimation process than the original value or log-transformed value. The National Cholesterol Education Program and Laboratory Standardization Panel established the goal that a single serum total

cholesterol measurement should be accurate within ±8.9 percent. The Health Care Financing Administration (HCFA) has also established similar testing requirements for total cholesterol (±10 percent), authored by the Centers for Disease Control and Prevention (Oppenheim et al. 1994). Hence, we chose $\sigma^2 = 0.3^2$ for the measurement error of the standardized FPG value, which corresponds to 0.09 for the variance ratio of measurement error to total cholesterol value.

For comparison, we also applied the ICM method, which ignores the measurement errors and uses a fixed threshold. In the ICM method, the fixed threshold value was set to be 240 mg/dL. We used the simple bootstrap sampling method with 200 replications to estimate the standard error of the regression parameter estimates for the ICM method. We investigated the robustness of the ICM model to the choice of threshold (200 and 270mg/dL); the effect directions remained unchanged but the effect sizes varied; significant factors related to time-to-hypercholesterolemia differed according to the threshold used.

The variance estimate for effect size is somewhat sensitive to the choice of bandwidth, so we employed a subsampling bootstrap with sample size of 500 subjects and 350 repetitions and adjusted the standard error based on the bootstrap by multiplying by the factor $\sqrt{500/10,236}$. In simulation data, the subsample bootstrap based standard error precisely estimates the true standard error. The average of the bootstrap-based estimates is also very close to the estimate of our method. The bootstrap-based estimate and the adjusted standard error are presented in Table 4.2.

In the ARIC Study data, African-Americans, having parental history of CHD, and high baseline total cholesterol have 1.56, 1.31, and 1.03 times greater hazard of hypercholesterolemia than people with the opposite characteristics, respectively. When baseline total cholesterol increases by 1 unit, the hazard of hypercholesterolemia increases by a factor of 0.025. There is significant interaction effect between age and

sex; men is very likely to be high-risk for hypercholesterolemia than women; as getting older, hazard ratio for men to women in hypercholesterolemia increases from 1.56 to 1.59, respectively; however, it decreases to 1.43 after 60-year old because of menopause.

To investigate the goodness-of-fit of our model, we generated predicted total cholesterol values using the formula in (4.17) based on the estimates. Using the predicted values, we suggest two graphical methods for model diagnosis. First, a quantile-quantile plot is generated to compare the distribution of the real observed total cholesterol values with the predicted distribution (left panel in Figure 4.1). The quantile-quantile plot shows that the distribution of predicted values matches the distribution of observations very closely. Secondly, we calculated the residuals by subtracting the predicted means from the real total cholesterol values (right panel in Figure 4.1). The residual plot in Figure 4.1 shows a fairly good fit and the residuals are randomly scattered around 0. The estimated cumulative hazard function at the last time points is less stable, so the residual of the observation at the last time is relatively large.

4.6 Concluding Remarks

We proposed a semiparametric extreme-value regression model for the highly skewed distribution of a biomarker subject to measurement error, estimated the model parameters using the marginal likelihood, and implemented computation via the pseudo-EM algorithm. In a numerical study, the proposed method shows good accuracy and acceptable coverage probability unless measurement error dominates observed values. The method is illustrated through an application to data from the ARIC Study.

The proposed model is based on the marginal likelihood method, so it is not guaranteed to satisfy semiparametric efficiency. To enhance the efficiency of the regression

parameter estimate, we can consider weighted estimating equations based on score functions. The weights would be optimal when they lead to the lower bound of the asymptotic variance. However, for ease of calculation we can consider alternative weights to reduce variance of the regression parameters and to be piecewise constants over time. Then based on the estimates from the pseudo-EM algorithm and the optimal weight, we would be able to obtain a weighted sandwich variance, which is smaller than the proposed asymptotic variance.

The proposed model can be extended in various ways; instead of a linear model for the time-invariant threshold effect, we can incorporate a time-dependent and nonparametric function for the threshold effect in the model; when a covariance structure for true biomarker values is postulated, a semiparametric maximum likelihood method can be constructed; the proposed model can also be extended to a frailty model accounting for random effects for clusters.

Figure 4.1: Quantile-Quantile and Residual Plots

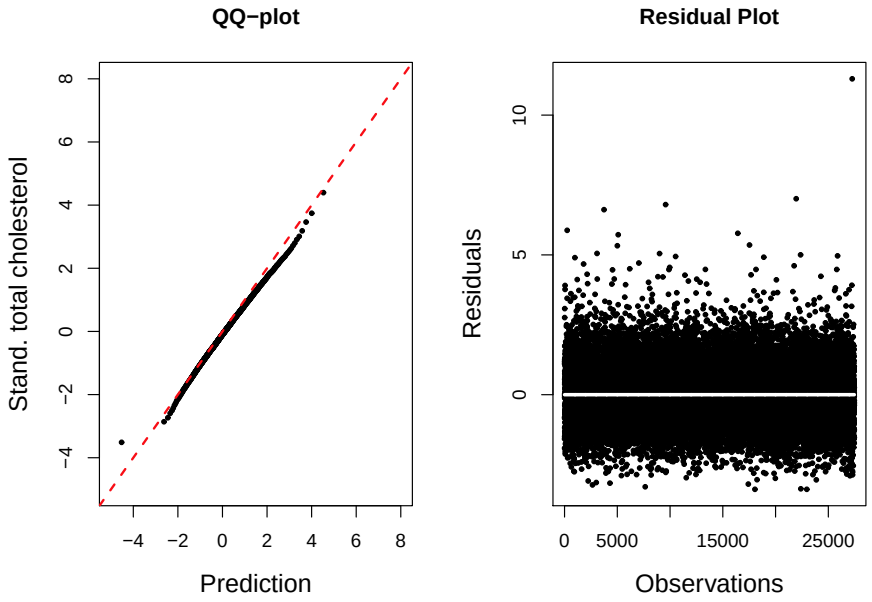


Table 4.1: Simulation Result

MEV.	Para-meter	ICM method: 90% quantile Bias	SE.	RMSE.	80% quantile Bias	SE.	RMSE.	70% quantile Bias	SE.	RMSE.	Our method: Bias	SE.	SEE.	CP.
						n=200								
σ^2 = 0.25	μ = 1.0	–	–	–	–	–	–	–	–	–	0.023	0.063	0.063	0.951
	β_1 = 0.3	-0.001	0.379	2.39	-0.006	0.274	1.73	-0.006	0.227	1.43	0.008	0.159	0.158	0.957
	β_2 = 0.3	-0.012	0.582	2.36	-0.021	0.429	1.74	-0.033	0.362	1.47	-0.001	0.246	0.246	0.959
σ^2 = 1.0	μ = 1.0	–	–	–	–	–	–	–	–	–	0.023	0.073	0.074	0.975
	β_1 = 0.3	-0.008	0.407	2.15	-0.032	0.274	1.44	-0.038	0.221	1.16	0.004	0.189	0.193	0.972
	β_2 = 0.3	-0.033	0.569	1.91	-0.041	0.410	1.38	-0.055	0.334	1.12	0.001	0.297	0.303	0.976
						n=400								
σ^2 = 0.25	μ = 1.0	–	–	–	–	–	–	–	–	–	0.012	0.044	0.045	0.963
	β_1 = 0.3	0.004	0.272	2.47	-0.003	0.196	1.78	-0.014	0.162	1.47	0.007	0.110	0.114	0.964
	β_2 = 0.3	-0.037	0.419	2.37	-0.025	0.302	1.71	-0.018	0.255	1.44	-0.002	0.176	0.177	0.950
σ^2 = 1.0	μ = 1.0	–	–	–	–	–	–	–	–	–	0.013	0.055	0.051	0.966
	β_1 = 0.3	-0.015	0.265	2.01	-0.034	0.178	1.35	-0.046	0.157	1.20	0.006	0.132	0.134	0.971
	β_2 = 0.3	-0.063	0.414	1.96	-0.047	0.308	1.46	-0.057	0.239	1.13	-0.007	0.211	0.209	0.962

MEV., measurement error variance; RMSE., ratio of mean square error,
SE., standard error; SSE., standard error estimate (median value); CP., 95 % coverate probability;
Ratio of MSE, mean squared error ratio of ICM method to our method; ρ, correlations between observed biomarker values;
When σ^2 = 0.25, the variance ratio=0.16, ρ=0.87; when σ^2 = 1.0, the variance ratio=0.6, ρ=0.62.

Table 4.2: Application to the ARIC Study Data

	ICM method			Our method		
	Estimate	SE*.	p-value	Estimate	SE**.	p-value
Threshold effect	-	-	-	1.576	0.0218	<0.0001
Race=African-Americans	0.409	0.096	<0.0001	0.130	0.0269	<0.0001
Gender=male	0.160	0.180	0.3729	0.447	0.0353	<0.0001
Hypertension	-0.279	0.134	0.0369	-0.022	0.0238	0.3458
History of Parents' CHD	-0.078	0.097	0.4251	0.274	0.0215	<0.0001
Age (50-59 yrs) at visit 1	0.451	0.240	0.0608	0.728	0.0293	<0.0001
Age (≥60 yrs) at visit 1	0.634	0.237	0.0076	0.714	0.0352	<0.0001
Male× age (50-59 yrs)	-0.562	0.238	0.0183	-0.714	0.0465	<0.0001
Male× age (≥60 yrs)	-0.938	0.221	<0.0001	-0.802	0.0552	<0.0001
Total cholesterol at visit 1 (mg/dL)	0.017	0.005	0.0003	0.025	0.0008	<0.0001

SE*, simple bootstrap-based standard error;
SE**, adjusted subsample bootstrap-based standard error;
measurement error variance = 0.3^2, $n = 10,236$, observations = $27,467$

CHAPTER5: WEIGHTED PSEUDO-LIKELIHOOD FOR ADJUSTING INFORMATIVE DIAGNOSIS: AN APPLICATION TO TIME-TO-HYPERCHOLESTEROLEMIA IN THE ARIC STUDY

5.1 Introduction

The Atherosclerosis Risk in Communities (ARIC) study was designed to investigate the causes of atherosclerosis. Hypercholesterolemia is a crucial risk factor for cardiovascular disease; hence, assessing risk factors associated with time-to- hypercholesterolemia also is of interest. Total cholesterol was measured at each clinic visit to determine hypercholesterolemia status. However, total cholesterol is subject to measurement error and even if an accurate total cholesterol level can be taken, the most appropriate threshold value to determine hypercholesterolemia may vary from patient to patient. In our previous work provided in Chapter 3 and 4, we introduced a threshold-dependent proportional hazards model to study the association between risk factors and time to hypercholesterolemia while accounting for measurement errors in the total cholesterol level.

The time-to-hypercholesterolemia data from the ARIC study poses another difficulty to statistical analysis: some subjects were diagnosed with hypercholesterolemia outside of this study (we call this "externally diagnosed"). After being diagnosed, these subjects may have started therapeutic treatment for hypercholesterolemia. As a result, their total cholesterol levels at subsequent visits may be lower than if they had not been diagnosed with hypercholesterolemia Figure 5.1 illustrates that the mean trend of total cholesterol levels in the subpopulation with external diagnosis and complete cases drops

off over time, and it has the similar mean to the mean total cholesterol level at visit 4 in the subpopulation with no external diagnosis and complete cases. In this study, there were 1,546 (13.2%) externally diagnosed subjects out of 11,718 subjects satisfying inclusion and exclusion criteria. Table 5.1 summarizes the number of subjects with externally diagnosed hypercholesterolemia by each visit.

Our previous analysis in Section 4.5 did not include these subjects because the total cholesterol level after being externally diagnosed was potentially attenuated. So, the conclusions from our previous study only apply to the subjects who were not externally diagnosed. However, the externally diagnosed subjects were likely to be high-risk for the disease. Therefore, it is prudent to incorporate them into the analysis. We can regard it as informative missing data problem.

Horvitz and Thompson (1952) proposed a method for survey data analysis accounting for different proportions of observations within strata by using inverse probability weights (IPW), which are the inverse of the inclusion probability in sampling data analysis, and then the method can be applied to the missing data problem. Rotnitzky and Robins (1995), Robins and Rotnitzky (1995), and Robins, Rotnitzky, and Zhao (1994; 1995) proposed weighted estimating equations in a regression setting using inverse probability weighting when data are missing at random (MAR). Robbins and Ritov (1997) showed the estimator is doubly robust, that is, it remains consistent when either a model for the missingness mechanism or the score vector for the missing data given the observed data is correctly specified. Rotnitzky et al. (1998) extended this method to address non-ignorable nonresponse in either the covariates or the outcomes by using augmented orthogonal inverse probability weighting. Breslow and Wellner (2006) considered a weighted likelihood estimator for semiparametric models with data from complex probability samples, and Li et al. (2008) proposed a weighted likelihood method for grouped interval censored data in case-cohort studies.

In Section 5.2, we extent the pseudo-likelihood approach proposed in Chapter 4 to a weighted pseudo-likelihood estimator to account for the sampling bias using inverse probability weighting, and also present the inference procedures using a weighted pseudo-likelihood EM algorithm for parameter estimation. Variance estimation is also proposed in Section 5.2. Asymptotic results for the proposed method are established in Section 5.3. We examine the numerical performance of the method for finite samples through a simulation study in Section 5.4. Application to the data of the ARIC study is presented in Section 5.5, and a discussion of the proposed method and related future work are in Section 5.6.

5.2 Method

5.2.1 Weighted Pseudo-Likelihood

As in our previous work in Chapter 4, we apply the threshold dependent model considering measurement error for longitudinal data to obtain the distribution of observed biomarker values. However, the proportion excluded may be non-ignorable (about 13.18 %) and it probably leads to selection bias. This issue can be viewed as informative missingness. To account for the selection bias, we weight observations by the inverse probability that subjects are not externally diagnosed in the previous estimation approach (Robins et al. 1995). We restrict attention to monotone missing data patterns because participants would control their total cholesterol values after the external diagnosis. We also assume that the external diagnosis occurs at random because the presence or absence of the external diagnosis depends on the previous (observed) total cholesterol value and is independent of the future (unobserved) total cholesterol values.

For subject i, let $Y_i^*(t)$ and $Y_i(t)$ be the true biomarker value and observed biomarker value at continuous time $t \geq 0$, respectively. We assume $Y_i^*(t)$ is non-decreasing over

the study period, which is plausible for many chronic diseases such as diabetes, hypertension, and hypercholesterolemia as they are usually irreversible for aged patients without additional medication or treatment. Let $\boldsymbol{X}_i$ be time-invariant covariates in regards to potential risk factors associated disease and $\boldsymbol{Z}_i(t)$ be time-varying auxiliary information related to subject i's health status and eventually the external disease diagnosis. Then denote $(\boldsymbol{Z}_i(t)^T, Y_i(t))$ as $\boldsymbol{A}_i(t)$ and define the accumulated information $\{\boldsymbol{Z}_i(s)^T, Y_i(s)\}_{s>0}^t$ as $\overline{\boldsymbol{A}}(t)$. Let $N_i(t) = 1$ if subject i has a pre-scheduled visit at time t, $N_i(t) = 0$, otherwise. We assume that the visit time V_i of subject i is independent of $Y_i(V_i)$ given $\boldsymbol{X}_i$. The visit time V_i for subject i is practically discrete: $(v_{i0}, v_{i1}, \ldots, v_{in_i})$, where $v_{i0} = 0$, baseline visit. Define $R_{ij} = 1$ if subject i has no external diagnosis as hypercholesterolemia at jth follow-up visit, and this implies $R_{i(j-1)} = 1$ for $j \geq 1$. We simplify the notation of $\overline{A}(v_{ij})$ to $\overline{A}_{ij}$. Note that $R_{i0} = 1$ because of the inclusion criteria. The observed data from n i.i.d subjects are $\{\boldsymbol{X}_i, Y_i(v_{ij}), Z_i(v_{ij}), N_i(v_{ij}), v_{ij}, R_{ij} \mid i = 1, \ldots, n, j = 0, ..., n_i\}$, and is abbreviated as $\{\aleph_i \mid i = 1, \ldots, n\}$ hereafter.

We assume that there exists for $j \geq 1$, a known function of unknown parameter(s) $\boldsymbol{\alpha}_0$ and $\overline{\boldsymbol{A}}_{i(j-1)}$, $\pi_{ij}(\boldsymbol{\alpha})$ taking values $(0, 1]$ such that

$$\pi_{ij}(\boldsymbol{\alpha}) = P\left[R_{ij} = 1 \mid R_{i(j-1)} = 1, \boldsymbol{X}_i, \overline{\boldsymbol{A}}_{i(j-1)}\right]. \tag{5.1}$$

Typically, a logistic function would be chosen for $\pi_{ij}(\boldsymbol{\alpha})$. The probability of no external diagnosis by the jth follow-up visit is $\pi_{i1}(\boldsymbol{\alpha}) \times \cdots \times \pi_{ij}(\boldsymbol{\alpha})$ denoted by $\overline{\pi}_{ij}(\boldsymbol{\alpha})$. The observed partial likelihood for $\{\pi_{ij}(\boldsymbol{\alpha})\}$ is then

$$L_n(\boldsymbol{\alpha}) = \prod_{i}^{n} \prod_{j=1}^{n_i} \left(\pi_{ij}(\boldsymbol{\alpha})^{R_{ij}} \left[1 - \pi_{ij}(\boldsymbol{\alpha})\right]^{1-R_{ij}}\right)^{R_{i(j-1)}}. \tag{5.2}$$

Assuming that the missing model is correct, we add the inverse probability weights

to the observed pseudo-likelihood function:

$$l_n^{wps}(\boldsymbol{\theta}, \Lambda, \boldsymbol{\alpha} \mid \aleph_i) = \sum_{i=1}^{n}\sum_{j=1}^{n_i} R_{ij}\overline{\pi}_{ij}(\boldsymbol{\alpha})^{-1} \log \int_{-\infty}^{\infty} \exp\left\{-\Lambda(v_{ij})e^{\boldsymbol{\beta}^T \boldsymbol{X}_i - \mu\xi}\right\} \times \Lambda(v_{ij})\mu \exp(\boldsymbol{\beta}^T \boldsymbol{X}_i - \mu\xi)\frac{1}{\sigma}\phi\left\{\frac{Y_i(v_{ij}) - \xi}{\sigma}\right\} d\xi, \quad (5.3)$$

where $\boldsymbol{\theta} = (\mu, \boldsymbol{\beta}^T)^T$ and $\phi(\cdot)$ is the standard normal density function. The weight used in (5.3) can be interpreted: as the number of subjects the observation would represent.

5.2.2 Inference Procedure

To obtain the partial likelihood estimate $\widehat{\boldsymbol{\alpha}}$, we solve the following partial score equation, and the score function is derived through differentiation of the partial likelihood in (5.2) with respect to α,

$$S_n(\boldsymbol{\alpha}) = \partial \log L_n(\alpha)/\partial\alpha = \sum_{i=1}^{n}\sum_{j=1}^{n_i}\{R_{ij} - R_{i(j-1)}\pi_{ij}(\boldsymbol{\alpha})\}\{\partial \text{logit}\left\{\pi_{ij}(\boldsymbol{\alpha})\right\}/\partial\alpha\} = 0. \quad (5.4)$$

Note that $S_n(\boldsymbol{\alpha})$ simplifies to $\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left\{R_{ij} - R_{i(j-1)}\pi_{ij}(\boldsymbol{\alpha})\right\} g\left\{\overline{\boldsymbol{A}}_{i(j-1)}\right\}$ if $\pi_{ij}(\boldsymbol{\alpha})$ follows the logistic regression model, $\text{logit}\{\pi_{ij}(\boldsymbol{\alpha})\} = \boldsymbol{\alpha}^T g\left\{\overline{\boldsymbol{A}}_{i(j-1)}\right\}$ for some known vector function $g(\cdot)$. The estimate $\widehat{\boldsymbol{\alpha}}$ is incorporated in the weighted pseudo-likelihood.

We consider the log complete pseudo-likelihood with weights. Specifically, we estimate $\Lambda(t)$ as a step function with jumps at the observed times. Let $v_{(1)} < \ldots < v_{(K)}$ be uniquely ordered observed times of $\{v_{ij} \mid i = 1, \ldots, n, j = 1, \ldots, n_i\}$ and $\Lambda_k = \Lambda_0(V_{(k)})$ and $V_{(0)} = 0$.

To facilitate the maximization, we adopt the weighted-pseudo-EM algorithm by

treating threshold values ξ as missing data. Then the weighted complete log pseudo-likelihood function is

$$l_c^{wps}\{\theta, \boldsymbol{\alpha}, \Lambda\} = \sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{j=1}^{n_i} \frac{I(v_{ij} = v_{(k)})R_{ij}}{\overline{\pi}_{ij}\{\widehat{\boldsymbol{\alpha}}\}}\Big[-\Lambda_k \exp(\boldsymbol{\beta}^T \boldsymbol{X}_i - \mu\xi_{ij}) + \log\Lambda_k + \log\mu + \boldsymbol{\beta}^T\boldsymbol{X}_i - \mu\xi_{ij} - \frac{1}{2}\log\sigma^2 - \frac{\{Y_i(v_{ij}) - \xi_{ij}\}^2}{2\sigma^2}\Big]. \quad (5.5)$$

In the maximization step of the lth iteration of the EM algorithm, we first maximize the conditional expectation of the weighted complete log pseudo-likelihood function given observed data over Λ_k's. We then update $\boldsymbol{\theta}$ via the Newton-Raphson algorithm. Specifically, we maximize $Q(\Lambda)$ defined by

$$Q_w(\Lambda) = \sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{j=1}^{n_i} \frac{I(V_{ij} = V_{(k)})R_{ij}}{\overline{\pi}_{ij}\{\widehat{\boldsymbol{\alpha}}\}} E\left\{-\Lambda_k \exp(\boldsymbol{\beta}^T\boldsymbol{X}_i - \mu\xi_{ij}) + \log\Lambda_k \mid \boldsymbol{\aleph}_i, \boldsymbol{\theta}^{(l)}\right\}. \quad (5.6)$$

Since $Q_w(\Lambda)$ is a concave function over a convex cone satisfying $\Lambda_1 \le \ldots \le \Lambda_K$, this maximization can be carried out using one of the many existing algorithms for convex optimization. To update $\boldsymbol{\theta}$, we apply the following one-step Newton-Raphson algorithm, $\boldsymbol{\theta}^{(l+1)} = \boldsymbol{\theta}^{(l)} + E\left(-\partial^2 l_c^{wps}/(\partial\theta)^2 \mid \boldsymbol{\aleph}(t), \boldsymbol{\theta}^{(l)}\right)^{-1}_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(l)}} E\left(\partial l_c^{wps}/\partial\theta \mid \boldsymbol{\aleph}(t), \boldsymbol{\theta}^{(l)}\right)_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(l)}}$. The conditional expectations are calculated in the expectation step of the EM algorithm based on the following expression

$$E\{g(\xi)|\boldsymbol{\aleph}_i(V_{ij}), \boldsymbol{\theta}^{(l)}\} = \frac{I(v_{ij} = v_{(k)})\int_{-\infty}^{\infty} g(\xi)\exp(-\Lambda_k e^{\boldsymbol{\beta}^T\boldsymbol{X}_i - \mu\xi})e^{-\mu\xi}\phi\left\{\frac{Y_i(v_{ij})-\xi}{\sigma}\right\}d\xi}{\int_{-\infty}^{\infty}\exp(-\Lambda_k e^{\boldsymbol{\beta}^T\boldsymbol{X}_i-\mu\xi})e^{-\mu\xi}\phi\left\{\frac{Y_i(v_{ij})-\xi}{\sigma}\right\}d\xi}, \quad (5.7)$$

where the $g(\xi)$'s to be calculated are ξ, ξ^2, $e^{-\mu\xi}$, $e^{-\mu\xi}\xi$, and $e^{-\mu\xi}\xi^2$. This integration can be approximated by the Gauss-Hermite quadrature (Davis 1984):

$$\sum_{k=1}^{N}\left(\sqrt{2}\sigma w_k g\{\sqrt{2}\sigma w_k + Y_i(v_{ij})\}\exp\left[-\Lambda_0(v_{ij})e^{\boldsymbol{\beta}^T X_i - \mu\{\sqrt{2}\sigma z_k + Y_i(v_{ij})\}}\right]e^{-\mu\{\sqrt{2}\sigma z_k + Y_i(v_{ij})\}}\right),$$

where N is the number of the quadratures and ω_k and z_k are weights and abscissae for the Gauss-Hermite quadrature, respectively. This EM loop is repeated until $\left|\boldsymbol{\theta}^{(l+1)} - \boldsymbol{\theta}^{(l)}\right|$ is smaller than a pre-specified criterion. We denote the final estimators as $\widehat{\boldsymbol{\theta}} = (\widehat{\mu}, \widehat{\boldsymbol{\beta}}^T)^T$ and $\widehat{\Lambda}$.

5.2.3 Variance Estimation

To derive the weighted marginal influence functions leading to asymptotic variance estimation for the regression parameter $\boldsymbol{\theta}$, we need the complete log pseudo-likelihood function with the weights. For subject i, the weighted pseudo-score functions with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ are given by

$$\begin{aligned} l_c^{wps}\{\theta, \boldsymbol{\alpha}, \Lambda\} &= \int_0^\infty \frac{R_j}{\overline{\pi}_j\{\boldsymbol{\alpha}\}}\Big[-\Lambda(v)\exp(\boldsymbol{\beta}^T\boldsymbol{X} - \mu\xi) + \log\Lambda(v) + \log\mu \\ &\quad + \boldsymbol{\beta}^T\boldsymbol{X} - \mu\xi - \frac{1}{2}\log\sigma^2 - \frac{\{Y(v)-\xi\}^2}{2\sigma^2}\Big] dN(v), \qquad (5.8) \\ \dot{l}_\mu^{wps}\{\boldsymbol{\theta}, \boldsymbol{\alpha}, \Lambda\} &= E_\xi\{\partial l_c^{wps}/\partial\mu \mid \aleph\} \\ &= \int_0^\infty \frac{R_j}{\overline{\pi}_j\{\boldsymbol{\alpha}\}}\left[\mu^{-1} - E_\xi\left\{\kappa(v)\xi \mid \aleph\right\}\right] dN(v), \qquad (5.9) \end{aligned}$$

$$\begin{aligned} \dot{l}_\beta^{wps}\{\boldsymbol{\theta}, \boldsymbol{\alpha}, \Lambda\} &= E_\xi\{\partial l_c^{wps}/\partial\beta \mid \aleph\} \\ &= \boldsymbol{X}\int_0^\infty \frac{R_j}{\overline{\pi}_j\{\boldsymbol{\alpha}\}} E_\xi\left\{\kappa_i(v) \mid \aleph\right\} dN(v), \qquad (5.10) \end{aligned}$$

where $\kappa(v) = 1 - \Lambda(v)\exp(\boldsymbol{\beta}^T\boldsymbol{X} - \mu\xi)$ and $N(v)$ denotes the counting process associated with prescheduled visiting time, so $\int_{0\le s\le v} dN(s) = j$,that is, the number of occasions. The random variable of R_j and the parameter of π_j follow the same definitions of the subject-specific variable and parameter R_{ij} and π_{ij}, respectively.

Let $\{P_{\boldsymbol{\theta},\Lambda_\eta}\}$ be a regular parametric subfamily of models, $\{P_{\boldsymbol{\theta},\Lambda} \mid P_{\boldsymbol{\theta},\Lambda} \ll m, m$ is Lebesque measure$\}$ and set $\partial/\partial\eta|_{\eta=0}\Lambda_\eta(v)=h(v)$ for $v > 0$ and $h(v) \in L_2(P_V)$. Then the

score operator for Λ is

$$\dot{l}^{wps}_{\Lambda}\{\boldsymbol{\theta},\boldsymbol{\alpha},\Lambda\}[h(v)] = \int_0^{\infty} \frac{R_j}{\overline{\pi}_j\{\boldsymbol{\alpha}\}} h(v) E(\kappa(v) \mid \boldsymbol{\aleph})/\Lambda(v) dN(v). \tag{5.11}$$

We define $h^*_{\mu}(v)$ and $h^*_{\beta}(v)$:

$$\begin{aligned}
h_1(v) &= E\Big\{W_1(v)e^{\boldsymbol{\beta}^T\boldsymbol{X}}\Big(E\{\xi \mid \boldsymbol{\aleph}\}E\{e^{-\mu\xi} \mid \boldsymbol{\aleph}\} - 2E\{\xi e^{-\mu\xi} \mid \boldsymbol{\aleph}\} \\
&\quad -e^{\boldsymbol{\beta}^T\boldsymbol{X}}\Lambda(v)\big[E\{\xi e^{-\mu\xi} \mid \boldsymbol{\aleph}\}E\{e^{-\mu\xi} \mid \boldsymbol{\aleph}\} - E\{\xi e^{-2\mu\xi} \mid \boldsymbol{\aleph}\}\big]\Big) \mid V = v\Big\}, \\
h_2(v) &= E\Big\{W_1(v)\boldsymbol{X}e^{\boldsymbol{\beta}^T\boldsymbol{X}}\Big(\Lambda(v)e^{\boldsymbol{\beta}^T\boldsymbol{X}}\big[E\{e^{-\mu\xi} \mid \boldsymbol{\aleph}\}^2 - E\{e^{-2\mu\xi} \mid \boldsymbol{\aleph}\}\big] + E\{e^{-\mu\xi} \mid \boldsymbol{\aleph}\}\Big) \mid V = v\Big\}, \\
h_3(v) &= E\Big(W_1(v)\Lambda(v)^{-2} + W_1(v)e^{2\boldsymbol{\beta}^T\boldsymbol{X}}\big[E\{e^{-\mu\xi} \mid \boldsymbol{\aleph}\}^2 - E\{e^{-2\mu\xi} \mid \boldsymbol{\aleph}\}\big] \mid V = v\Big), \\
h^*_{\mu}(v) &= h_1(v)/h_3(v), \\
h^*_{\beta}(v) &= h_2(v)/h_3(v),
\end{aligned}$$

and $W_1(v)$=$R_j/\overline{\pi}_j(\boldsymbol{\alpha})$.

In the asymptotic proofs given later, we show that the asymptotic covariance of $\widehat{\boldsymbol{\theta}}$ has a sandwich variance form. For the sandwich variance form, we need to summarize notations. Denote $P\{\ddot{l}^{wps}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\theta}_0,\boldsymbol{\alpha}_0,\Lambda_0) - \ddot{l}^{wps}_{\Lambda\boldsymbol{\theta}}(\boldsymbol{\theta}_0,\boldsymbol{\alpha}_0,\Lambda_0)[h^*_{\theta}]\}$ as $\boldsymbol{D}_w$, and

$$\begin{aligned}
\dot{l}^{wps}_{\mu\alpha}\{\boldsymbol{\theta},\boldsymbol{\alpha},\Lambda\} &= \int_{v=0}^{\infty} W_2(v)\big[\mu^{-1} - E\{\kappa(v)\xi \mid \boldsymbol{\aleph}\}\big] dN(v), \\
\dot{l}^{wps}_{\beta\alpha}\{\boldsymbol{\theta},\boldsymbol{\alpha},\Lambda\} &= \int_{v=0}^{\infty} \boldsymbol{X}W_2(v)E\{\kappa(v) \mid \boldsymbol{\aleph}\} dN(v), \\
\dot{l}^{wps}_{\Lambda\alpha}\{\boldsymbol{\theta},\boldsymbol{\alpha},\Lambda\}[h(v)] &= \int_{v=0}^{\infty} W_2(v)h(v)E\{\kappa(v) \mid \boldsymbol{\aleph}\}/\Lambda(v) dN(v),
\end{aligned}$$

where $W_2(v) = -R_j\overline{\pi}_j\{\boldsymbol{\alpha}\}^{-2}\partial\overline{\pi}_v(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}$, and $\int_{0\le s\le v} dN(s) = j$. The partial score function for α is

$$S(\boldsymbol{\alpha}) = \partial \log L(\alpha)/\partial\alpha = \int_0^{\infty} \{R_j - R_{(j-1)}\pi_j(\boldsymbol{\alpha})\}\{\partial \text{logit}\{\pi_j(\boldsymbol{\alpha})\}/\partial\alpha\} dN(t).$$

Then if $\boldsymbol{D}_w$ is invertible, the information matrix for $\widehat{\boldsymbol{\theta}}$ is

$$\begin{aligned} I_w &= \boldsymbol{D}_w^{-1} P\left\{(\boldsymbol{M}_\theta + \boldsymbol{M}_\alpha)(\boldsymbol{M}_\theta + \boldsymbol{M}_\alpha)^T\right\}\left\{\boldsymbol{D}_w^{-1}\right\}^T \\ &= \boldsymbol{D}_w^{-1} P\left\{(\boldsymbol{M}_\theta \boldsymbol{M}_\theta^T + 2\boldsymbol{M}_\theta \boldsymbol{M}_\alpha^T + \boldsymbol{M}_\alpha \boldsymbol{M}_\alpha^T\right\}\left\{\boldsymbol{D}_w^{-1}\right\}^T, \end{aligned} \tag{5.12}$$

where $\boldsymbol{M}_\alpha = P\left\{\ddot{l}^{wps}_{\boldsymbol{\theta}\boldsymbol{\alpha}}(\boldsymbol{\theta}_0, \boldsymbol{\alpha}_0, \Lambda_0) - \ddot{l}^{wps}_{\Lambda\boldsymbol{\alpha}}(\boldsymbol{\theta}_0, \boldsymbol{\alpha}_0, \Lambda_0)[h^*_\theta]\right\} P\left\{-\partial^2 \log L/(\partial\boldsymbol{\alpha})^2\right\}^{-1} S(\boldsymbol{\alpha}_0)$ and $\boldsymbol{M}_\theta = \dot{l}^{wps}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \boldsymbol{\alpha}_0, \Lambda_0) - \dot{l}^{wps}_{\Lambda}(\boldsymbol{\theta}_0, \boldsymbol{\alpha}_0, \Lambda_0)[h^*_\theta]$.

We estimate the information matrix in (5.12) using the estimated parameters $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\Lambda})$ and empirical measure $\mathbb{P}_n = n^{-1}\sum_{i=1}^n$ instead of P.

5.3 Asymptotic Results

In this section, we provide asymptotic results for the proposed estimators and the proofs are summarized in Appendix C. Let $\boldsymbol{\theta}_0$ and $\boldsymbol{\alpha}_0$ denote the true regression parameters, and let Λ_0 denote the cumulative hazard function. The asymptotic results will use the following conditions:

(A1) The finite-dimensional parameter spaces $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$ are compact subsets of the domains of $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$, respectively.

(A2) The covariate $\boldsymbol{X}$ has bounded support with probability 1. If $\boldsymbol{\beta}^T\boldsymbol{X} + \alpha = 0$ almost surely (a.s.), then $\boldsymbol{\beta} = 0$ and $\alpha = 0$.

(A3) The support of the visit time, V, is an interval $\mathscr{S}[V] = [l_V, u_V]$, with $0 < l_V \leq u_V < \infty$.

(A4) The number of the visit times, $\int_0^{u_V} dN(V)$ is P_V-almost surely finite.

(A5) The cumulative hazard function Λ_0 has strictly positive derivative on $\mathscr{S}[V]$.

(A6) The auxiliary information $\boldsymbol{A}(t)$ has bounded support with probability 1. If $\boldsymbol{\alpha}^T g\{\overline{\boldsymbol{A}}(t)\} + \gamma = 0$ almost surely (a.s.), then $\boldsymbol{\alpha} = 0$ and $\gamma = 0$, where $g(\cdot)$ is a known bounded function.

(A7) $\pi_V(\alpha) \geq c > 0$ for all $\alpha \in \boldsymbol{\Theta}_2$ and some constant c.

(A8) The link function for $\pi_V(\boldsymbol{\alpha})$ in the generalized linear model, $H(\cdot)$ is one-to-one. There exists a measurable function $m(\cdot)$ such that $|H^{-1}(\boldsymbol{\alpha}_1^T g\{\overline{\boldsymbol{A}}(t)\}) - H^{-1}(\boldsymbol{\alpha}_2^T g\{\overline{\boldsymbol{A}}(t)\})| \leq m\{\overline{\boldsymbol{A}}(t)\}|\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2|$ and $P|m|^2 < \infty$.

The assumptions that parameter, covariate, and visit time are bounded in $(A1)$, $(A2)$, and $(A3)$, respectively are standard. Condition (A2) and (A6) ensure the identifiability of $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$, respectively. Condition (A7) ensures that each subject i, the probability of no external diagnosis during the follow-up period is bounded away from zero, so the inverse weight is bounded. Condition (A8) means that the derivative of the inverse function of the link function is uniformly bounded by the measurable function, $m(\cdot)$. Link functions commonly used in generalized linear models satisfy Condition (A8) on $\boldsymbol{\Theta}_2$, particularly the logit function. These conditions hold naturally in most applications.

For convergence of the estimators to the true parameters, we need to define a topology. Let the bounded regression parameter space $\boldsymbol{\Theta}_1 \times \boldsymbol{\Theta}_2 (\subset R^{d_1} \times R^{d_2})$ be equipped with the Euclidean topology. Regarding the infinite dimensional nonparametric space, let $\mathscr{F}$ be the set of all Borel subprobability measures on $\mathscr{S}[V]$. Then $\mathscr{F}$ can be equipped with the vague topology by defining that, for any sequence $F_n \in \mathscr{F}$ and $F \in \mathscr{F}$, F_n converges vaguely to F if and only if

$$\int f dF_n \to \int f dF \quad \text{for every} \quad f \in C_0(\mathscr{S}[V]),$$

where $C_0(\mathscr{S}[V])$ is the set of all continuous functions that vanish outside a compact subset of $\mathscr{S}[V]$. The product space $\boldsymbol{\Theta}_1 \times \boldsymbol{\Theta}_2 \times \mathscr{F}$ equipped with the product topology is equivalent to convergence of $(\widehat{\boldsymbol{\theta}}^T, \widehat{\boldsymbol{\alpha}}^T, \widehat{F})$ to $(\boldsymbol{\theta}^T, \boldsymbol{\alpha}^T, F)$ in those respective domains.

Theorem 5.3.1. *(Consistency of the MLE) Suppose that conditions, (A1), (A2), (A3), (A4), (A6), (A7), and (A8) are satisfied, then* $\widehat{\boldsymbol{\theta}}$ *converges to* $\boldsymbol{\theta}_0$ *a.s., and if* $v \in \mathscr{S}[V]$ *is a continuity point of* Λ_0, $\widehat{\Lambda}(v)$ *converges to* $\Lambda_0(v)$ *a.s. Moreover, if* Λ_0 *is continuous, then* $\sup_{v\in\mathscr{S}[V]}|\widehat{\Lambda}(v) - \Lambda_0(v)|$ *converges to 0 a.s.*

Before discussing the overall convergence rate, we define the distance d on $R^{d_1} \times R^{d_2} \times \Phi$ as follows: $d\{(\boldsymbol{\theta}_1, \boldsymbol{\alpha}_1, \Lambda_1), (\boldsymbol{\theta}_2, \boldsymbol{\alpha}_2, \Lambda_2)\} = |\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2| + |\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2| + \|\Lambda_1 - \Lambda_2\|_{2,P}$, where $|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2|$ and $|\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2|$ are the Euclidean distance in R^{d_1} and R^{d_2}, $\|\Lambda_1 - \Lambda_2\|_{2,P_V} = \left(\int (\Lambda_1(v) - \Lambda_2(v))^2 dP_V\right)^{1/2}$, where P_V is the marginal probability measure of the measurement time variable V.

Theorem 5.3.2. *(Rate of convergence) Suppose that conditions (A1), (A2), (A3), (A4), and (A7) are satisfied. Then* $d\left\{(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\Lambda}), (\boldsymbol{\theta}_0, \boldsymbol{\alpha}_0, \Lambda_0)\right\} = O_p(n^{-1/3})$.

The convergence rate we found is applied to prove the asymptotic normality of the regression parameter MLE, $\widehat{\boldsymbol{\theta}}$.

Theorem 5.3.3. *(Asymptotic normality) Suppose that* θ_0 *is an interior point of* Θ *and that conditions* $(A1)$–$(A8)$ *are satisfied. Then*

$$n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = -n^{1/2}(\mathbb{P}_n - P)\widetilde{\psi}^{wps}(\boldsymbol{\aleph}) + o_p(1) \to N(0, I_w(\theta_0)) \quad \textit{in distribution,}$$

where P *is the probability measure, that is,* $P\widetilde{\psi}^{wps}(\boldsymbol{\aleph}) = \int \widetilde{\psi}^{wps}(\boldsymbol{\aleph}) dP$, $\mathbb{P}_n$ *is the empirical measure of* $\boldsymbol{\aleph}_i$, $i = 1, \ldots, n$, *that is,* $\mathbb{P}_n\widetilde{\psi}^{wps}(\boldsymbol{\aleph}) = n^{-1}\sum_{i=1}^{n}\widetilde{\psi}^{wps}(\boldsymbol{\aleph})$, $\widetilde{\psi}^{wps}(\boldsymbol{\aleph})$ *is the marginal influence function defined as* $\widetilde{\psi}^{wps} = \boldsymbol{D}_w^{-1}\{\boldsymbol{M}_\theta + \boldsymbol{M}_\alpha\}$, *and* $I_w(\theta_0)$ *is the information in (*5.12*).*

Theorem 5.3.4. *(Consistency of the asymptotic variance estimator) When the bandwidth h_n satisfies that h_n and $\log\{n/(nh_n)\}$ converge to 0 as $n \to \infty$, then $\widehat{I}_w$ converges to $I(\boldsymbol{\theta}_0)$ in probability.*

5.4 Simulation Study

We consider scenarios of longitudinal data with random measurement time points. The number of measurement times per subject is three, baseline and two follow-ups, and the measurement times are independently generated from the normal distribution with means of 0.5, 1.2, and 1.9, respectively and with the common standard deviation of 0.1 since study visit window is usually fixed, but each visit time varies across subjects. Two covariates are included in the model: one is generated from the Bernoulli distribution with probability 0.5, and the other is from the normal distribution with mean 0 and variance 0.1. The true values for (β_1, β_2) are set as (0.3, 0.3), and the true cumulative baseline hazard is assumed to be $2t^{1/5}$.

Consequently, the true biomarker value is generated by:

$$Y_i^*(v_{ij}) \quad = \quad \mu^{-1}\left\{\boldsymbol{\beta}^T \boldsymbol{X}_i + \log \Lambda_0(v_{ij}) - \log\left(-\log p_i\right)\right\} \text{ for } 1 \le i \le n, 0 \le j \le 2, \quad (5.13)$$

where p_i are independent draws from the uniform distribution (0,1) for each i.

The observed biomarker value is generated by

$$Y_i(v_{ij}) = Y_i^*(v_{ij}) + \epsilon_i(v_{ij}), \quad (5.14)$$

where $\epsilon(v_{ij})$ is independently generated from the normal distribution with zero mean and some finite variance for all i and j and is independent of $Y_i^*(v_{ij})$. We consider two measurement error variances of 0.25 and 1.0 and two sample sizes from 300 (600 follow-ups) to 600 (1,200 follow-ups. For each combination we conducted 1,000 repetitions.

We apply the following logistic regression model for the missing rule with the missing rate of 13% based on the previous biomarker values: logit{Pr(observable at visit j)}=3.0-0.25$Y_i(v_{i(j-1)})$. That is to say we calculate the probability of non-missingness, $\overline{\pi}_{ij}$ for each subject i at measurement times j=1 and 2. We then we generate a uniform [0,1] random variable for each observation, u_{ij}, and include observations at visit j when $u_{ij} < \overline{\pi}_{ij}$. If subject i has missing value at visit j, then the following observations are set to be missing, which simulates a monotone missing pattern.

For each simulated dataset, the proposed weighted-pseudo-EM algorithm was used to estimate the parameters accounting for the predicted probability of non-missingness. The initial values used for β and $\Lambda_0(t)$ in the algorithm were 0's and observed times, respectively. In the maximization-step, the spectral projected gradient method was used for the constrained optimization in (3.5). The convergence criterion for the algorithm was set as 10^{-10}. In the simulations, we noticed that threshold effect of μ was sensitive to the initial values. Therefore, we first calculated the profile likelihood of μ using the same algorithm except that μ was held at some fixed value. We then carried out a grid search to find the maximum likelihood estimate for μ.

Table 5.2 shows that bias of the regression coefficient estimates $\widehat{\boldsymbol{\beta}}$ is small, whereas the bias of the threshold effect estimate $\widehat{\mu}$ is relative large, but this is acceptable. The bias of the estimates decreases as the sample size increases or the variance ratio decreases. The estimated variance estimates (summarized by those median value) tend to overestimate the empirical variance. Out of 1,000 sets of the simulated data, the asymptotic variance estimate is likely to be overestimated in 10-15 % of datasets because of unstable cumulative hazard function estimates at the last observation times. Hence, when the asymptotic variance is calculated using the MLEs, we excluded 2% observations from the simulated data, which corresponds to the last observation times. Consequently, the coverage probability is greater than 95 %.

5.5 Application

The Atherosclerosis Risk in Communities (ARIC) study is a population-based cohort study from four U.S. communities, Forsyth County, NC, Jackson, MS, suburbs of Minneapolis, MN, and Washington County, MD, and participants underwent a baseline examination in 1987-1989, had three follow-up examinations at approximately three-year intervals, and a further examination in 2011-2013. The ARIC Study was designed to investigate the causes of atherosclerosis, and hypercholesterolemia is a crucial risk factor for cardiovascular disease. Hypercholesterolemia, that is, high blood cholesterol, is not a disease but a metabolic derangement that can be secondary to many diseases and contributes to many diseases, most notably cardiovascular disease. Hence, assessing risk factors associated with time-to-hypercholesterolemia is of interest. The participants were predominantly white or African-American: the few participants of other races are excluded from the analysis. Subjects with complete covariates and at least one valid follow-up visit data are included in the analysis

The time-to-hypercholesterolemia data from the ARIC study pose the informative missing data as we described in Section 5.1. In the preceding analysis, 1,546 (13.2%) out of 11,718 subjects satisfying inclusion and exclusion criteria were excluded due to an external diagnosis. To account for this informative and monotone missingness, we apply the weighted pseudo-likelihood approach to the ARIC data.

In the first step, we employ a logistic regression model for the external diagnosis outcome R_{ij} for $1 \leq i \leq n$ and $1 \leq j \leq 3$ to predict the probability for no external diagnosis based on the baseline and previous auxiliary information. We distinguish missing visits or drop-out with the external diagnosis and let $R_{ij} = .$ when subject i is missing or drop out at visit j. As predictors for the probability of no external diagnosis, we include sex, race, hypertension, previous coronary heart disease (CHD) history, parents CHD history, former smoking, high-density lipoprotein (HDL), and previous total cholesterol

level. When the previous visit is missing, then the last previous cholesterol value is carried forward to impute the missing previous cholesterol value. In addition, we take account of age effect, visit effect (categorical time lag effect), and interaction effect between previous total cholesterol level and visit.

By removing insignificant predictors, we reach the following final logistic regression model for the probability of no external diagnosis as hypercholesterolemic in table 5.3. Whites, people with hypertension, people with previous CHD history, people with parents CHD history, and former smoker are more likely to be externally diagnosed as hypercholesterolemia than people with the opposite characteristics. As previous cholesterol or HDL at baseline is higher, the probability of the external diagnosis is higher. The interquartile range of the weight obtained from the predicted probability for no external diagnosis is 1.02 to 1.12, and the weight greater than 99.9% quantile ranges from 9.25 to 40.77 with one extreme value of 361.15.

In the model, we consider baseline covariates including race, gender, hypertension, parents coronary heart disease (CHD) history, categorized age (<50,50-60, ≥60), and total cholesterol. These variables are generally regarded as major factors associated with hypercholesterolemia. All subjects with complete data for the baseline covariates are included in the analysis. Demographic characteristics of the subjects included in the analysis set include average age of 53.9 years (range 44-66 years), white race 9,166 (78.2%), and women 6,498 (55.5%). The average total cholesterol at baseline is 213.6 (±40.7) mg/dL, and the number of the participants with hypertension and parental history of CHD are 3,473 (29.6 %) and 4,719(40.3%), respectively.

To enhance estimations, we standardized the total cholesterol by the sample mean of $205.7 mg/dl$ and standard deviation of $37.3 mg/dl$ so that it has zero mean and the unit variance. The observation time is scaled to (0,1]. The standardized value and the rescaled visit time better facilitate the estimation process than either the original

values or log-transformed values.

The National Cholesterol Education Program and Laboratory Standardization Panel established the goal that a single serum total cholesterol measurement should be accurate within ±8.9 percent. The Health Care Financing Administration (HCFA) has also established similar testing requirements for total cholesterol (±10 percent), authored by the Centers for Disease Control and Prevention (Oppenheim et al. 1994). Hence, we chose $\sigma^2 = 0.3^2$ for the measurement error of the standardized total cholesterol value, which corresponds to 0.09 for the variance ratio of measurement error to total cholesterol value.

For comparison, we applied the unweighted pseudo-likelihood method to the subsample (11,718) with no external diagnosis or data before external diagnosis during the follow-up period. The variance estimate for effect size is somewhat sensitive to the choice of bandwidth, so we employed a subsampling bootstrap with sample size of 500 subjects and 400 repetitions then adjusted the standard error based on the bootstrap by multiplying the factors $\sqrt{500/11,718}$. In simulation data, the subsample bootstrap based standard error precisely estimates the true standard error. The bootstrap-based estimate and the adjusted standard error are presented in Table 5.4.

Overall, there is no marked difference between the effect sizes from the unweighted and weighted pseudo-likelihood method, and variance from the weighted pseudo-likelihood estimates is greater than variance of the unweighted estimates because of the variation due to the estimated weight. In the ARIC Study data, African-Americans, having parental history of CHD, and high baseline total cholesterol have 1.15, 1.32, and 1.02 times greater hazard of hypercholesterolemia than people with the opposite characteristics, respectively. When baseline total cholesterol level increases by 1 unit, the hazard for hypercholesterolemia increases by a factor of 0.024. There is significant interaction effect between age and sex; men is very likely to be high-risk for hypercholesterolemia

than women; up to 60-year old, hazard ratio for men to women in hypercholesterolemia is 1.67; however, it decreases to 1.51 after 60-year old because of menopause.

5.6 Concluding Remarks

We proposed a weighted pseudo-likelihood estimator based on a marginal semiparametric regression model for analyzing time-to-event of longitudinal biomarkers with missing data. We estimated the weight using a logistic regression model for the informative and monotone missing response. Parameter estimation was carried out by the weighted-pseudo-EM algorithm. The weighted estimator requires greater sample sizes to perform as well as the unweighted estimator for complete data because the weight estimation causes increased variation. The proposed method appears to be fairly accurate, but variance estimates are likely to be conservative. The method was illustrated through an application to data from the ARIC study.

The proposed model is based on the marginal likelihood method, so it is not guaranteed to satisfy semiparametric efficiency and is not doubly robust, that is, the estimator is consistent only when the model chosen for the missing data is correct. However, developing a doubly-robust estimator is a challenging open problem because of the score functions with respect to infinite dimensional nuisance parameters.

The proposed model can be extended in various ways: We could consider general missing mechanism rather than monotone missing patterns; When a covariance structure for the true biomarker values is postulated, a weighted semiparametric maximum likelihood method could be constructed.

Figure 5.1: Mean Trend of Total Cholesterol Levels in Subpopulation with Complete Follow-Ups

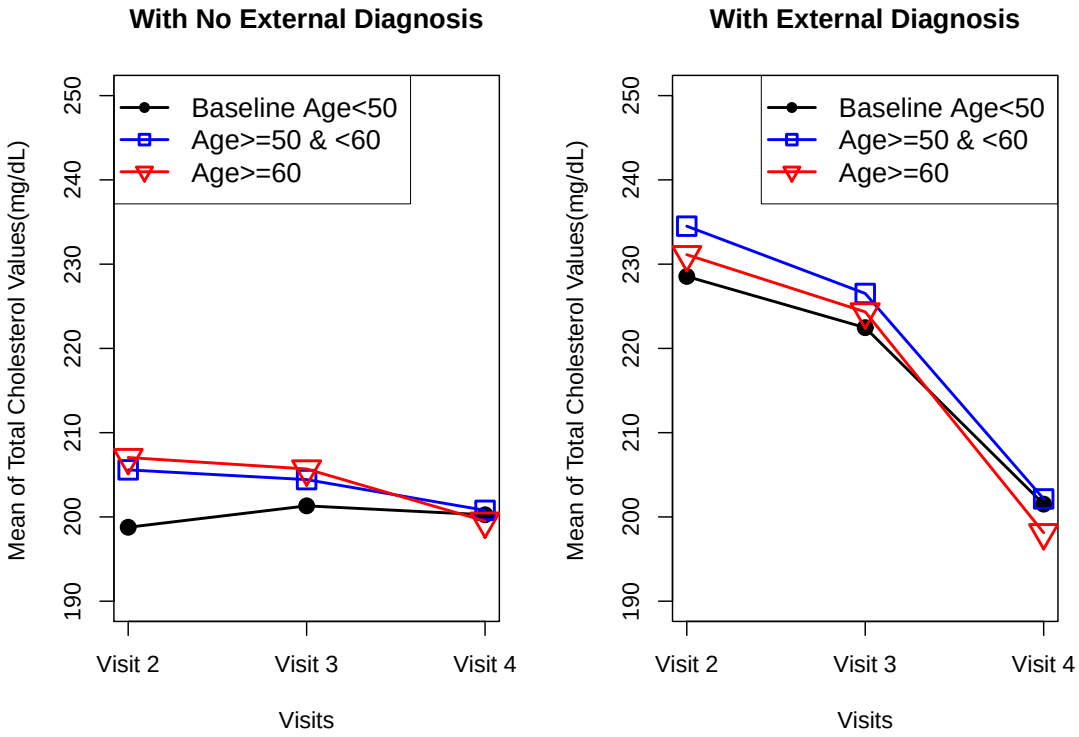


Table 5.1: Prevalence of Externally Diagnosed Hypercholesterolemia

Visit 2	Visit 3	Visit 4
n=11,699	n=10,487	n=9,571
507 (4.33%)	974 (9.25%)	1,546 (15.92%)

Table 5.2: Simulation Result When Missing Rate is 13%

Sample Size	Measurement Error Variance	Parameter	Bias	SE.	SEE. (median)	CP.
n=300	σ^2 = 0.25	μ = 1.0	0.042	0.055	0.061	0.962
		β_1 = 0.3	0.013	0.144	0.184	0.983
		β_2 = 0.3	0.011	0.215	0.249	0.978
n=300	σ^2 = 1.0	μ = 1.0	0.049	0.073	0.077	0.980
		β_1 = 0.3	0.010	0.175	0.289	0.983
		β_2 = 0.3	0.026	0.278	0.362	0.983
n=600	σ^2 = 0.25	μ = 1.0	0.033	0.040	0.042	0.966
		β_1 = 0.3	0.010	0.102	0.129	0.986
		β_2 = 0.3	0.006	0.156	0.175	0.975
n=600	σ^2 = 1.0	μ = 1.0	0.037	0.055	0.052	0.967
		β_1 = 0.3	0.008	0.121	0.210	0.984
		β_2 = 0.3	0.000	0.190	0.246	0.978

SE.: standard error; SEE. is standard error estimate; CP. is coverage probability.

Table 5.3: Logistic Regression for the Probability of No External Diagnosis as Hypercholesterolemia

Predictor	Estimate	SE.	p-value
Intercept	2.403	0.130	<0.0001
Male	-0.058	0.064	0.3670
African-Americans	0.773	0.087	<0.0001
Hypertension	-0.471	0.059	<0.0001
Previous CHD history	-1.088	0.104	<0.0001
Parents CHD history	-0.174	0.056	0.002
Former Smoking	-0.151	0.059	0.011
HDL (mg/dL)	0.026	0.002	<0.0001
Visit 4	-0.697	0.066	<0.0001
Previoius Cholesterl (40mg/dL) at Visit 2	-0.919	0.038	<0.0001
Previoius Cholesterl (40mg/dL) at Visit 3	-1.048	0.047	<0.0001
Previoius Cholesterl (40mg/dL) at Visit 4	-0.999	0.049	<0.0001

SE : Standard Error

Table 5.4: Application to the ARIC Study Data

	unweighted EM method			weighted EM method		
Risk Factors	Estimate	SE.	p-value	Estimate	SE.	p-value
Threshold effect	1.576	0.0204	<0.0001	1.573	0.0227	<0.0001
Race=African-Americans	0.137	0.0243	<0.0001	0.143	0.0372	0.0002
Gender=male	0.454	0.0330	<0.0001	0.514	0.0491	<0.0001
Hypertension	-0.013	0.0219	0.5528	-0.015	0.0319	0.6480
History of Parents' CHD	0.267	0.0201	<0.0001	0.277	0.0382	<0.0001
Age (50-59 yrs) at visit 1	0.725	0.0300	<0.0001	0.765	0.0475	<0.0001
Age (≥60 yrs) at visit 1	0.696	0.0341	<0.0001	0.758	0.0539	<0.0001
Male×Age (50-59 yrs) at visit 1	-0.737	0.0451	<0.0001	-0.766	0.0664	<0.0001
Male×Age (≥60 yrs) at visit 1	-0.819	0.0528	<0.0001	-0.862	0.0720	<0.0001
Total cholesterol at visit 1 (mg/dL)	0.025	0.0007	<0.0001	0.024	0.0008	<0.0001
Measurement error variance is 0.3^2 and n=11,718						

Table 5.5: Age Distribution by Whether or Not Being Externally Diagnosed with Hypercholesterolemia

External Diagnosis	Variable	Age (year) at Baseline			Total
		<50	≥50 and <60	≥60	(n=11,718)
No	Frequency	2,978	5,083	2,111	10,172
	Percentage	29.3%	50.0%	20.8%	
Yes	Frequency	325	838	383	1,546
	Percentage	21.0%	54.0%	24.8%	

CHAPTER6: SUMMARY AND FUTURE WORK

We had proposed statistical methods for several problems arising in longitudinal and observational studies when time-to-disease occurrence defined by biomarkers is the outcome variable of interest, namely, interval censored data, measurement error in biomarker values, determination of diagnostic cutoff point for biomarkers, and informative external diagnosis. First, we restricted attention to the first follow-up after baseline. Observed biomarker values were analyzed separately as true values and measurement error. An additive model was applied to account for biomarker values subject to measurement error by assuming that measurement error follows a zero-mean and finite variance Gaussian process and is independent of the true biomarker values. Assuming that the true underlying trend of biomarker values is non-decreasing over time, and observation time is independent of time-to-disease occurrence, we adopted generalized extreme value distributions to construct a stochastic model for the time-varying true biomarker values. Then we constructed the marginal observed likelihood for the observed biomarker values using a mixture of a normal distribution and a generalized extreme value distribution. This marginal model is equivalent to a class of proportional hazards models for threshold-dependent time-to-event. For the marginal likelihood, we considered all probabilities that disease occurs for each threshold and integrated over all of the information. By considering all possible threshold values we simultaneously resolved the problems of unobservable disease occurrence time and flexible threshold.

We thoroughly investigated this marginal model in Chapter 3 via simulation studies and asymptotic properties establishment. The marginal likelihood estimator for the

simulations was both accurate and semiparametrically efficient. The explicit asymptotic variance formula and estimation were presented. The variance estimates based on the observed information matrix approximated the true variance in finite samples well.

This marginal likelihood was extended to the pseudo-likelihood as provided in Chapter 4 by ignoring correlations between biomarker values within a subject. Compared to the marginal likelihood, the pseudo-likelihood estimator is more stable and accurate in data with considerable measurement error. In addition, we improved efficiency slightly as the correlation within a subject is reduced. For inference of the regression parameter estimates, we derived the variance formula using a sandwich form. The variance estimate approximated the true variance in finite samples well. However, this pseudo-likelihood model is not guaranteed to satisfy semiparametric efficiency.

To adjust for informative external diagnosis, the pseudo-likelihood estimator was extended to the weighted pseudo-likelihood estimator in Chapter 5 by employing inverse probability weighting in the pseudo-likelihood model. We applied a marginal structure model to predict the probabilities. The proposed method appears to be fairly accurate, but the asymptotic variance estimates are likely to be over-conservative.

All three estimators are consistent and the regression parameter estimators satisfy asymptotic normality. In both the marginal and pseudo-likelihood model, the proposed estimators are more accurate and efficient than Pan's (1999) method, which is a proportional hazards model with a fixed threshold for interval censored data.

Through the three model, we were able to resolve the four issues in statistical analysis of interest. The proposed class of proportional hazards model for threshold-dependent time-to-disease occurrence can not directly estimate the diagnosis cutoff point for biomarkers, but can provide evidence about whether or not the threshold should differ across sub-populations by testing the interaction between the threshold effect and corresponding risk factors. We considered all possible ranges of thresholds,

which may be regarded as a non-informative prior to the threshold. In practice, we may need to restrict our scientific interest to narrower range of thresholds, so informative prior models could be chosen.

All proposed models were illustrated on either the main ARIC study data or the diabetes data. Compared to Pan's (1999) method, we found it easier identify risk factors because the proposed method is more efficient. The estimated effect sizes and directions agree well with the previous studies.

We may extend our models in several directions. First, the models and associated estimators could be implemented for multiple outcomes in biomedical studies. For example, both high blood pressure and hypercholesterolemia could be simultaneously important. We assumed that observation time is independent of time-to-event since the ARIC study scheduled visits ahead of time; however observation time can be dependent on the event of interest. Therefore, we could consider models accounting for the dependency of the observation time on a subject's health status. For semiparametric models for time-to-event data the baseline hazard functions are time-dependent and assumed to be unknown; however, time-varying coefficients as well as baseline information would be needed to account for time-dependent latent variables. The proposed method could be extended to models including time-varying risk factors or time-varying thresholds.

APPENDIX A: Technical Details for Chapter 3

A.1 Identifiability and Derivation of Efficient Score Functions

Lemma A.1.1. *The folowing model is identifiable, that is, if probability at Θ_1 is equal to the one at Θ_2, then $\Theta_1 = \Theta_2$.*

$$\prod_{i=1}^{n} \int_{-\infty}^{\infty} \exp\{-\Lambda_0(V_i) e^{\boldsymbol{\beta}^T X_i - \mu\xi}\} \Lambda_0(V_i) \mu \exp(\boldsymbol{\beta}^T X_i - \mu\xi) \frac{1}{\sigma} \phi\left\{ \frac{Y_i(V_i) - \xi}{\sigma} \right\} d\xi,$$

where $\phi(\cdot)$ is the standard normal density function.

Proof. We let the likelihood function for $\boldsymbol{\Theta}_k$,

$$f(\boldsymbol{\Theta}_k) = \int_{-\infty}^{\infty} \exp\big(- \Lambda_k(V) e^{\boldsymbol{\beta}_k^T X - \mu_k \xi} \big) \Lambda_k(V) \mu_k e^{\boldsymbol{\beta}_k^T X - \mu_k \xi} \frac{1}{\sigma} \phi\big(\frac{Y(V) - \xi}{\sigma} \big) d\xi, \quad k = 1, 2 \tag{6.1}$$

and suppose that two likelihood functions with $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$ are the same, that is, $f(\boldsymbol{\Theta}_1) = f(\boldsymbol{\Theta}_2)$. For any $W = (Y(V), V, X)$, the likelihood function $f(\boldsymbol{\Theta}_k)$ can be simplified as following:

$$\int_{-\infty}^{\infty} \exp\big\{ - \Lambda_k(V) e^{\boldsymbol{\beta}_k^T X - \mu_k \xi} \big\} \Lambda_k(V) \mu_k e^{\boldsymbol{\beta}_k^T X - \mu_k \xi - \frac{\xi^2}{2\sigma^2}} \exp\left\{ - \frac{Y(V)\xi}{\sigma^2} \right\} d\xi. \tag{6.2}$$

Since $Y(V)$ can be arbitrary real number, the integration in (6.2) are bilateral Laplace transformation of $g_k(\xi)$, where

$$g_k(\xi) = \exp\left\{ -\Lambda_k(V) e^{\boldsymbol{\beta}_k^T X - \mu_k \xi} \right\} \Lambda_k(V) \mu_k \exp\left(\boldsymbol{\beta}_k^T X - \mu_k \xi - \xi^2 / 2\sigma^2 \right), \quad k = 1, 2.$$

Hence $f(\boldsymbol{\Theta}_1) = f(\boldsymbol{\Theta}_2)$ implies $g_1 = g_2$ by one-to-one property of Laplace transformation.

Thus, for any ξ,

$$
\begin{aligned}
&-\Lambda_1(V)e^{\boldsymbol{\beta}_1^T X-\mu_1\xi} + \log\left(\Lambda_1(V)\right) + \log\mu_1 + \boldsymbol{\beta}_1^T X - \mu_1\xi \\
&= -\Lambda_2(V)e^{\boldsymbol{\beta}_2^T X-\mu_2\xi} + \log\left(\Lambda_2(V)\right) + \log\mu_2 + \boldsymbol{\beta}_2^T X - \mu_2\xi.
\end{aligned}
$$

Examining the behavior when ξ goes to $-\infty$, we obtain $\mu_1 = \mu_2$ and $\Lambda_1(V)e^{\boldsymbol{\beta}_1^T X} = \Lambda_2(V)e^{\boldsymbol{\beta}_2^T X}$. So by(A2), $\boldsymbol{\beta}_1=\boldsymbol{\beta}_2$ and $\Lambda_1(V)=\Lambda_2(V)$. □

Lemma A.1.2. *The efficient score function for* $\boldsymbol{\theta}$ *is*

$$
l^*_{\boldsymbol{\theta}}(\boldsymbol{\theta},\Lambda,\boldsymbol{W}) = \begin{pmatrix} \mu^{-1} - E(\kappa\xi \mid \boldsymbol{W}) - E(\kappa \mid \boldsymbol{W})\frac{E(E(\kappa|\boldsymbol{W})\{\mu^{-1}-E(\kappa\xi|\boldsymbol{W})\}|V)}{E(E(\kappa|\boldsymbol{W})^2|V)} \\ E(\kappa|\boldsymbol{W})\left[\boldsymbol{X} - \frac{E(\boldsymbol{X}E(\kappa|\boldsymbol{W})^2|V)}{E(E(\kappa|\boldsymbol{W})^2|V)}\right] \end{pmatrix}, \tag{6.3}
$$

where $\kappa=1-\Lambda(V)e^{\boldsymbol{\beta}^T\boldsymbol{X}-\mu\xi}$.

Proof. First, we have

$$
\begin{aligned}
\dot{l}_{\mu}(\boldsymbol{\theta},\Lambda,\boldsymbol{W}) &= E(\partial/\partial\mu l_c \mid \boldsymbol{W}) = \mu^{-1} - E_{\xi}(\kappa\xi \mid \boldsymbol{W}), \\
\dot{l}_{\boldsymbol{\beta}}(\boldsymbol{\theta},\Lambda,\boldsymbol{W}) &= E(\partial/\partial\beta l_c \mid \boldsymbol{W}) = \boldsymbol{X}E_{\xi}(\kappa \mid \boldsymbol{W}),
\end{aligned}
$$

where l_c denotes the log complete likelihood function based on a single observation.

Let $\{P_{\boldsymbol{\theta},\Lambda_\eta}\}$ be a regular parametric subfamily of models, $\{P_{\boldsymbol{\theta},\Lambda} \mid P_{\boldsymbol{\theta},\Lambda} \ll m, m$:Lebesque measure$\}$ and set $\partial/\partial\eta\,|_{\eta=0}\,\Lambda_\eta(V)=h(V)$ for $V>0$ and $h(V)\in L_2(F)$, then we have a score operator for Λ:

$$
\dot{l}_{\Lambda}(\boldsymbol{\theta},\Lambda,\boldsymbol{W})[h(V)] = h(V)/\Lambda(V) - h(V)e^{\boldsymbol{\beta}^T\boldsymbol{X}}E(e^{-\mu\xi} \mid \boldsymbol{W}) = h(V)/\Lambda(V)E(\kappa \mid \boldsymbol{W}).
$$

To obtain the efficient score function for μ, we need to find $h^*_\mu(V)$ satisfying

$$E\Big[\left\{\dot{l}_\mu(\boldsymbol{\theta},\Lambda,\boldsymbol{W}) - \dot{l}_\Lambda(\boldsymbol{\theta},\Lambda,\boldsymbol{W})[h^*_\mu(V)]\right\}\dot{l}_\Lambda(\boldsymbol{\theta},\Lambda,\boldsymbol{W})[h(V)]\Big] = 0, \text{ for every } h(V).$$

That is,

$$\begin{aligned}
&E_{\boldsymbol{W}}\left[\{\dot{l}_\mu(\boldsymbol{\theta},\Lambda,\boldsymbol{W}) - \dot{l}_\Lambda(\boldsymbol{\theta},\Lambda,\boldsymbol{W})[h^*_\mu])\}\dot{l}_\Lambda(\boldsymbol{\theta},\Lambda,\boldsymbol{W})[h]\right] \\
&= E\left[\left\{\mu^{-1} - E(\kappa\xi \mid \boldsymbol{W}) - h^*_\mu(V)\Lambda(V)^{-1}E(\kappa \mid \boldsymbol{W})\right\}\left\{h(V)\Lambda(V)^{-1}E(\kappa \mid \boldsymbol{W})\right\}\right] \\
&= E_V\Big(h(V)\Lambda(V)^{-1}E_{Y,X}\left[E(\kappa \mid \boldsymbol{W})\{\mu^{-1} - E(\kappa\xi \mid \boldsymbol{W})\} - h^*_\mu(V)\Lambda(V)^{-1}E(\kappa \mid \boldsymbol{W})^2 \mid V\right]\Big) \\
&= 0.
\end{aligned}$$

Therefore, $E_{Y,X}\Big[E(\kappa \mid \boldsymbol{W})\{\mu^{-1} - E(\kappa\xi \mid \boldsymbol{W})\} - h^*_\mu(V)\Lambda(V)^{-1}E(\kappa \mid \boldsymbol{W})^2 \mid V\Big] = 0.$
We then obtain

$$h^*_\mu(V) = \Lambda(V)E\Big[E(\kappa \mid \boldsymbol{W})\left\{\mu^{-1} - E(\kappa\xi \mid \boldsymbol{W})\right\} \mid V\Big]\Big/E\left\{E(\kappa \mid \boldsymbol{W})^2 \mid V\right\}. \tag{6.4}$$

Using the $h^*_\mu(V)$ in (6.4), the efficient score function for μ is

$$\begin{aligned}
l^*_\mu(\boldsymbol{\theta},\Lambda,\boldsymbol{W}) &= \dot{l}_\mu(\boldsymbol{\theta},\Lambda,\boldsymbol{W}) - \dot{l}_\Lambda(\boldsymbol{\theta},\Lambda,\boldsymbol{W})[h^*_\mu(V)] \\
&= \mu^{-1} - E(\kappa\xi \mid \boldsymbol{W}) - E(\kappa \mid \boldsymbol{W})\frac{E\Big[E(\kappa \mid \boldsymbol{W})\{\mu^{-1} - E(\kappa\xi \mid \boldsymbol{W})\} \mid V\Big]}{E\Big\{E(\kappa \mid \boldsymbol{W})^2 \mid V\Big\}}.
\end{aligned}$$

Similarly, the efficient score function for $\boldsymbol{\beta}$ is obtained by solving equation:

$$\begin{aligned}
&E_{\boldsymbol{W}}[\{\dot{l}_\beta(\boldsymbol{\theta},\Lambda,\boldsymbol{W}) - \dot{l}_\Lambda(\boldsymbol{\theta},\Lambda,\boldsymbol{W})[h^*_\beta]\}\dot{l}_\Lambda(\boldsymbol{\theta},\Lambda,\boldsymbol{W})[h]] \\
&= E_{\boldsymbol{W}}[\{\boldsymbol{X}E(\kappa \mid \boldsymbol{W}) - h^*_\beta(V)\Lambda(V)^{-1}E(\kappa \mid \boldsymbol{W})\}\{h(V)\Lambda(V)^{-1}E(\kappa \mid \boldsymbol{W})\}] \\
&= E_V[h(V)\Lambda(V)^{-1}E_{Y,X}\{XE(\kappa \mid \boldsymbol{W})^2 - h^*_\beta(V)\Lambda(V)^{-1}E(\kappa \mid \boldsymbol{W})^2 \mid V\}] \\
&= 0.
\end{aligned}$$

Thus, $E_{Y,\boldsymbol{X}}\{\boldsymbol{X}E(\kappa \mid \boldsymbol{W})^2 - h^*_\beta(V)\Lambda(V)^{-1}E(\kappa \mid \boldsymbol{W})^2 \mid V\} = 0$. So,

$$E\{\boldsymbol{X}E(\kappa \mid \boldsymbol{W})^2 \mid V\} = h^*_\beta(V)\Lambda(V)^{-1}E\{E(\kappa \mid \boldsymbol{W})^2 \mid V\}.$$

We hence obtain $h^*_\beta(V)$ as

$$h^*_\beta(V) = \Lambda(V)E\{\boldsymbol{X}E(\kappa \mid \boldsymbol{W})^2 \mid V\}/E\{E(\kappa \mid \boldsymbol{W})^2 \mid V\}. \tag{6.5}$$

So the efficient score function for $\boldsymbol{\beta}$ is

$$l^*_\beta(\boldsymbol{\theta}, \Lambda, \boldsymbol{W}) = \dot{l}_\beta(\boldsymbol{\theta}, \Lambda, \boldsymbol{W}) - \dot{l}_\Lambda(\boldsymbol{\theta}, \Lambda, \boldsymbol{W})[h^*_\beta] = E(\kappa \mid \boldsymbol{W})\left[\boldsymbol{X} - \frac{E\{\boldsymbol{X}E(\kappa \mid \boldsymbol{W})^2 \mid V\}}{E\{E(\kappa \mid \boldsymbol{W})^2 \mid V\}}\right].$$

□

Lemma A.1.3. *The information operator $E\left[(\dot{l}_\theta, \dot{l}_\Lambda)^*(\dot{l}_\theta, \dot{l}_\Lambda)\right]$, which maps $\Theta \times H$ to the dual space of $\Theta \times H$ (equivalent to $\Theta \times H$), is continuously invertible at θ_0, Λ_0, where $\dot{l}^*_{\boldsymbol{\theta}}$ and $\dot{l}^*_\Lambda$ are the adjoint operators of the linear operators, $\dot{l}_{\boldsymbol{\theta}}$ and $\dot{l}_\Lambda$,respectively.*

Proof. It suffices to show that $E(\dot{l}^*_\Lambda \dot{l}_\Lambda)$, and $E(\dot{l}^*_{\boldsymbol{\theta}} \dot{l}_{\boldsymbol{\theta}}) - E(\dot{l}^*_{\boldsymbol{\theta}} \dot{l}_\Lambda)E(\dot{l}^*_\Lambda \dot{l}_\Lambda)^{-1}E(\dot{l}^*_\Lambda \dot{l}_{\boldsymbol{\theta}})$ (denote as the matrix A) are invertible. By taking linear operator of $\dot{l}_\Lambda$, we obtain

$$E(\dot{l}^*_\Lambda \dot{l}_\Lambda[h, \widetilde{h}]) = E\left[\tilde{h}(V)h(V)\Lambda(V)^{-2}E\left\{E(\kappa \mid \boldsymbol{W})^2 \mid V\right\}\right].$$

Therefore, $E(\dot{l}^*_\Lambda \dot{l}_\Lambda[h]) = h(V)\Lambda(V)^{-2}E\left\{E(\kappa \mid \boldsymbol{W})^2 \mid V\right\}$ so is invertible from $L_2(P_V)$ to $L_2(P_V)$.

If A is singular, then there exists some non-zero vector ν such that $\nu^T A\nu = 0$. Therefore,it gives $E\{(\nu^T \dot{l}_{\boldsymbol{\theta}} - \dot{l}_\Lambda[h])^{\otimes 2}\} = 0$, where $h = \nu^T E[\dot{l}^*_\Lambda \dot{l}_\Lambda]^{-1}E[\dot{l}^*_\Lambda \dot{l}_{\boldsymbol{\theta}}]$. As the

result, $\nu^T \dot{l}_{\boldsymbol{\theta}} - \dot{l}_\Lambda[h] = 0$ almost surely. We obtain

$$E\left[\left\{\nu^T(-\xi, \boldsymbol{X}) - h(V)/\Lambda_0(V)\right\}\left\{1 - \Lambda_0(V)e^{\boldsymbol{\beta}_0^T \boldsymbol{X} - \mu_0 \xi}\right\} + \nu_1/\mu_0 \mid \boldsymbol{W}\right] = 0,$$

where ν_1 is the component of ν corresponding to μ. The left-hand side can be treated as a Laplace transformation of some function of ξ so we immediately conclude

$$\left\{\nu^T(-\xi, \boldsymbol{X}) - h(V)/\Lambda_0(V)\right\}\left\{1 - \Lambda_0(V)e^{\beta_0^T \boldsymbol{X} - \mu_0 \xi}\right\} + \nu_1/\mu_0 = 0.$$

Since ξ is arbitrary, $\nu_1 = 0, \quad \nu_{-1}^T \boldsymbol{X} = 0, \quad h(V)/\Lambda_0(V) = 0, \quad$ where $\quad \nu_{-1} = (\nu_2, \ldots, \nu_d)$. Therefore, $\nu = 0$, and this leads to the contradiction. □

A.2 Proof of Asymptotic Results

Proof of Theorem 3.4.1. :

We reparametrize F=$1 - e^{-\Lambda}$, $\theta = \theta$ and let

$$\begin{aligned} & f(\theta, F \mid W) \\ &= -\int_{-\infty}^{\infty} \{1 - F(V)\}^{\exp(\beta^T \boldsymbol{X} - \mu\xi)} \log\{1 - F(V)\}\, \mu e^{\beta^T \boldsymbol{X} - \mu\xi} \frac{1}{\sigma} \phi\left\{\frac{Y(V) - \xi}{\sigma}\right\} d\xi. \end{aligned} \tag{6.6}$$

Let $0 < \alpha < 1$ be a fixed constant throughout the proof. By concavity of the log function, the model identifiability and Jensen's inequality,

$$E\left(\log\left[1 + \alpha\left\{\frac{f(\theta, F \mid \boldsymbol{W})}{f(\boldsymbol{\theta}_0, F_0 \mid \boldsymbol{W})} - 1\right\}\right]\right) < 0. \tag{6.7}$$

For an open neighborhood $\mathscr{N}$ around $(\boldsymbol{\theta}, F)$, define $\bar{f}(\boldsymbol{W} \mid \mathscr{N})$=$\sup_{\boldsymbol{\theta}, F \in \mathscr{N}} f(\boldsymbol{\theta}, F \mid \boldsymbol{W})$. For a sequence of open balls $\mathscr{N}_\epsilon$ with radius ϵ shrinking to $(\boldsymbol{\theta}, F)$ as ϵ goes to 0, we have $\bar{f}(\boldsymbol{W} \mid \mathscr{N}_\epsilon) \to f(\boldsymbol{\theta}, F \mid \boldsymbol{W})$. By (6.7), for ϵ sufficiently small, there is an $\eta_\epsilon > 0$

so that

$$E\left(\log\left[1+\alpha\left\{\frac{\bar{f}(\boldsymbol{W} \mid \mathscr{N}_\epsilon)}{f(\boldsymbol{\theta}_0, F_0 \mid \boldsymbol{W})}-1\right\}\right] \wedge \eta_\epsilon \right) < 0. \tag{6.8}$$

On the other hand, since $(\widehat{\boldsymbol{\theta}}_n, \widehat{F}_n)$ is the maximum likelihood estimator, we have

$$\sum_{i=1}^{n} \log f(\widehat{\boldsymbol{\theta}}_n, \widehat{F}_n \mid W_i) \geq \sum_{i=1}^{n} \log f(\boldsymbol{\theta}_0, F_0 \mid W_i).$$

By concavity of log function, this implies

$$\sum_{i=1}^{n} \log\left[1+\alpha\left\{\frac{f(\widehat{\theta}_n, \widehat{F}_n \mid W_i)}{f(\boldsymbol{\theta}_0, F_0 \mid W_i)}-1\right\}\right] \geq 0. \tag{6.9}$$

For any vaguely open neighborhood $\mathscr{N}_0$ of the true $(\boldsymbol{\theta}_0, F_0)$, its complement in $\Theta \times \mathscr{F}$ is a vaguely closed subset of a compact set, hence also vaguely compact. Then open cover $\left\{\mathscr{N}_{(\boldsymbol{\theta},F)} \mid (\boldsymbol{\theta}, F) \notin \mathscr{N}_0\right\}$ of this complement has a finite subcover $\mathscr{N}_{(\theta_1,F_1)}$, …, $\mathscr{N}_{(\theta_k,F_m)}$. If $(\widehat{\boldsymbol{\theta}}_n, \widehat{F}_n)$ is not in $\mathscr{N}_0$, it is in one of the subcovers. By (6.9), we have

$$\left\{(\widehat{\boldsymbol{\theta}}, \widehat{F} \notin \mathscr{N}_0)\right\} \subset \cup_{k=1}^{m}\left(\sum_{i=1}^{n} \log\left[1+\alpha\left\{\frac{\bar{f}(\boldsymbol{W} \mid \mathscr{N}_{(\theta_k,F_k)})}{f(\boldsymbol{\theta}_0, F_0 \mid W_i)}-1\right\}\right] \wedge \eta_{(\theta_k,F_k)} \geq 0 \right).$$

The probability of each of the sets in the union is the probability that an average of uniformly bounded and independent random variables is non-negative. However, these random variables have negative expectation by (6.8). By Hoeffding's inequality, each of the probabilities is of the order $e^{-\epsilon n}$ where ϵ can be chosen equal to $2\mu^2/(\eta_0 - \log(1-\alpha))^2$. Here $\eta_0 = \max\left\{\eta_{(\theta_k,F_k)} \mid 1 \leq k \leq m\right\}$, and μ is any negative number that is greater than the expectation in (6.8). This is true for all $n \geq 1$. Hence,

$$\sum_{n=1}^{\infty} P\left\{(\widehat{\boldsymbol{\theta}}_n, \widehat{F}_n) \notin \mathscr{N}_0\right\} < \infty.$$

By the Borel-Cantelli lemma, it follows that, with probability 1, $(\widehat{\boldsymbol{\theta}}_n, \widehat{F}_n) \in \mathscr{N}_0$ for

all n sufficiently large. By the definition of our product topology, this implies that $\widehat{\boldsymbol{\theta}}_n$ converges to $\boldsymbol{\theta}_0$ in $P_{(\boldsymbol{\theta}_0,F_0)}$-almost surely and $\widehat{F}_n$ converges to F_0 in $P_{(\boldsymbol{\theta}_0,F_0)}$-almost surely. In particular, if F_0 is continuous, this implies

$$\lim_{n\to\infty} \sup_{v\in\mathscr{S}[V]} |\widehat{F}_n(v) - F_0(v)| = 0 \quad P_{(\boldsymbol{\theta}_0,F_0)}\text{-almost surely.}$$

Since $\widehat{\Lambda}=-\log(1-\widehat{F}_n)$, Theorem 1 is proved. □

Once consistency of $\widehat{\theta}$ is proved, we can concentrate on a neighborhood of $\boldsymbol{\theta}_0$. For any $\eta > 0$, let $B(\boldsymbol{\theta}_0, \eta)$ be the ball centered at $\boldsymbol{\theta}_0$ with radius η. If $\boldsymbol{\theta}_0$ is on the boundary of Θ, then take $B(\boldsymbol{\theta}_0, \eta) \cap \Theta$ instead of $B(\boldsymbol{\theta}_0, \eta)$. Then $B(\boldsymbol{\theta}_0, \eta)$ is included in $\boldsymbol{\Theta}$. We suppose that *condition*3 is satisfied so that Λ_0 is bounded and away from 0 on $\mathscr{S}[V]$. Since we have proved that $\widehat{\Lambda}$ converges on $\mathscr{S}[V]$, we may restrict $\widehat{\Lambda}$ to the following class of functions:

$$\Phi = \{\Lambda \mid \Lambda \text{ is non-decreasing and } 0 < 1/M \le \Lambda(t) \le M < \infty \text{ for all } t \in \mathscr{S}[V]\},$$

where M is a large positive constant.

For any probability measure P, define $L_2(P) = \left\{g \mid \int g^2 dP < \infty\right\}$. Let $\|\cdot\|_{2,P}$ be the usual $L_{2,P}$ norm. For any subclass $\mathscr{F}$ of $L_2(P)$, define the bracketing number $N_{[\,]}(\epsilon, \mathscr{F}, L_2(P))$ as infimum of cardinal numbers for $\left\{g_i^L, g_i^U \mid g_i^L \le g \le g_i^U, g \in \mathscr{F}, \text{ for some } i, \text{ and } \|g_i^U - g_i^L\| \le \epsilon\right\}$. By the following lemma, we figure out size of the class for likelihood functions of our interest.

Lemma A.2.1. *Let*

$$\mathscr{H} = \left\{\log f(\boldsymbol{\theta}, \Lambda\big|\boldsymbol{W})|\boldsymbol{\theta} \in B(\boldsymbol{\theta}_0, \eta), \Lambda \in \Phi\right\}, \tag{6.10}$$

where $f(\boldsymbol{\theta}, \Lambda|\boldsymbol{W})$ *is the re-parametrized function with* Λ *instead of* F *from (*6.6*). Suppose that (A2) is satisfied. Then there exists a constant* $C > 0$ *such that*

$$\sup_Q N_{[\,]}(\epsilon, \mathscr{H}, L_2(Q)) \leq C(1/\epsilon^d)e^{1/\epsilon} \quad \textit{for all} \quad \epsilon > 0,$$

where d *is the dimension of* $\boldsymbol{\theta}$*. Hence, for* ϵ *small enough, we have*

$$\sup_Q \log N_{[\,]}(\epsilon, \mathscr{H}, L_2(Q)) \leq C(1/\epsilon).$$

Here Q *runs through the class of all probability measures.*

The proof of Lemma $A.2.1$ is adapted from Huang (1996), where the author provided the order of the entropy for a class of log-likelihood functions over bounded parameter space in current status data.

Proof of Lemma $A.2.1$. We first calculate the order of the bracketing number for class $\mathscr{H}'$, where $\mathscr{H}'=\{f \,|\log f \in \mathscr{H}\}$. It is known that for the class of functions,

$$\Phi = \{\Lambda \mid \Lambda \text{ is non-decreasing and } 0 < 1/M \leq \Lambda(v) \leq M < \infty \text{ for all } v \in \mathscr{S}[V]\},$$

where M is a constant and it is known that its ϵ bracketing number is of the order of $m = N_{[\,]}(\epsilon, \Phi, L_2(P)) = O(e^{1/\epsilon})$. This means that the class Φ has the finite entropy. Let $\Lambda_i^{*L} = \Lambda_i^L - \epsilon$ and $\Lambda_i^{*U} = \Lambda_i^U + \epsilon$ for $1 \leq i \leq m$. Then for any $\Lambda \in \Phi$, we have, for some i, $\Lambda_i^{*L} + \epsilon \leq \Lambda \leq \Lambda_i^{*U} - \epsilon$ and $\|\Lambda_i^{*U} - \Lambda_i^{*L}\|_{2,P_V} \leq 3\epsilon$. Since Φ is uniformly bounded away from 0, we can choose ϵ small enough such that all the bracketing functions stay away from 0.

For the true $\mu_0 > 0$, we can find a constant δ_0 such that $\mu_0 > \delta_0 > 0$. Related to $\boldsymbol{\theta}$, we can also choose k points $(\beta_1, \mu_1) = \theta_1, \ldots, (\boldsymbol{\beta}_k, \mu_k) = \boldsymbol{\theta}_k$ in $B(\theta_0, \eta)$ such that for any $(\boldsymbol{\beta}, \mu) \in B(\theta_0, \eta)$, $|\boldsymbol{\beta}_j - \boldsymbol{\beta}| < \delta_1$ and $|\mu_j - \mu| < \delta_2$ for given constants $\delta_1 > 0$ and $0 < \delta_2 < \delta_0$,

since Θ_0 is compact by (A1). Then for any $(\boldsymbol{\beta}, \mu) \in B(\theta_0, \eta)$ and for some $1 \le j \le k$, there exist a positive constant C_1 such that $|(\boldsymbol{\beta}_j - \boldsymbol{\beta})^T \boldsymbol{X} - (\mu_j - \mu)\xi| \le C_1\delta_1 + \delta_2|\xi|$.

Using the chosen k regression parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$ and m cumulative hazard functions $\{\Lambda_i^{*L}, \Lambda_i^{*U}\}_{i=1}^m$, we will show how to construct upper and lower envelope functions for the log likelihood functions belonging to (6.10). For any $(\boldsymbol{\theta}, \Lambda) \in B(\boldsymbol{\theta}_0, \eta) \times \Phi$, we can choose $\boldsymbol{\theta}_j$,$\{\Lambda_i^{*L}(V), \Lambda_i^{*L}(V)\}$ satisfying

$$\begin{aligned}
& \exp\big\{-\Lambda^{*U}(V)e^{\boldsymbol{\beta}_j^T\boldsymbol{X}+C_1\delta_1-\mu_j\xi+\delta_2|\xi|}\big\}\Lambda^{*L}(V)(\mu_j-\delta_2|\xi|)e^{\boldsymbol{\beta}_j^T\boldsymbol{X}-C_1\delta_1-\mu_j\xi-\delta_2|\xi|} \\
\le\ & \exp\big\{-\Lambda(V)e^{\boldsymbol{\beta}^T\boldsymbol{X}-\mu\xi}\big\}\Lambda(V)\mu e^{\boldsymbol{\beta}^T\boldsymbol{X}-\mu\xi} \\
\le\ & \exp\big\{-\Lambda^{*L}(V)e^{\boldsymbol{\beta}_j^T\boldsymbol{X}-C_1\delta_1-\mu_j\xi-\delta_2|\xi|}\big\}\Lambda^{*U}(V)(\mu_j+\delta_2|\xi|)e^{\boldsymbol{\beta}_j^T\boldsymbol{X}+C_1\delta_1-\mu_j\xi+\delta_2|\xi|}.
\end{aligned}$$

It is well known that the minimum value of k can be on the order of $O(\epsilon^{-d})$.

Then we let

$$\begin{aligned}
f_{ij}^{*L} &= \int \exp\big\{-\Lambda^{*U}(V)e^{\boldsymbol{\beta}_j^T\boldsymbol{X}+C_1\delta_1-\mu_j\xi+\delta_2|\xi|}\big\}\Lambda^{*L}(V)\{\mu_j-\delta_2|\xi|\}e^{\boldsymbol{\beta}_j^T\boldsymbol{X}-C_1\delta_1-\mu_j\xi-\delta_2|\xi|} \\
&\quad \times\phi\Big(\frac{Y(V)-\xi}{\sigma}\Big)d\xi, \\
f_{ij}^{*U} &= \int \exp\big\{-\Lambda^{*L}(V)e^{\boldsymbol{\beta}_j^T\boldsymbol{X}-C_1\delta_1-\mu_j\xi-\delta_2|\xi|}\big\}\Lambda^{*U}(V)\{\mu_j+\delta_2|\xi|\}e^{\boldsymbol{\beta}_j^T\boldsymbol{X}+C_1\delta_1-\mu_j\xi+\delta_2|\xi|} \\
&\quad \times\phi\Big(\frac{Y(V)-\xi}{\sigma}\Big)d\xi,
\end{aligned}$$

so f_{ij}^{*L} and f_{ij}^{*U} are finite envelope functions for $f(\boldsymbol{\theta}, \Lambda \mid W)$, which is $\log(f) \in \mathscr{H}$. Finally, we need to show that $|f_{ij}^{*U} - f_{ij}^{*L}|$ can be small enough to be less than an arbitrary constant ϵ.

$$|f_{ij}^{*U} - f_{ij}^{*L}| \le \int \big(C_2|\Lambda^{*U}(V) - \Lambda^{*L}(V)| + C_3\delta_1 + C_4\delta_2|\xi|\big)\phi\left\{\frac{Y(V)-\xi}{\sigma}\right\}d\xi,$$

for some constant C_2, C_3, and C_4. Thus,

$$\begin{aligned}\|f_{ij}^{*U} - f_{ij}^{*L}\|_{2,P} &\leq C_2\|\Lambda^{*U}(v) - \Lambda^{*L}(v)\|_{2,P_V} + C_3\delta_1 + C_4'\delta_2 \\ &\leq 3C_2\epsilon + C_3\delta_1 + C_4'\delta_2,\end{aligned}$$

for some constant C_4'. This implies that there exist f_{ij}^{*U}, f_{ij}^{*L}, $i = 1, \ldots, m$ and $j = 1, \ldots, k$ such that, for any $f \in \mathscr{H}'$. $f_{ij}^{*L} \leq f \leq f_{ij}^{*U}$, for some $1 \leq i \leq m$, $1 \leq j \leq k$, and $|f_{ij}^{*U} - f_{ij}^{*L}|_{2,P} \leq 3C_2\epsilon + C_3\delta_1 + C_4\delta_2'$. This means that the bracketing number $N_{[\,]}(\epsilon, \mathscr{H}', L_2(P))$ for the class $\mathscr{H}'$ is of order $mk = O(\epsilon^{-d}e^{1/\epsilon})$. Note that since log function on the domain bounded away from 0 is Lipschitz continuous and any function $f \in \mathscr{H}'$ is bounded and away from 0, by (A2) and (A3), the bracketing number of $\mathscr{H}$ is dominated by the bracketing number of the class $\mathscr{H}'$. □

Proof of Theorem 3.4.2. We apply Theorem 3.4.1 of van der Vaart and Wellner (1996) to the prove the results. Specifically, we need to check the following conditions: let $0 \leq \delta < \eta$ be arbitrary and C be a generic constant, then for $\delta < d\left((\boldsymbol{\theta}, \boldsymbol{\theta}_0), (\Lambda, \Lambda_0)\right) \leq \eta$,

$$(i) \quad \sup_{\delta/2 < d\{(\boldsymbol{\theta},\boldsymbol{\theta}_0),(\Lambda,\Lambda_0)\} \leq \delta, \boldsymbol{\theta} \in \Theta_0} E\big\{l(\boldsymbol{\beta}, \Lambda, \boldsymbol{W}) - l(\beta_0, \Lambda_0, \boldsymbol{W})\big\} \leq -\delta^2,$$

$$(ii) \quad E^* \sup_{\delta/2 < d\{(\boldsymbol{\theta},\boldsymbol{\theta}_0),(\Lambda,\Lambda_0)\} \leq \delta, \boldsymbol{\theta} \in \boldsymbol{\Theta}_0} n^{1/2}|(\mathbb{P}_n - P)\{l(\beta, \Lambda, \boldsymbol{W}) - l(\beta_0, \Lambda_0, \boldsymbol{W})\}| \leq C\psi(\delta),$$

for function ψ such that $\delta \to \psi(\delta)/\delta^\alpha$ is increasing on (δ, η) for some $\alpha < 2$.

For the first condition, we perform the Taylor expansion to obtain

$$E\big\{l(\boldsymbol{\theta}_0, \Lambda_0) - l(\boldsymbol{\theta}, \Lambda)\big\} = E\big[l\left\{(1-\epsilon)\boldsymbol{\theta}_0 + \epsilon\boldsymbol{\theta}, (1-\epsilon)\Lambda_0 + \epsilon\Lambda\right\}\big]\Big|_1^0$$

$$= -\frac{1}{2}\frac{\partial^2}{\partial\epsilon^2}\Big|_{\epsilon=\epsilon^*} E\left[l\left\{(1-\epsilon)\boldsymbol{\theta}_0 + \epsilon\theta, (1-\epsilon)\Lambda_0 + \epsilon\Lambda\right\}\right],$$

for some $\epsilon^* \in (0,1)$. For η small enough, we note that right-hand side is equal to

$$\left[\{I(\boldsymbol{\theta}_0,\Lambda_0)+o(1)\}\,(\boldsymbol{\theta}-\boldsymbol{\theta}_0,\Lambda-\Lambda_0)(\boldsymbol{\theta}-\boldsymbol{\theta}_0,\Lambda-\Lambda_0)\right],$$

where $I(\boldsymbol{\theta}_0,\Lambda_0)$ is the information operator in Lemma $A.1.3$.

Since the information operator is invertible linear operator and uniformly bounded and away from 0, for some constant C,

$$\|I(\boldsymbol{\theta}_0,\Lambda_0)(a,h)(a,h)\| \geq C\{|a|^2+\|h\|^2_{2,P_V}\}.$$

Hence, for some constant C,

$$\left[\{I(\boldsymbol{\theta}_0,\Lambda_0)+o(1)\}\,(\boldsymbol{\theta}-\boldsymbol{\theta}_0,\Lambda-\Lambda_0)(\boldsymbol{\theta}-\boldsymbol{\theta}_0,\Lambda-\Lambda_0)\right] \geq C(|\boldsymbol{\theta}-\boldsymbol{\theta}_0|^2+\|\Lambda-\Lambda_0\|^2_{2,P_V}).$$

Thus, condition (i) holds.

For the second condition, by Lemma$A.2.1$, for some constants C and M,

$$J_{[\,]}(\eta,\mathscr{H},L_2(P)) \leq \int_0^\eta \sqrt{1+C\epsilon^{-1}}d\epsilon \leq \int_0^\eta M\epsilon^{-1/2}d\epsilon = M\eta^{1/2}.$$

Then, according to Lemma 3.4.2 of van der Vaart and Wellner (1996),

$$E^* \sup_{d\{(\boldsymbol{\theta},\Lambda),(\boldsymbol{\theta}_0,\Lambda_0)\}\leq\eta} \left|n^{1/2}(P_n-P)\left\{l(\boldsymbol{\theta},\Lambda\mid\boldsymbol{W})-l(\boldsymbol{\theta}_0,\Lambda_0\mid\boldsymbol{W})\right\}\right| = O(1)\eta^{1/2}\left(1+\frac{\eta^{1/2}}{\eta^2 n^{1/2}}M\right).$$

Finally, let

$$\psi(\eta) = \eta^{1/2}\left(1+\frac{\eta^{1/2}}{\eta^2 n^{1/2}}\right).$$

Then $\phi(\delta)/\delta^\alpha$ is an increasing function for some $0<\alpha<1/2$, so the condition (ii) is satisfied. In addition, since $(\widehat{\boldsymbol{\theta}},\widehat{\Lambda})$ maximizes $l(\boldsymbol{\theta},\Lambda)$, $Pl(\widehat{\boldsymbol{\theta}},\widehat{\Lambda},\boldsymbol{W}) \geq Pl(\beta_0,\Lambda_0\mid\boldsymbol{W})$ is also satisfied. When $r_n = n^{1/3}$, then $n^{2/3}\psi(n^{-1/3}) = O(n^{1/2})$ for every n.

Hence, all the conditions of Theorem 3.4.1 of van der Vaart and Wellner (1996) are satisfied. This implies

$$d\{(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda}), (\boldsymbol{\theta}_0, \Lambda_0)\} = O_p(n^{-1/3}).$$

□

Proof of Theorem 3.4.3. We will prove that asymptotic distribution of the MLE $\widehat{\boldsymbol{\theta}}$ is normal distribution with mean 0 and semiparametric efficient variance by following the approach (p.1007) of Zeng, Yin, and Ibrahim (2005). For simplicity of notations, $\dot{l}_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \Lambda, \boldsymbol{W})$ and $\dot{l}_{\Lambda}(\boldsymbol{\theta}, \Lambda, \boldsymbol{W})$ also denoted as $\dot{l}_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \Lambda)$ and $\dot{l}_{\Lambda}(\boldsymbol{\theta}, \Lambda)$, respectively.

Since $(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda})$ are maximum likelihood estimator for $(\boldsymbol{\theta}, \Lambda)$, we immediately obtain that

$$\mathbb{P}_n\{\dot{l}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda}) - \dot{l}_{\Lambda}(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda})[h^*]\} = 0.$$

Thus,

$$\mathbb{G}_n\{\dot{l}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda}) - \dot{l}_{\Lambda}(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda})[h^*]\} = -n^{1/2}P\{\dot{l}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda}) - \dot{l}_{\Lambda}(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda})[h^*]\},$$

where $\mathbb{G}_n = n^{1/2}(\mathbb{P}_n - P)$.

Let us consider the following two classes of functions:

$$\{\dot{l}_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \Lambda) - \dot{l}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \Lambda_0) \quad | \quad |\boldsymbol{\theta} - \boldsymbol{\theta}_0| \le \eta \text{ and } \|\Lambda - \Lambda_0\|_{2,P_V} \le \eta\} \quad \text{and}$$
$$\{\dot{l}_{\Lambda}(\boldsymbol{\theta}, \Lambda)[h^*] - \dot{l}_{\Lambda}(\boldsymbol{\theta}_0, \Lambda_0)[h^*] \quad | \quad |\boldsymbol{\theta} - \boldsymbol{\theta}_0| \le \eta \text{ and } \|\Lambda - \Lambda_0\|_{2,P_V} \le \eta\},$$

where η is near 0. The entropy numbers for the two classes are of order $1/\eta$ and this implies that these two classes are P-Donsker. Hence, $\dot{l}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda}) - \dot{l}_{\Lambda}(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda})[h^*]$ belongs to a P-Donsker class. This leads to

$$\mathbb{G}_n\{\dot{l}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \Lambda_0) - \dot{l}_{\Lambda}(\boldsymbol{\theta}_0, \Lambda_0)[h^*]\} + o_p(1) = -n^{1/2}P\{\dot{l}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda}) - \dot{l}_{\Lambda}(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda})[h^*]\}. \qquad (6.11)$$

We perform a Taylor's series expansion of the right side in (6.11) at $(\boldsymbol{\theta}_0, \Lambda_0)$:

$$\begin{aligned}
& \mathbb{G}_n\{\dot{l}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \Lambda_0) - \dot{l}_{\Lambda}(\boldsymbol{\theta}_0, \Lambda_0)[h^*]\} + o_p(1) \\
= \; & -n^{1/2} P\{\ddot{l}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \Lambda_0) - \ddot{l}_{\Lambda\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \Lambda_0)[h^*]\}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\
& -n^{1/2} P\{\ddot{l}_{\boldsymbol{\theta}\Lambda}(\boldsymbol{\theta}_0, \Lambda_0)[\widehat{\Lambda} - \Lambda_0] - \ddot{l}_{\Lambda\Lambda}(\boldsymbol{\theta}_0, \Lambda_0)[h^*, \widehat{\Lambda} - \Lambda_0]\} \\
& +n^{1/2} O(|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0|^2 + \|\widehat{\Lambda} - \Lambda_0\|^2_{2,P_V}). \qquad (6.12)
\end{aligned}$$

Here $\ddot{l}_{\boldsymbol{\theta}\Lambda}(\boldsymbol{\theta}_0, \Lambda_0)[\widehat{\Lambda} - \Lambda_0]$ is the derivative of $\dot{l}_{\boldsymbol{\theta}}$ along the path $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, $\Lambda = \Lambda_0 + \epsilon(\widehat{\Lambda} - \Lambda_0)$, and $\ddot{l}_{\Lambda\Lambda}(\boldsymbol{\theta}_0, \Lambda_0)[h^*, \widehat{\Lambda} - \Lambda_0]$ is the derivative of $\dot{l}_{\Lambda}[h^*]$ along the path $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, $\Lambda = \Lambda_0 + \epsilon(\widehat{\Lambda} - \Lambda_0)$.

The second term on the right side of (6.12) is 0, because the second term can be re-expressed as

$$-n^{1/2} P\left[\left\{\dot{l}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \Lambda_0)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) - \dot{l}_{\Lambda}(\boldsymbol{\theta}_0, \Lambda_0)[h^*]\right\}\left(\dot{l}_{\Lambda}[\widehat{\Lambda} - \Lambda_0]\right)\right] \qquad (6.13)$$

and h^* satisfies that the equation (6.13) is 0. The third term on the right side of (6.12) is $o_p(1)$, because of the convergence rate for $(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda})$. Hence,

$$\begin{aligned}
& -n^{1/2} P\left\{\ddot{l}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \Lambda_0) - \ddot{l}_{\boldsymbol{\theta}\Lambda}(\boldsymbol{\theta}_0, \Lambda_0)[h^*]\right\}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\
& = \mathbb{G}_n\left\{\dot{l}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \Lambda_0) - \dot{l}_{\Lambda}(\boldsymbol{\theta}_0, \Lambda_0)[h^*]\right\} + o_p(1). \qquad (6.14)
\end{aligned}$$

We show that the matrix $P\left\{\ddot{l}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \Lambda_0) - \ddot{l}_{\boldsymbol{\theta}\Lambda}(\boldsymbol{\theta}_0, \Lambda_0)[h^*]\right\}$ is non-singular. Suppose that the matrix is singular, then there exist a non-0 vector b such that

$$b^T P\left\{\ddot{l}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \Lambda_0) - \ddot{l}_{\boldsymbol{\theta}\Lambda}(\boldsymbol{\theta}_0, \Lambda_0)[h^*]\right\} b = 0,$$

that is, $P\left\{\left(b^T\dot{l}_{\boldsymbol{\theta}} - b^T\dot{l}_{\Lambda}[h^*]\right)^2\right\} = 0$. Then

$$b^T\dot{l}_{\boldsymbol{\theta}} - b^T\dot{l}_{\Lambda}[h^*] = b^T\begin{pmatrix} \mu^{-1} - E_{\xi}(\kappa\xi \mid W) - h^*_{\mu}(V)\Lambda(V)^{-1}E(\kappa \mid W) \\ XE_{\xi}(\kappa \mid W) - h^*_{\beta}(V)\Lambda(V)^{-1}E(\kappa \mid W) \end{pmatrix} = 0 \quad \text{almost surely,}$$

where $\kappa = 1 - \Lambda(V)e^{\beta^T X - \mu\xi}$ and $h^*_{\mu}(V)$ and $h^*_{\boldsymbol{\theta}}(V)$ are in (6.4) and (6.5), respectively. We obtain a contradiction that $b = 0$ with similar argument in the proof of Lemma$A$.1.3.

Finally, from (6.14), we obtain that

$$\begin{aligned} n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) &= -\left[P\left\{\ddot{l}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \Lambda_0) - \ddot{l}_{\boldsymbol{\theta}\Lambda}(\boldsymbol{\theta}_0, \Lambda_0)[h^*]\right\}\right]^{-1}\mathbb{G}_n\left\{\dot{l}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \Lambda_0) - \dot{l}_{\Lambda}(\boldsymbol{\theta}_0, \Lambda_0)[h^*]\right\} \\ &\quad + o_p(1). \end{aligned}$$

Therefore, $n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges to a normal distribution and has influence function given by

$$\left[P\left\{\dot{l}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \Lambda_0) - \dot{l}_{\boldsymbol{\theta}\Lambda}(\boldsymbol{\theta}_0, \Lambda_0)[h^*]\right\}\right]^{-1}\left\{\dot{l}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \Lambda_0) - \dot{l}_{\Lambda}(\boldsymbol{\theta}_0, \Lambda_0)[h^*]\right\}.$$

Because this influence function is on the linear space spanned by the score functions $\dot{l}_{\boldsymbol{\theta}}$ and $\dot{l}_{\Lambda}[h]$, the influence function is the same as the efficient influence function for $\boldsymbol{\theta}_0$. Hence the asymptotic variance of $n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ attains the semiparametric efficiency bound. □

Proof of Theorem 3.4.4. First, by the uniform convergence of $(\widehat{\theta}, \widehat{\Lambda})$ almost surely, we conclude that $\widehat{E}(\kappa|\boldsymbol{W})$ and $\widehat{E}(\kappa\xi|\boldsymbol{W})$ converges to $E(\kappa|\boldsymbol{W})$ and $E(\kappa\xi|\boldsymbol{W})$ uniformly in $\boldsymbol{W}$ with probability one. Therefore,

$$\sup_v \left| \frac{\sum_{i=1}^n K_{h_n}(V_i - v)\widehat{E}(\kappa|\boldsymbol{W}_i)}{\sum_{i=1}^n K_{h_n}(V_i - v)} - \frac{\sum_{i=1}^n K_{h_n}(V_i - v)E(\kappa|\boldsymbol{W}_i)}{\sum_{i=1}^n K_{h_n}(V_i - v)} \right| \to 0$$

where $K_{h_n}(x) = h_n^{-1}\exp\{-x^2/h_n\}$. On the other hand, following the general results in Hansen (2008), under the conditions for h_n, we obtain

$$\sup_v \left| \frac{\sum_{i=1}^n K_{h_n}(V_i - v)E(\kappa|\boldsymbol{W}_i)}{\sum_{i=1}^n K_{h_n}(V_i - v)} - E[E(\kappa|\boldsymbol{W})|V = v] \right| \to 0$$

with probability one. Similarly, we can show the uniform convergence of $\widehat{E}(\boldsymbol{X}E(\kappa|\boldsymbol{W})^2|V = v)$ to $E(\boldsymbol{X}E(\kappa|\boldsymbol{W})^2|V = v)$ and the uniform convergence of $\widehat{E}(E(\kappa|\boldsymbol{W})\{\mu^{-1} - E(\kappa\xi|\boldsymbol{W})\}|V = v)$ to $E(E(\kappa|\boldsymbol{W})\{\mu^{-1} - E(\kappa\xi|\boldsymbol{W})\}|V = v)$. Consequently, $\widehat{R}_1(v)$ and $\widehat{R}_2(v)$ converge to $R_1(v)$ and $R_2(v)$ respectively and uniformly in v. This immediately gives that $\widehat{l}^*_{i\theta}$ converges uniformly in $\boldsymbol{W}_i$ to $l^*(\boldsymbol{\theta})$. Thus, the result of this theorem holds. □

APPENDIX B: Technical Details for Chapter 4

B.1 Identifiability and Derivation of Pseudo-Efficient Score Functions

Lemma B.1.1. *The model in (4.3) is identifiable, that is, if the probability at Θ_1 is equal to that at Θ_2, then $\Theta_1 = \Theta_2$.*

Proof. We first show model identifiability in the case with only one observation time. Suppose that two likelihood functions with $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$ are the same:

$$\begin{aligned}&\int_{-\infty}^{\infty} \exp\left\{-\Lambda_1(t)e^{\boldsymbol{\beta}_1^T\boldsymbol{X}-\mu_1\xi}\right\}\Lambda_1(t)\mu_1 e^{\boldsymbol{\beta}_1^T\boldsymbol{X}-\mu_1\xi}\frac{1}{\sigma}\phi\left\{\frac{Y(t)-\xi}{\sigma}\right\}d\xi\\ &= \int_{-\infty}^{\infty} \exp\left\{-\Lambda_2(t)e^{\boldsymbol{\beta}_2^T\boldsymbol{X}-\mu_2\xi}\right\}\Lambda_2(t)\mu_2 e^{\boldsymbol{\beta}_2^T\boldsymbol{X}-\mu_2\xi}\frac{1}{\sigma}\phi\left\{\frac{Y(t)-\xi}{\sigma}\right\}d\xi. \qquad (6.15)\end{aligned}$$

For any $\aleph = (Y(v), v, \boldsymbol{X})$, the models in (6.15) can be simplified as follows: For $k = 1, 2$,

$$\int_{-\infty}^{\infty} \exp\left\{-\Lambda_k(v)e^{\boldsymbol{\beta}_k^T\boldsymbol{X}-\mu_k\xi}\right\}\Lambda_k(v)\mu_k e^{\boldsymbol{\beta}_k^T\boldsymbol{X}-\mu_k\xi-\frac{\xi^2}{2\sigma^2}}\exp\left\{-\frac{Y(v)\xi}{\sigma^2}\right\}d\xi. \qquad (6.16)$$

Since $Y(v)$ can be an arbitrary real number, the integration in (6.16) is a bilateral Laplace transformation of $f_{kv}(\xi)$, where

$$f_{kv}(\xi) = \exp\left\{-\Lambda_k(v)e^{\boldsymbol{\beta}_k^T\boldsymbol{X}-\mu_k\xi}\right\}\Lambda_k(v)\mu_k\exp\{\boldsymbol{\beta}_k^T\boldsymbol{X}-\mu_k\xi-\xi^2/(2\sigma^2)\}, \quad k = 1, 2. \quad (6.17)$$

This implies $f_{1v} = f_{2v}$ by the one-to-one property of Laplace transformation. Thus, for any ξ,

$$\begin{aligned}&-\Lambda_1(v)e^{\boldsymbol{\beta}_1^T\boldsymbol{X}-\mu_1\xi} + \log\{\Lambda_1(v)\} + \log\mu_1 + \boldsymbol{\beta}_1^T\boldsymbol{X} - \mu_1\xi\\ &= -\Lambda_2(v)e^{\boldsymbol{\beta}_2^T\boldsymbol{X}-\mu_2\xi} + \log\{\Lambda_2(v)\} + \log\mu_2 + \boldsymbol{\beta}_2^T\boldsymbol{X} - \mu_2\xi. \qquad (6.18)\end{aligned}$$

Examining the behavior when $\xi \to -\infty$, we obtain $\mu_1 = \mu_2$ and $\Lambda_1(v)e^{\boldsymbol{\beta}_1^T \boldsymbol{X}} = \Lambda_2(v)e^{\boldsymbol{\beta}_2^T \boldsymbol{X}}$. So by (A2), $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ and $\Lambda_1(v) = \Lambda_2(v)$. Hence, we proved the model identifiability in case with only one observation.

Now suppose that the number of observation times is $J > 1$, and

$$\begin{aligned}
&\prod_{j=1}^{J} \int_{-\infty}^{\infty} \exp\left\{-\Lambda_1(v_j)e^{\boldsymbol{\beta}_1^T \boldsymbol{X} - \mu_1 \xi_j}\right\} \Lambda_1(v_j)\mu_1 e^{\boldsymbol{\beta}_1^T \boldsymbol{X} - \mu_1 \xi_j} \frac{1}{\sigma}\phi\left\{\frac{Y(v_j) - \xi_j}{\sigma}\right\} d\xi_j \\
&= \prod_{j=1}^{J} \int_{-\infty}^{\infty} \exp\left\{-\Lambda_2(v_j)e^{\boldsymbol{\beta}_2^T \boldsymbol{X} - \mu_2 \xi_j}\right\} \Lambda_2(v_j)\mu_2 e^{\boldsymbol{\beta}_2^T \boldsymbol{X} - \mu_2 \xi_j} \frac{1}{\sigma}\phi\left\{\frac{Y(v_j) - \xi_j}{\sigma}\right\} d\xi_j \quad (6.19)
\end{aligned}$$

By the one-to-one property of the Laplace bilateral transformation, the equation in (6.19) implies that $\prod_{j=1}^{J} f_{1v_j}(\xi_j) = \prod_{j=1}^{J} f_{2v_j}(\xi_j)$ for any $\xi_1, \ldots, \xi_J$.

Then this leads to $\sum_{j=1}^{J} \log\{f_{1v_j}(\xi_j)\} = \sum_{j=1}^{J} \log\{f_{2v_j}(\xi_j)\}$ for any $\xi_1, \ldots, \xi_J$. Supposing that the ξ's are all 0 except ξ_j and examining the behavior when $\xi_j \to -\infty$, we complete the proof by the same argument as in the case with only one observation. □

We denote the log-likelihood at time v by

$$l(\boldsymbol{\theta}, \Lambda(v) \mid \aleph) = \log\left[\int_{-\infty}^{\infty} - \exp\left\{-\Lambda(v)e^{\boldsymbol{\beta}^T \boldsymbol{X} - \mu\xi}\right\} \Lambda(v)\mu e^{\boldsymbol{\beta}^T \boldsymbol{X} - \mu\xi} \frac{1}{\sigma}\phi\left\{\frac{Y(v) - \xi}{\sigma}\right\} d\xi\right]$$

hereafter. Then, the log pseudo-likelihood is re-expressed as $l^{ps} = \int l\{\boldsymbol{\theta}, \Lambda(v) \mid \aleph\} dN(v)$.

Lemma B.1.2. *The linear operator at time v, $E\left[\left\{\dot{l}_\theta(v), \dot{l}_\Lambda(v)\right\}^* \left\{\dot{l}_\theta(v), \dot{l}_\Lambda(v)\right\}\right]$, which maps $\Theta \times H$ to the dual space of $\Theta \times H$ (equivalent to $\Theta \times H$), is continuously invertible at $(\theta_0, \Lambda_0(v))$, where $\dot{l}^*_{\boldsymbol{\theta}}(v)$ and $\dot{l}^*_\Lambda(v)$ are the adjoint operators of the linear operators at time v, $\dot{l}_{\boldsymbol{\theta}}(v)$ and $\dot{l}_\Lambda(v)$, respectively. Here, $\dot{l}_\theta(v) = \partial l(v)/\partial\theta$.*

Proof. It suffices to show that $E\left\{\dot{l}^*_\Lambda(v)\dot{l}_\Lambda(v)\right\}$, and
$E\left\{\dot{l}^*_{\boldsymbol{\theta}}(v)\dot{l}_{\boldsymbol{\theta}}(v)\right) - E(\dot{l}^*_{\boldsymbol{\theta}}(v)\dot{l}_\Lambda(v)\right\} E\left\{\dot{l}^*_\Lambda(v)\dot{l}_\Lambda(v)\right\}^{-1} E\left\{\dot{l}^*_\Lambda(v)\dot{l}_{\boldsymbol{\theta}}(v)\right\}$ (denoted as the matrix

$A(v)$) are invertible. By taking linear operator of $\dot{l}_\Lambda$, we obtain

$$E(\dot{l}^*_\Lambda(v)\dot{l}_\Lambda(v)[h,\widetilde{h}]) = E\left[\tilde{h}(v)h(v)\Lambda(v)^{-2}E\left\{E(\kappa(v)\mid \aleph)^2 \mid v\right\}\right].$$

Therefore, $E\left\{\dot{l}^*_\Lambda(v)\dot{l}_\Lambda(v)[h]\right\} = h(v)\Lambda(v)^{-2}E\left\{E(\kappa(v)\mid \aleph)^2 \mid v\right\}$ and so is invertible from $L_2(P_V)$ to $L_2(P_V)$.

If $A(v)$ is singular, then there exists some non-zero vector $\boldsymbol{b}$ such that $\boldsymbol{b}^T A(v)\boldsymbol{b} = 0$. Therefore, $E\{(\boldsymbol{b}^T\dot{l}_{\boldsymbol{\theta}} - \dot{l}_\Lambda[h])^{\otimes 2}\} = 0$, where $h = \boldsymbol{b}^T E[\dot{l}^*_\Lambda\dot{l}_\Lambda]^{-1}E(\dot{l}^*_\Lambda\dot{l}_{\boldsymbol{\theta}})$. As a result, $\boldsymbol{b}^T\dot{l}_{\boldsymbol{\theta}} - \dot{l}_\Lambda[h] = 0$ almost surely. We obtain

$$E\left[\left\{\boldsymbol{b}^T(-\xi,\boldsymbol{X}) - h(v)/\Lambda_0(v)\right\}\left\{1-\Lambda_0(v)e^{\boldsymbol{\beta}_0^T\boldsymbol{X}-\mu_0\xi}\right\} + b_1/\mu_0 \mid \aleph\right] = 0,$$

where b_1 is the component of $\boldsymbol{b}$ corresponding to μ. The left-hand side can be treated as a Laplace transformation of some function of ξ so we immediately conclude

$$\left\{\boldsymbol{b}^T(-\xi,\boldsymbol{X}) - h(v)/\Lambda_0(v)\right\}\left\{1-\Lambda_0(v)e^{\beta_0^T\boldsymbol{X}-\mu_0\xi}\right\} + b_1/\mu_0 = 0.$$

Since ξ is arbitrary, $b_1 = 0$, $\boldsymbol{b}_{-1}^T\boldsymbol{X} = 0$, $h(v)/\Lambda_0(v) = 0$, where $\boldsymbol{b}_{-1} = (b_2,\ldots,b_d)$. Therefore, $\boldsymbol{b} = 0$, and this leads to a contradiction. □

B.2 Proof of Asymptotic Results

Proof of Theorem 4.3.1. From the model (4.3), we let

$$\mathscr{H} = \left\{l^{ps}(\boldsymbol{\theta},\Lambda,\aleph) \mid \boldsymbol{\theta}\in\boldsymbol{\Theta},\Lambda(t)\in\boldsymbol{\Phi}^*\right\}, \tag{6.20}$$

where the parameter space, $\boldsymbol{\Phi}^* = \{\Lambda(t) \mid \Lambda(t) = -\log S(t)$, $S(t)$ is a non-increasing function with $S(0) = 1$, $S(t)\geq 0\}$. If the following conditions are satisfied, we prove

asymptotic consistency of $(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda})$ by Theorem 2.12 (Kosorok 2008).

(a) For any sequence $\{(\boldsymbol{\theta}_n, \Lambda_n)\} \in \boldsymbol{\Theta} \times \boldsymbol{\Phi}^*, \underline{\lim}_{n\to\infty} l^{ps}(\boldsymbol{\theta}_n, \Lambda_n \mid \boldsymbol{\aleph}) \geq l^{ps}(\boldsymbol{\theta}_0, \Lambda_0 \mid \boldsymbol{\aleph})$ implies $d\left((\boldsymbol{\theta}_n, \Lambda_n), (\boldsymbol{\theta}_0, \Lambda_0)\right) \to 0$,

(b) $\mathbb{P}_n l^{ps}(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda}) = \sup_{(\boldsymbol{\theta},\Lambda)\in\boldsymbol{\Theta}\times\boldsymbol{\Phi}^*} \mathbb{P}_n l^{ps}(\boldsymbol{\theta}, \Lambda) - o_p(1)$,

(c) $\sup_{(\boldsymbol{\theta},\Lambda)\in\boldsymbol{\Theta}\times\boldsymbol{\Phi}^*} |\mathbb{P}_n l^{ps}(\boldsymbol{\theta}, \Lambda) - P l^{ps}(\boldsymbol{\theta}, \Lambda)| \to 0$ in probability, as $n \to \infty$.

Condition (a) is satisfied by identifiability of the marginal likelihood proved in Lemma B.1.1, and condition (b) is satisfied because $(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda})$ is the MLE. For the last condition, we calculate the bracket covering number for the class in (6.20). For any $(\mu_1, \boldsymbol{\beta}_1, \Lambda_1)$, $(\mu_2, \boldsymbol{\beta}_2, \Lambda_2) \in \boldsymbol{\Theta}\times\boldsymbol{\Phi}^*$ such that $\sup_{v\in\mathscr{I}[V]}|\Lambda_1(v)-\Lambda_2(v)| < \epsilon$, $|\mu_1-\mu_2| < \epsilon$, and $|\boldsymbol{\beta}_1-\boldsymbol{\beta}_2| < \epsilon$, for $\epsilon > 0$, we wish to set boundaries for the bracket covering number for the class $\mathscr{H}$. There exists a positive constant C_1 such that $|(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)^T \boldsymbol{X} - (\mu_1 - \mu_2)\xi| \leq C_1\epsilon + \epsilon|\xi|$. Then for some positive constants, C_2,C_3, and C_4,

$$
\begin{aligned}
&|l^{ps}(\mu_1, \boldsymbol{\beta}_1, \Lambda_1, \boldsymbol{\aleph}) - l^{ps}(\mu_2, \boldsymbol{\beta}_2, \Lambda_2, \boldsymbol{\aleph})| \\
&\leq \sum_{v:dN(v)=1} |l\{\mu_1, \boldsymbol{\beta}_1, \Lambda_1(v), \boldsymbol{\aleph}\} - l\{\mu_2, \boldsymbol{\beta}_2, \Lambda_2(v), \boldsymbol{\aleph}\}|, \\
&\leq \sum_{v:dN(v)=1} \int \{C_2|\Lambda_1(v) - \Lambda_2(v)| + C_3\epsilon + C_4\epsilon|\xi|\}\, \phi\left\{\frac{Y(v) - \xi}{\sigma}\right\} d\xi.
\end{aligned}
$$

Thus, $\|l^{ps}(\mu_1, \boldsymbol{\beta}_1, \Lambda_1 \mid \boldsymbol{\aleph}) - l^{ps}(\mu_2, \boldsymbol{\beta}_2, \Lambda_2 \mid \boldsymbol{\aleph})\|_{1,P} \leq C_2'\epsilon + C_3'\epsilon + C_4'\epsilon$ for some positive constants, C_2', C_3', and C_4'. We obtain that $\log N_{[\,]}(O(1)\epsilon, \mathscr{H}, L_1(P)) \leq \log N_{[\,]}(\epsilon, \boldsymbol{\Theta} \times \boldsymbol{\Phi}^*, \|\cdot\|_{l^\infty}) \leq O(1/\epsilon)$. Hence, $\mathscr{H}$ is P-Glivenko-Cantelli class, and Theorem 3.4.1 is proved. □

Once consistency of $\widehat{\boldsymbol{\theta}}$ is established, we can concentrate on a neighborhood of $\boldsymbol{\theta}_0$. For any $\eta > 0$, let $B(\boldsymbol{\theta}_0, \eta)$ be the ball with radius η centered at $\boldsymbol{\theta}_0$. If $\boldsymbol{\theta}_0$ is on the

boundary of $\boldsymbol{\Theta}$, then take $B(\boldsymbol{\theta}_0, \eta) \cap \boldsymbol{\Theta}$ instead of $B(\boldsymbol{\theta}_0, \eta)$. Then $B(\boldsymbol{\theta}_0, \eta)$ is included in $\boldsymbol{\Theta}$. We suppose that condition (A3) is satisfied so that Λ_0 is bounded and away from 0 on $\mathscr{S}[V]$. Since we have proved that $\widehat{\Lambda}$ converges on $\mathscr{S}[V]$, we may restrict $\widehat{\Lambda}$ to the following class of functions:

$$\Phi = \{\Lambda \mid \Lambda \text{ is non-decreasing and } 0 < 1/M \leq \Lambda(t) \leq M < \infty \text{ for all } t \in \mathscr{S}[V]\}, \quad (6.21)$$

where M is a large positive constant. For any probability measure P, define $L_2(P) = \left\{g \mid \int g^2 dP < \infty\right\}$. Let $\|\cdot\|_{2,P}$ be the usual $L_{2,P}$ norm. For any subclass $\mathscr{F}$ of $L_2(P)$, define the bracketing number $N_{[\,]}(\epsilon, \mathscr{F}, L_2(P))$ as the infimum of the cardinal numbers for $\left\{g_i^L, g_i^U \mid g_i^L \leq g \leq g_i^U, g \in \mathscr{F}, \text{ for some } i, \text{ and } \|g_i^U - g_i^L\| \leq \epsilon\right\}$.

By the following lemma, we determine the size of the class for marginal likelihood functions of interest.

Lemma B.2.1. *Let*

$$\mathscr{H} = \{l^{ps}(\boldsymbol{\theta}, \Lambda, \aleph(t)) \mid \boldsymbol{\theta} \in B(\boldsymbol{\theta}_0, \eta), \Lambda \in \Phi\}. \quad (6.22)$$

Suppose that (A2) is satisfied. Then there exists a constant $C > 0$ such that

$$\sup_Q N_{[\,]}(\epsilon, \mathscr{H}, L_2(Q)) \leq C(1/\epsilon^d)e^{1/\epsilon} \quad \textit{for all} \quad \epsilon > 0,$$

where d is the dimension of $\boldsymbol{\theta}$. Hence, for ϵ small enough, we have

$$\sup_Q \log N_{[\,]}(\epsilon, \mathscr{H}, L_2(Q)) \leq C(1/\epsilon).$$

Here Q runs through the class of all probability measures.

The proof of Lemma B.2.1 is adapted from Huang (1996), where the author provided

the order of the entropy for a class of log-likelihood functions over bounded parameter space in current status data.

Proof of Lemma B.2.1. We first calculate the order of the bracketing number for class $\mathscr{H}'$, where $\mathscr{H}'=\{f \,|\, \log f \in \mathscr{H} \text{ and } \int_0^\infty dN(V) = 1\}$. It is known that for the class of functions

$$\Phi = \left\{\Lambda \middle| \Lambda \text{ is non-decreasing and } 0 < 1/M \le \Lambda(t) \le M < \infty \text{ for all } t \in \mathscr{S}[V]\right\},$$

where M is a constant, its ϵ bracketing number is of the order of $m = N_{[\,]}(\epsilon, \Phi, L_2(P)) = O(e^{1/\epsilon})$. This means that the class Φ has finite entropy. Let $\Lambda_i^{*L} = \Lambda_i^L - \epsilon$ and $\Lambda_i^{*U} = \Lambda_i^U + \epsilon$ for $1 \le i \le m$. Then for any $\Lambda \in \Phi$, we have, for some i, $\Lambda_i^{*L} + \epsilon \le \Lambda \le \Lambda_i^{*U} - \epsilon$ and $\|\Lambda_i^{*U} - \Lambda_i^{*L}\|_{2,P_V} \le 3\epsilon$. Since Φ is uniformly bounded away from 0, we can choose ϵ small enough such that all the bracketing functions stay away from 0.

For the true $\mu_0 > 0$, we can find a constant δ_0 such that $\mu_0 > \delta_0 > 0$. Related to $\boldsymbol{\theta}$, we can also choose k points $(\boldsymbol{\beta}_1, \mu_1) = \boldsymbol{\theta}_1, \ldots, (\boldsymbol{\beta}_k, \mu_k) = \boldsymbol{\theta}_k$ in $B(\boldsymbol{\theta}_0, \eta)$ such that for any $(\boldsymbol{\beta}, \mu) \in B(\boldsymbol{\theta}_0, \eta)$, $\left|\boldsymbol{\beta}_j - \boldsymbol{\beta}\right| < \delta_1$ and $\left|\mu_j - \mu\right| < \delta_2$ for given constants $\delta_1 > 0$ and $0 < \delta_2 < \delta_0$, since $\boldsymbol{\Theta}_0$ is compact by (A1). Then for any $(\boldsymbol{\beta}, \mu) \in B(\boldsymbol{\theta}_0, \eta)$ and for some $1 \le j \le k$, there exists a positive constant C_1 such that $\left|(\boldsymbol{\beta}_j - \boldsymbol{\beta})^T \boldsymbol{X} - (\mu_j - \mu)\xi\right| \le C_1\delta_1 + \delta_2\left|\xi\right|$.

Using the chosen k regression parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$ and m cumulative hazard functions $\{\Lambda_i^{*L}, \Lambda_i^{*U}\}_{i=1}^m$, we will show how to construct upper and lower envelope functions for the log marginal likelihood functions belonging to (6.22). For any $(\boldsymbol{\theta}, \Lambda) \in B(\boldsymbol{\theta}_0, \eta) \times \Phi$, we can choose $\boldsymbol{\theta}_j$,$\{\Lambda_i^{*L}(v), \Lambda_i^{*U}(v)\}$ satisfying

$$\begin{aligned}
& \exp\left\{-\Lambda_i^{*U}(v) e^{\boldsymbol{\beta}_j^T \boldsymbol{X} + C_1\delta_1 - \mu_j\xi + \delta_2|\xi|}\right\} \Lambda_i^{*L}(v)(\mu_j - \delta_2|\xi|) e^{\boldsymbol{\beta}_j^T \boldsymbol{X} - C_1\delta_1 - \mu_j\xi - \delta_2|\xi|} \\
\le\ & \exp\left\{-\Lambda(v) e^{\boldsymbol{\beta}^T \boldsymbol{X} - \mu\xi}\right\} \Lambda(v)\mu e^{\boldsymbol{\beta}^T \boldsymbol{X} - \mu\xi} \\
\le\ & \exp\left\{-\Lambda_i^{*L}(v) e^{\boldsymbol{\beta}_j^T \boldsymbol{X} - C_1\delta_1 - \mu_j\xi - \delta_2|\xi|}\right\} \Lambda_i^{*U}(v)(\mu_j + \delta_2|\xi|) e^{\boldsymbol{\beta}_j^T \boldsymbol{X} + C_1\delta_1 - \mu_j\xi + \delta_2|\xi|}.
\end{aligned}$$

It is well known that the minimum value of k can be on the order of $O(\epsilon^{-d})$.

Then we let

$$\begin{aligned} f_{ij}^{*L} &= \int \exp\left\{-\Lambda_i^{*U}(v)e^{\boldsymbol{\beta}_j^T\boldsymbol{X}+C_1\delta_1-\mu_j\xi+\delta_2|\xi|}\right\}\Lambda_i^{*L}(v)(\mu_j-\delta_2|\xi|)e^{\boldsymbol{\beta}_j^T\boldsymbol{X}-C_1\delta_1-\mu_j\xi-\delta_2|\xi|} \\ &\quad\times\phi\left\{\frac{Y(v)-\xi}{\sigma}\right\}d\xi, \qquad (6.23) \\ f_{ij}^{*U} &= \int \exp\left\{-\Lambda_i^{*L}(v)e^{\boldsymbol{\beta}_j^T\boldsymbol{X}-C_1\delta_1-\mu_j\xi-\delta_2|\xi|}\right\}\Lambda_i^{*U}(v)(\mu_j+\delta_2|\xi|)e^{\boldsymbol{\beta}_j^T\boldsymbol{X}+C_1\delta_1-\mu_j\xi+\delta_2|\xi|} \\ &\quad\times\phi\left\{\frac{Y(v)-\xi}{\sigma}\right\}d\xi, \qquad (6.24) \end{aligned}$$

so f_{ij}^{*L} and f_{ij}^{*U} are finite envelope functions for $f(\boldsymbol{\theta},\Lambda \mid \aleph) \in \mathscr{H}'$ and any time v. Finally, we need to show that $\|f_{ij}^{*U}-f_{ij}^{*L}\|_{2,P}$ can be small enough to be less than an arbitrary constant ϵ. For some constant C_2, C_3, and C_4,

$$\left|f_{ij}^{*U}-f_{ij}^{*L}\right| \le \int \left(C_2\left|\Lambda^{*U}(v)-\Lambda^{*L}(v)\right|+C_3\delta_1+C_4\delta_2|\xi|\right)\phi\Big(\frac{Y(v)-\xi}{\sigma}\Big)d\xi.$$

Thus,

$$\begin{aligned} \|f_{ij}^{*U}-f_{ij}^{*L}\|_{2,P} &\le C_2\left\|\Lambda^{*U}(v)-\Lambda^{*L}(v)\right\|_{2,P_V}+C_3\delta_1+C_4'\delta_2 \\ &\le 3C_2\epsilon+C_3\delta_1+C_4'\delta_2, \end{aligned}$$

for some constant C_4'. This implies that there exist f_{ij}^{*L}, f_{ij}^{*U}, $i=1,\ldots,m$ and $j=1,\ldots,k$ such that, for any $f \in \mathscr{H}'$. $f_{ij}^{*L} \le f \le f_{ij}^{*U}$, for some $1 \le i \le m$, $1 \le j \le k$, and $\|f_{ij}^{*U}-f_{ij}^{*L}\|_{2,P} \le 3C_2\epsilon+C_3\delta_1+C_4\delta_2'$. This means that the bracketing number $N_{[\,]}(\epsilon,\mathscr{H}',L_2(P))$ for the class $\mathscr{H}'$ is of order $mk=O(\epsilon^{-d}e^{1/\epsilon})$. Note that the log function on the domain bounded away from 0 is Lipschitz continuous, and any function $f \in \mathscr{H}'$ is bounded and away from 0; the bracket, $\left[f_{ij}^{*L},f_{ij}^{*U}\right]$ for $f \in \mathscr{H}'$ covers f at any $v \in \mathscr{S}[V]$, that is, $\int_0^{u_V}\log f_{ij}^{*L}dN(v) \le \sum_{v:dN(v)=1}\log f(v) \le \int_0^{u_V}\log f_{ij}^{*U}dN(v)$. Hence, by the assumptions

(A2), (A3), and (A4), the bracketing number of $\mathscr{H}$ is dominated by the bracketing number of the class $\mathscr{H}'$. □

Proof of Theorem 4.3.2. We apply Theorem 3.4.1 of van der Vaart and Wellner (1996) to the prove the results. Specifically, we need to check the following conditions: Let $0 \le \delta < \eta$ be arbitrary and C be a generic constant, then for $\delta < d\left((\boldsymbol{\theta}, \boldsymbol{\theta}_0), (\Lambda, \Lambda_0)\right) \le \eta$,

$$
\begin{aligned}
&(i) \quad \sup_{\delta/2 < d\{(\boldsymbol{\theta},\boldsymbol{\theta}_0),(\Lambda,\Lambda_0)\} \le \delta, \boldsymbol{\theta} \in \Theta_0} P\left\{ l^{ps}(\boldsymbol{\theta}, \Lambda, \boldsymbol{\aleph}) - l^{ps}(\boldsymbol{\theta}_0, \Lambda_0, \boldsymbol{\aleph}) \right\} \le -\delta^2, \\
&(ii) \quad E^* \sup_{\delta/2 < d\{(\boldsymbol{\theta},\boldsymbol{\theta}_0),(\Lambda,\Lambda_0)\} \le \delta, \boldsymbol{\theta} \in \boldsymbol{\Theta}_0} n^{1/2} |(\mathbb{P}_n - P)\{ l^{ps}(\boldsymbol{\theta}, \Lambda, \boldsymbol{\aleph}) - l^{ps}(\boldsymbol{\theta}_0, \Lambda_0, \boldsymbol{\aleph}) \}| \le C\psi(\delta),
\end{aligned}
$$

for function ψ such that $\delta \to \psi(\delta)/\delta^{\alpha}$ is increasing on (δ, η) for some $\alpha < 2$.

For the first condition, we perform the Taylor expansion to obtain

$$
\begin{aligned}
P\left\{ l^{ps}(\boldsymbol{\theta}_0, \Lambda_0) - l^{ps}(\boldsymbol{\theta}, \Lambda) \right\} &= P\Big[\int l\left\{ (1-\epsilon)\boldsymbol{\theta}_0 + \epsilon\boldsymbol{\theta}, (1-\epsilon)\Lambda_0 + \epsilon\Lambda \right\} dN(v) \Big]\Big|_1^0 \\
&= -\frac{1}{2} \frac{\partial^2}{\partial \epsilon^2}\Big|_{\epsilon=\epsilon^*} P\left[\int l\left\{ (1-\epsilon)\boldsymbol{\theta}_0 + \epsilon\theta, (1-\epsilon)\Lambda_0 + \epsilon\Lambda \right\} dN(v) \right],
\end{aligned}
$$

for some $\epsilon^* \in (0,1)$. For η small enough, we note that right-hand side is equal to

$$
P \int \left\{ E\left[\left\{ \dot{l}_\theta(v), \dot{l}_\Lambda(v) \right\}^* \left\{ \dot{l}_\theta(v), \dot{l}_\Lambda(v) \right\} \right] + o(1) \right\} (\boldsymbol{\theta} - \boldsymbol{\theta}_0, \Lambda - \Lambda_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0, \Lambda - \Lambda_0) dN(v),
$$

where $E\left[\left\{ \dot{l}_\theta(v), \dot{l}_\Lambda(v) \right\}^* \left\{ \dot{l}_\theta(v), \dot{l}_\Lambda(v) \right\} \right]$ is the linear operator in Lemma B.1.2.

Since the linear operator is an invertible linear operator and uniformly bounded and away from 0, for some constant C,

$$
\left\| E\left[\left\{ \dot{l}_\theta(v), \dot{l}_\Lambda(v) \right\}^* \left\{ \dot{l}_\theta(v), \dot{l}_\Lambda(v) \right\} \right] (a,h)(a,h) \right\| \ge C\{ |a|^2 + \|h\|_{2,P_V}^2 \}.
$$

Hence, for some constant C,

$$\begin{aligned} & P\int\left\{E\left[\left\{\dot{l}_\theta(v),\dot{l}_\Lambda(v)\right\}^*\left\{\dot{l}_\theta(v),\dot{l}_\Lambda(v)\right\}\right]+o(1)\right\}(\boldsymbol{\theta}-\boldsymbol{\theta}_0,\Lambda-\Lambda_0)(\boldsymbol{\theta}-\boldsymbol{\theta}_0,\Lambda-\Lambda_0)dN(v) \\ \geq\ & C(|\boldsymbol{\theta}-\boldsymbol{\theta}_0|^2+\|\Lambda-\Lambda_0\|^2_{2,P_V}). \end{aligned}$$

Thus, condition (i) holds.

For the second condition, by Lemma$B$.2.1, for some constants C and M,

$$J_{[\,]}(\eta,\mathscr{H},L_2(P))\leq\int_0^\eta\sqrt{1+C\epsilon^{-1}}d\epsilon\leq\int_0^\eta M\epsilon^{-1/2}d\epsilon=M\eta^{1/2}.$$

Then, according to Lemma 3.4.2 of van der Vaart and Wellner (1996),

$$E^*\sup_{d\{(\boldsymbol{\theta},\Lambda),(\boldsymbol{\theta}_0,\Lambda_0)\}\leq\eta}\left|n^{1/2}(P_n-P)\left\{l^{ps}(\boldsymbol{\theta},\Lambda\mid\aleph)-l^{ps}(\boldsymbol{\theta}_0,\Lambda_0\mid\aleph)\right\}\right|=O(1)\eta^{1/2}\left(1+\frac{\eta^{1/2}}{\eta^2n^{1/2}}M\right).$$

Finally, let

$$\psi(\eta)=\eta^{1/2}\left(1+\frac{\eta^{1/2}}{\eta^2n^{1/2}}\right).$$

Then $\phi(\delta)/\delta^\alpha$ is an increasing function for some $0<\alpha<1/2$, so the condition (ii) is satisfied. In addition, since $(\widehat{\boldsymbol{\theta}},\widehat{\Lambda})$ maximizes $l^{ps}(\boldsymbol{\theta},\Lambda)$, $Pl^{ps}(\widehat{\boldsymbol{\theta}},\widehat{\Lambda},\aleph)\geq Pl^{ps}(\boldsymbol{\theta}_0,\Lambda_0,\aleph)$ is also satisfied. When $r_n=n^{1/3}$, then $n^{2/3}\psi(n^{-1/3})=O(n^{1/2})$ for every n.

Hence, all the conditions of Theorem 3.4.1 of van der Vaart and Wellner (1996) are satisfied. This implies

$$d\{(\widehat{\boldsymbol{\theta}},\widehat{\Lambda}),(\boldsymbol{\theta}_0,\Lambda_0)\}=O_p(n^{-1/3}).$$

□

Proof of Theorem 4.3.3. We will prove that asymptotic distribution of the MLE, $\widehat{\boldsymbol{\theta}}$, is normal distribution with mean 0 and variance in (4.14) by following the approach on page 1007 of Zeng, Yin, and Ibrahim (2005). For simplicity of notation, $\dot{l}_{\boldsymbol{\theta}}(\boldsymbol{\theta},\Lambda,\aleph)$ and

$\dot{l}^{ps}_{\Lambda}(\boldsymbol{\theta}, \Lambda, \aleph)$ are denoted by $\dot{l}^{ps}_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \Lambda)$ and $\dot{l}^{ps}_{\Lambda}(\boldsymbol{\theta}, \Lambda)$, respectively.

Since $(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda})$ are maximum likelihood estimators for $(\boldsymbol{\theta}, \Lambda)$, we immediately obtain that $\mathbb{P}_n\{\dot{l}^{ps}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda}) - \dot{l}^{ps}_{\Lambda}(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda})[h^*]\} = 0$. Thus,

$$\mathbb{G}_n\left\{\dot{l}^{ps}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda}) - \dot{l}^{ps}_{\Lambda}(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda})[h^*]\right\} = -n^{1/2}P\left\{\dot{l}^{ps}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda}) - \dot{l}^{ps}_{\Lambda}(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda})[h^*]\right\},$$

where $\mathbb{G}_n = n^{1/2}(\mathbb{P}_n - P)$.

Let us consider the following two classes of functions:

$$\begin{aligned}
&\left\{\dot{l}_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \Lambda) - \dot{l}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \Lambda_0) \;\;\middle|\;\; |\boldsymbol{\theta} - \boldsymbol{\theta}_0| \le \eta \text{ and } \|\Lambda - \Lambda_0\|_{2,P_V} \le \eta\right\} \quad \text{and} \\
&\left\{\dot{l}_{\Lambda}(\boldsymbol{\theta}, \Lambda)[h^*] - \dot{l}_{\Lambda}(\boldsymbol{\theta}_0, \Lambda_0)[h^*] \;\;\middle|\;\; |\boldsymbol{\theta} - \boldsymbol{\theta}_0| \le \eta \text{ and } \|\Lambda - \Lambda_0\|_{2,P_V} \le \eta\right\},
\end{aligned}$$

where η is near 0. The entropy numbers for the two classes are of order $1/\eta$ and this implies that these two classes are P-Donsker. Hence, $\dot{l}^{ps}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda}) - \dot{l}^{ps}_{\Lambda}(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda})[h^*]$ belongs to a P-Donsker class. This leads to

$$\mathbb{G}_n\left\{\dot{l}^{ps}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \Lambda_0) - \dot{l}_{\Lambda}(\boldsymbol{\theta}_0, \Lambda_0)[h^*]\right\} + o_p(1) = -n^{1/2}P\left\{\dot{l}^{ps}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda}) - \dot{l}^{ps}_{\Lambda}(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda})[h^*]\right\}. \quad (6.25)$$

We perform a Taylor's series expansion of the right side of (6.25) at $(\boldsymbol{\theta}_0, \Lambda_0)$:

$$\begin{aligned}
& \mathbb{G}_n\left\{\dot{l}^{ps}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \Lambda_0) - \dot{l}^{ps}_{\Lambda}(\boldsymbol{\theta}_0, \Lambda_0)[h^*]\right\} + o_p(1) \\
= \;& -n^{1/2}P\left\{\ddot{l}^{ps}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \Lambda_0) - \ddot{l}^{ps}_{\Lambda\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \Lambda_0)[h^*]\right\}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\
& -n^{1/2}P\left\{\ddot{l}^{ps}_{\boldsymbol{\theta}\Lambda}(\boldsymbol{\theta}_0, \Lambda_0)[\widehat{\Lambda} - \Lambda_0] - \ddot{l}^{ps}_{\Lambda\Lambda}(\boldsymbol{\theta}_0, \Lambda_0)[h^*, \widehat{\Lambda} - \Lambda_0]\right\} \\
& +n^{1/2}O\left(|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0|^2 + \|\widehat{\Lambda} - \Lambda_0\|^2_{2,P_V}\right). \qquad (6.26)
\end{aligned}$$

Here $\ddot{l}^{ps}_{\boldsymbol{\theta}\Lambda}(\boldsymbol{\theta}_0, \Lambda_0)[\widehat{\Lambda} - \Lambda_0]$ is the derivative of $\dot{l}_{\boldsymbol{\theta}}$ along the path $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, $\Lambda = \Lambda_0 + \epsilon(\widehat{\Lambda} - \Lambda_0)$, and $\ddot{l}^{ps}_{\Lambda\Lambda}(\boldsymbol{\theta}_0, \Lambda_0)[h^*, \widehat{\Lambda} - \Lambda_0]$ is the derivative of $\dot{l}_{\Lambda}[h^*]$ along the path $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, $\Lambda =$

$\Lambda_0 + \epsilon(\widehat{\Lambda} - \Lambda_0)$.

We need to find the $\boldsymbol{h}^*$ to make the second term on the right side of (6.26) be 0. The derivatives of the marginal score functions in (4.9), (4.10), and (4.11) with respect to Λ at direction of $h(v)$ are

$$\begin{aligned}\ddot{l}^{ps}_{\mu\Lambda}[h_\mu(v)] &= \int_0^\infty h_\mu(v)e^{\boldsymbol{\beta}^T\boldsymbol{X}}\Big(2E\{\xi e^{-\mu\xi} \mid \boldsymbol{\aleph}\} - E\{\xi \mid \boldsymbol{\aleph}\}E\{e^{-\mu\xi} \mid \boldsymbol{\aleph}\} + \Lambda(t)e^{\boldsymbol{\beta}^T\boldsymbol{X}} \\ &\quad \times\big[E\{\xi e^{-\mu\xi} \mid \boldsymbol{\aleph}\}E\{e^{-\mu\xi} \mid \boldsymbol{\aleph}\} - E\{\xi e^{-2\mu\xi} \mid \boldsymbol{\aleph}\}\big]\Big)dN(v), \qquad (6.27)\end{aligned}$$

$$\begin{aligned}\ddot{l}^{ps}_{\boldsymbol{\beta}\Lambda}[h_\beta(v)] &= -\boldsymbol{X}\int_0^\infty h_\beta(v)e^{\boldsymbol{\beta}^T\boldsymbol{X}}\Big(\Lambda(v)e^{\boldsymbol{\beta}^T\boldsymbol{X}}\big[E\{e^{-\mu\xi} \mid \boldsymbol{\aleph}\}^2 - E\{e^{-2\mu\xi} \mid \boldsymbol{\aleph}\}\big] \\ &\quad + E\{e^{-\mu\xi} \mid \boldsymbol{\aleph}\}\Big)dN(v), \qquad (6.28)\\ \ddot{l}^{ps}_{\Lambda\Lambda}[h(v), h^*_\theta(v)] &= -\int_0^\infty h(v)h^*_\theta(v)\Big(e^{2\boldsymbol{\beta}^T\boldsymbol{X}}\big[E\{e^{-\mu\xi} \mid \boldsymbol{\aleph}\}^2 - E\{e^{-2\mu\xi} \mid \boldsymbol{\aleph}\}\big] \\ &\quad + \Lambda(v)^{-2}\Big)dN(v). \qquad (6.29)\end{aligned}$$

Based on the derivatives in (6.27), (6.28), and (6.29), we obtain $h^*_\mu(v)$ and $h^*_\beta(v)$ in (4.12) and (4.13) such that

$$\begin{aligned}\sum_{v:dN(v)=1} E\Big(h(v)E\big[g_{1v}\{\boldsymbol{\aleph}\} - h^*_\mu(v)g_{3v}\{\boldsymbol{\aleph}\}\big] \mid V = v\Big) = 0,\\ \sum_{v:dN(v)=1} E\Big(h(v)E\big[g_{2v}\{\boldsymbol{\aleph}\} - h^*_\beta(v)g_{3v}\{\boldsymbol{\aleph}\}\big] \mid V = v\Big) = 0,\end{aligned}$$

respectively, and

$$\begin{aligned}g_{1v}\{\boldsymbol{\aleph}\} &= e^{\boldsymbol{\beta}^T\boldsymbol{X}}\Big(2E\{\xi e^{-\mu\xi} \mid \boldsymbol{\aleph}\} - E\{\xi \mid \boldsymbol{\aleph}\}E\{e^{-\mu\xi} \mid \boldsymbol{\aleph}\} \\ &\quad + \Lambda(v)e^{\boldsymbol{\beta}^T\boldsymbol{X}}\big[E\{\xi e^{-\mu\xi} \mid \boldsymbol{\aleph}\}E\{e^{-\mu\xi} \mid \boldsymbol{\aleph}\} - E\{\xi e^{-2\mu\xi} \mid \boldsymbol{\aleph}\}\big]\Big),\end{aligned}$$

$$
\begin{aligned}
g_{2v}\{\boldsymbol{\aleph}\} &= \boldsymbol{X} e^{\boldsymbol{\beta}^T \boldsymbol{X}} \Big(\Lambda(v) e^{\boldsymbol{\beta}^T \boldsymbol{X}} \left[E\{e^{-\mu\xi} \mid \boldsymbol{\aleph}\}^2 - E\{e^{-2\mu\xi} \mid \boldsymbol{\aleph}\} \right] + E(e^{-\mu\xi} \mid \boldsymbol{\aleph}) \Big), \\
g_{3v}\{\boldsymbol{\aleph}\} &= e^{2\boldsymbol{\beta}^T \boldsymbol{X}} \left[E\{e^{-\mu\xi} \mid \boldsymbol{\aleph}\}^2 - E\{e^{-2\mu\xi} \mid \boldsymbol{\aleph}\} \right] + \Lambda(v)^{-2}.
\end{aligned}
$$

The third term on the right side of (6.26) is $o_p(1)$ because of the convergence rate for $(\widehat{\boldsymbol{\theta}}, \widehat{\Lambda})$. Hence,

$$
\begin{aligned}
& -n^{1/2} P \left\{ \ddot{l}^{ps}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \Lambda_0) - \ddot{l}^{ps}_{\boldsymbol{\theta}\Lambda}(\boldsymbol{\theta}_0, \Lambda_0)[h^*_\theta] \right\} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\
&= \mathbb{G}_n \left\{ \dot{l}^{ps}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \Lambda_0) - \dot{l}^{ps}_{\Lambda}(\boldsymbol{\theta}_0, \Lambda_0)[h^*_\theta] \right\} + o_p(1). \qquad (6.30)
\end{aligned}
$$

It remains to show that $P\left\{ \ddot{l}^{ps}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \Lambda_0) - \ddot{l}^{ps}_{\boldsymbol{\theta}\Lambda}(\boldsymbol{\theta}_0, \Lambda_0)[h^*_{\boldsymbol{\theta}}(v)] \right\}$ is non-singular. Suppose that the matrix is singular, then there exists a non-0 vector $\boldsymbol{b} = (b_1, \ldots, b_d)$ for $d \geq 2$ such that

$$
\begin{aligned}
& \boldsymbol{b}^T P \left\{ \ddot{l}^{ps}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \Lambda_0) - \ddot{l}^{ps}_{\boldsymbol{\theta}\Lambda}(\boldsymbol{\theta}_0, \Lambda_0)[h^*_{\boldsymbol{\theta}}(v)] \right\} \boldsymbol{b} \\
&= P \left[\int \boldsymbol{b}^T \left\{ \ddot{l}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \Lambda_0) - \ddot{l}_{\boldsymbol{\theta}\Lambda}(\boldsymbol{\theta}_0, \Lambda_0)[h^*_{\boldsymbol{\theta}}(v)] \right\} \boldsymbol{b} dN(v) \right] \\
&= 0. \qquad (6.31)
\end{aligned}
$$

The equation in (6.31) is equivalent to $P\left[\int \{\boldsymbol{b}^T (\dot{l}_{\boldsymbol{\theta}} - \dot{l}_\Lambda[h^*])\}^2 dN(v) \right] = 0$. Then

$$
\boldsymbol{b}^T \dot{l}_{\boldsymbol{\theta}} - \boldsymbol{b}^T \dot{l}_\Lambda[h^*(v)] = \boldsymbol{b}^T \begin{pmatrix} \mu^{-1} - E_\xi\{\kappa(v)\xi \mid \boldsymbol{\aleph}\} - h^*_\mu(v)\Lambda(v)^{-1} E\{\kappa(v) \mid \boldsymbol{\aleph}\} \\ X E_\xi\{\kappa(v) \mid \boldsymbol{\aleph}\} - h^*_\beta(v)\Lambda(v)^{-1} E\{\kappa(v) \mid \boldsymbol{\aleph}\} \end{pmatrix} = 0 \quad \text{almost surely,}
$$

where $\kappa(v) = 1 - \Lambda(v) e^{\beta^T X - \mu\xi}$, and $h^*_\mu(v)$ and $h^*_{\boldsymbol{\theta}}(v)$ are in (4.12) and (4.13), respectively. We obtain :

$$
E\left[\boldsymbol{b}^T \left(-\xi - h^*_\mu(v)/\Lambda(v), \{\boldsymbol{X} - h^*_{\boldsymbol{\beta}}(v)/\Lambda(v)\}^T \right) \left\{ 1 - \Lambda(v) \exp(\boldsymbol{\beta}^T \boldsymbol{X} - \mu\xi) \right\} + b_1/\mu \mid \boldsymbol{\aleph} \right] = 0,
$$

and this expectation can be treated as the Laplace transformation of some function of

ξ, so we immediately conclude

$$\boldsymbol{b}^T \left(-\xi - h^*_\mu(v)\Lambda(v)^{-1}, \{\boldsymbol{X} - h^*_{\boldsymbol{\beta}}(v)\Lambda(v)^{-1}\}^T\right)\left\{1 - \Lambda(v)\exp(\boldsymbol{\beta}^T\boldsymbol{X} - \mu\xi)\right\} + b_1/\mu = 0.$$

Since ξ is arbitrary, we can obtain $b_1 = 0, (b_2, \ldots, b_d)^T\boldsymbol{X} = 0$, and $\boldsymbol{b}^T\boldsymbol{h}^*(v)/\Lambda(v) = 0$. Therefore, $\boldsymbol{b} = 0$, and this contradicts the assumption about singularity of the matrix.

Finally, from (6.30), we obtain that

$$\begin{aligned} n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) &= -\left[P\left\{\ddot{l}^{ps}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \Lambda_0) - \ddot{l}^{ps}_{\boldsymbol{\theta}\Lambda}(\boldsymbol{\theta}_0, \Lambda_0)[h^*]\right\}\right]^{-1} \mathbb{G}_n\left\{\dot{l}^{ps}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \Lambda_0) - \dot{l}^{ps}_{\Lambda}(\boldsymbol{\theta}_0, \Lambda_0)[h^*]\right\} \\ &\quad + o_p(1). \end{aligned}$$

Therefore, $n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges to a normal distribution and has marginal influence function given by $\left[P\left\{\dot{l}^{ps}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \Lambda_0) - \dot{l}^{ps}_{\boldsymbol{\theta}\Lambda}(\boldsymbol{\theta}_0, \Lambda_0)[h^*]\right\}\right]^{-1}\left\{\dot{l}^{ps}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \Lambda_0) - \dot{l}^{ps}_{\Lambda}(\boldsymbol{\theta}_0, \Lambda_0)[h^*]\right\}$. □

Proof of Theorem 4.3.4. To prove the consistency of the variance estimator provided in section 4.2.3, it is sufficient to show that $\widehat{\boldsymbol{D}}$ and $\widehat{\mathbf{A}}$ are consistent estimators, where $\boldsymbol{D} = P\left(\ddot{l}^{ps}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \Lambda_0) - \ddot{l}^{ps}_{\boldsymbol{\theta}\Lambda}(\boldsymbol{\theta}_0, \Lambda_0)[h^*_{\boldsymbol{\theta}}(v)]\right)$ and $\mathbf{A} = \dot{l}^{ps}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \Lambda_0) - \dot{l}^{ps}_{\Lambda}(\boldsymbol{\theta}_0, \Lambda_0)[h^*_{\boldsymbol{\theta}}(v)]$.

By the uniform convergence of $(\widehat{\boldsymbol{\Theta}}, \widehat{\Lambda})$ almost surely, we conclude that $\widehat{E}\{g(\xi) \mid \aleph\}$ converges to $E\{g(\xi) \mid \aleph\}$ uniformly in $\aleph$, where $g(\xi)$ can be $1 - \Lambda(v)\exp(\boldsymbol{\beta}^T\boldsymbol{X} - \mu\xi)$, ξ, ξ^2, $\exp(-\mu\xi)$, $\exp(-2\mu\xi)$, $\xi\exp(-\mu\xi)$, $\xi\exp(-2\mu\xi)$, $\xi^2\exp(-\mu\xi)$, and $\xi^2\exp(-2\mu\xi)$. Therefore,

$$\sup_v \left| \frac{\sum_{i=1}^n \sum_{j=1}^{n_i} K_{h_n}(v_{ij} - v)\widehat{E}(g(\xi) \mid \aleph_i)}{\sum_{i=1}^n \sum_{j=1}^{n_i} K_{h_n}(v_{ij} - v)} - \frac{\sum_{i=1}^n \sum_{j=1}^{n_i} K_{h_n}(v_{ij} - v)E(\kappa(v) \mid \aleph_i)}{\sum_{i=1}^n \sum_{j=1}^{n_i} K_{h_n}(v_{ij} - v)} \right| \to 0$$

where $K_{h_n}(x) = h_n^{-1}\exp\{-x^2/h_n\}$. On the other hand, following the general results in Hansen (2008), under the conditions for h_n, we obtain

$$\sup_v \left| \frac{\sum_{i=1}^n \sum_{j=1}^{n_i} K_{h_n}(v_{ij} - v)E\left\{g(\xi) \mid \aleph_i\right\}}{\sum_{i=1}^n \sum_{j=1}^{n_i} K_{h_n}(v_{ij} - v)} - E\left[E\left\{g(\xi)|\aleph\right\}|V = v\right] \right| \to 0$$

with probability one. Similarly, we can show the uniform convergence of the estimators for $E\left[\exp(\boldsymbol{\beta}^T\boldsymbol{X})E\{\xi \mid \aleph\}E\{\exp(-\mu\xi) \mid \aleph\} \mid V = v\right]$,

$E\left[\exp(\boldsymbol{\beta}^T\boldsymbol{X})E\{\xi\exp(-\mu\xi) \mid \aleph\} \mid V = v\right]$,

$E\left[\exp(2\boldsymbol{\beta}^T\boldsymbol{X})\Lambda(v)E\{\xi\exp(-\mu\xi) \mid \aleph\}E\{\exp(-\mu\xi) \mid \aleph\} \mid V = v\right]$,

$E\left[\exp(2\boldsymbol{\beta}^T\boldsymbol{X})\Lambda(v)E\{\xi\exp(-2\mu\xi) \mid \aleph\} \mid V = v\right]$,

$E\left[\boldsymbol{X}\exp(2\boldsymbol{\beta}^T\boldsymbol{X})E\{\exp(-\mu\xi) \mid \aleph\}^2 \mid V = v\right]$,

$E\left[\boldsymbol{X}\exp(C_1\boldsymbol{\beta}^T\boldsymbol{X})\Lambda(v)E\{\xi\exp(-C_2\mu\xi) \mid \aleph\} \mid V = v\right]$, and

$E\left[\boldsymbol{X}\exp(2\boldsymbol{\beta}^T\boldsymbol{X})E\{\exp(-2\mu\xi) \mid \aleph\} \mid V = v\right]$, where C_1 and C_2 are constants. Consequently, $\widehat{h}^*_\theta(v)$ converges to $h^*_\theta(v)$ uniformly in v. This immediately gives that $\widehat{\mathbf{A}}$ converges uniformly to $\mathbf{A}$ given $\aleph$. Thus, $n^{-1}\sum_{i=1}^n(\mathbf{A}_i\mathbf{A}_i^T)$ converges to $P(\mathbf{A}\mathbf{A}^T)$. Using a similar argument, it can be shown that $\widehat{\boldsymbol{D}}$ converges uniformly to $\boldsymbol{D}$ given $\aleph$. Hence, the result of the theorem holds. □

APPENDIX C: Technical Details for Chapter 5

Theorems in Chapter 5 are easily derived from theorems Chapter 4. So, we sketch the different proofs from those for theorems in Chapter 4.

C.1 Proof of Asymptotic Results

Proof of Theorem 5.3.1. The following conditions are needed to examined to apply Theorem 2.12 (Kosorok 2008):

(a) For any sequence $\{(\boldsymbol{\theta}_n, \boldsymbol{\alpha}_n, \Lambda_n)\} \in \boldsymbol{\Theta}_1 \times \boldsymbol{\Theta}_2 \times \boldsymbol{\Phi}^*$,

$\underline{\lim}_{n\to\infty} l^{wps}(\boldsymbol{\theta}_n, \boldsymbol{\alpha}_n, \Lambda_n \mid \aleph) \geq l^{wps}(\boldsymbol{\theta}_0, \boldsymbol{\alpha}_0, \Lambda_0 \mid \aleph)$

implies $d\left((\boldsymbol{\theta}_n, \boldsymbol{\alpha}_n, \Lambda_n), (\boldsymbol{\theta}_0, \boldsymbol{\alpha}_0, \Lambda_0)\right) \to 0$,

(b) $\mathbb{P}_n l^{ps}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\Lambda}) = \sup_{(\boldsymbol{\theta},\boldsymbol{\alpha},\Lambda)\in\boldsymbol{\Theta}_1\times\boldsymbol{\Theta}_2\times\boldsymbol{\Phi}^*} \mathbb{P}_n l^{ps}(\boldsymbol{\theta}, \boldsymbol{\alpha}, \Lambda) - o_p(1)$,

(c) $\sup_{(\boldsymbol{\theta},\boldsymbol{\alpha},\Lambda)\in\boldsymbol{\Theta}_1\times\boldsymbol{\Theta}_2\times\boldsymbol{\Phi}^*} |\mathbb{P}_n l^{wps}(\boldsymbol{\theta}, \boldsymbol{\alpha}\Lambda) - P l^{wps}(\boldsymbol{\theta}, \boldsymbol{\alpha}, \Lambda)| \to 0$ in probability, as $n \to \infty$.

Condition (a) is satisfied by identifiability of the marginal likelihood proved in Lemma B.1.1 and a generalized linear model with Condition (A2) and (A6). Condition (b) is satisfied by the continuous mapping theorem and MLE, $\widehat{\boldsymbol{\alpha}}$ of the partial likelihood in (5.2):

$$\mathbb{P}_n l^{wps}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\Lambda}) = \sup_{(\boldsymbol{\theta},\Lambda)\in\boldsymbol{\Theta}\times\boldsymbol{\Phi}^*} \mathbb{P}_n l^{wps}(\boldsymbol{\theta}, \widehat{\boldsymbol{\alpha}}, \Lambda) - o_p(1) = \sup_{(\boldsymbol{\theta},\boldsymbol{\alpha},\Lambda)\in\boldsymbol{\Theta}\times\boldsymbol{\Phi}^*} \mathbb{P}_n l^{wps}(\boldsymbol{\theta}, \boldsymbol{\alpha}, \Lambda) - o_p(1).$$

Define

$$\mathscr{H}_w = \left\{ l^{wps}(\boldsymbol{\theta}, \boldsymbol{\alpha}, \Lambda, \aleph) \mid (\boldsymbol{\theta}, \boldsymbol{\alpha}) \in \boldsymbol{\Theta}_1 \times \boldsymbol{\Theta}_2, \Lambda(t) \in \boldsymbol{\Phi}^* \right\}, \tag{6.32}$$

where the parameter space, $\boldsymbol{\Phi}^* = \{\Lambda(t) \mid \Lambda(t) = -\log S(t),\ S(t)$ is a non-increasing function with $S(0) = 1,\ S(t) \geq 0\}$.

We calculate the bracket covering number for the class in (6.32). For any $(\mu_1, \boldsymbol{\beta}_1, \boldsymbol{\alpha}_1, \Lambda_1)$ and $(\mu_2, \boldsymbol{\beta}_2, \boldsymbol{\alpha}_2, \Lambda_2) \in \boldsymbol{\Theta}_1 \times \boldsymbol{\Theta}_2 \times \boldsymbol{\Phi}^*$ such that $\sup_{v \in \mathscr{S}[V]} |\Lambda_1(v) - \Lambda_2(v)| < \epsilon$, $|\mu_1 - \mu_2| < \epsilon$, $|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2| < \epsilon$, and $|\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2| < \epsilon$ for $\epsilon > 0$, we wish to set boundaries for the bracket covering number for the class $\mathscr{H}_w$. There exist positive constants C_1 such that $|(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)^T \boldsymbol{X} - (\mu_1 - \mu_2)\xi| \leq C_1\epsilon + \epsilon|\xi|$ and C_2 such that $|(\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2)^T g\{\overline{\boldsymbol{A}}(t)\}| \leq C_2\epsilon$. Then for some positive constants, C_2, C_3,C_4, and C_5 and by the properties of the link function for the generalized linear model in Condition (A8),

$$
\begin{aligned}
&|l^{wps}(\mu_1, \boldsymbol{\beta}_1, \boldsymbol{\alpha}_1, \Lambda_1, \boldsymbol{\aleph}) - l^{wps}(\mu_2, \boldsymbol{\beta}_2, \boldsymbol{\alpha}_2, \Lambda_2, \boldsymbol{\aleph})| \\
&\leq \sum_{v:dN(v)=1} |\overline{\pi}(v)(\alpha_1)^{-1}[l\{\mu_1, \boldsymbol{\beta}_1, \Lambda_1(v), \boldsymbol{\aleph}\} - l\{\mu_2, \boldsymbol{\beta}_2, \Lambda_2(v), \boldsymbol{\aleph}\}]| \\
&+ \sum_{v:dN(v)=1} |\overline{\pi}(v)(\alpha_1)^{-1} l\{\mu_2, \boldsymbol{\beta}_2, \Lambda_2(v), \boldsymbol{\aleph}\} - \overline{\pi}(v)(\alpha_2)^{-1} l\{\mu_2, \boldsymbol{\beta}_2, \Lambda_2(v), \boldsymbol{\aleph}\}| \\
&\leq \sum_{v:dN(v)=1} \left[\int \{C_3|\Lambda_1(v) - \Lambda_2(v)| + C_4\epsilon + C_5\epsilon|\xi|\} \, \phi\left\{\frac{Y(v) - \xi}{\sigma}\right\} d\xi + m(\overline{\boldsymbol{A}}(t)) C_2\epsilon \right].
\end{aligned}
$$

Thus, $\|l^{wps}(\mu_1, \boldsymbol{\beta}_1, \boldsymbol{\alpha}_1, \Lambda_1 \mid \boldsymbol{\aleph}) - l^{wps}(\mu_2, \boldsymbol{\beta}_2, \boldsymbol{\alpha}_2, \Lambda_2 \mid \boldsymbol{\aleph})\|_{1,P} \leq (C_2' + C_3' + C_4' + C_5')\epsilon$ for some positive constants, C_2', C_3', C_4', and C_5'. We obtain that $\log N_{[\,]}(O(1)\epsilon, \mathscr{H}_w, L_1(P))$ $\leq \log N_{[\,]}(\epsilon, \boldsymbol{\Theta}_1 \times \boldsymbol{\Theta}_2 \times \boldsymbol{\Phi}^*, \|\cdot\|_{l^\infty}) \leq O(1/\epsilon)$. Hence, $\mathscr{H}_w$ is P-Glivenko-Cantelli class, and theorem 5.3.1 is proved. □

Once consistency of $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\alpha}}$ is established, we can concentrate on a neighborhoods of $\boldsymbol{\theta}_0$ and $\boldsymbol{\alpha}_0$. For any $\eta_1 > 0$ and $\eta_2 > 0$ let $B(\boldsymbol{\theta}_0, \eta_1)$ and $B(\boldsymbol{\alpha}_0, \eta_2)$ be the balls with radius η_1 and η_2 centered at $\boldsymbol{\theta}_0$ and $\boldsymbol{\alpha}_0$, respectively. If $\boldsymbol{\theta}_0$ or $\boldsymbol{\alpha}_0$ is on the boundary of $\boldsymbol{\Theta}_1$ or $\boldsymbol{\Theta}_2$, respectively, then take $B(\boldsymbol{\theta}_0, \eta_1) \cap \boldsymbol{\Theta}_1$ or $B(\boldsymbol{\alpha}_0, \eta_2) \cap \boldsymbol{\Theta}_2$ instead of $B(\boldsymbol{\theta}_0, \eta_1)$ or $B(\boldsymbol{\alpha}_0, \eta_2)$. Then $B(\boldsymbol{\theta}_0, \eta_1)$ and $B(\boldsymbol{\alpha}_0, \eta_2)$ are included in $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$, respectively. We suppose that condition (A3) is satisfied so that Λ_0 is bounded and away from 0 on $\mathscr{S}[V]$. Since we have proved that $\widehat{\Lambda}$ converges on $\mathscr{S}[V]$, we may restrict $\widehat{\Lambda}$ to the

following class of functions:

$$\Phi = \{\Lambda \mid \Lambda \text{ is non-decreasing and } 0 < 1/M \leq \Lambda(t) \leq M < \infty \text{ for all } t \in \mathscr{S}[V]\}, \quad (6.33)$$

where M is a large positive constant.

Using Lemma $B.2.1$, we determine the size of the class for weighted pseudo-likelihood functions of interest.

Lemma C.1.1. *Let*

$$\mathscr{H}_w = \{l^{wps}(\boldsymbol{\theta}, \boldsymbol{\alpha}, \Lambda, \aleph(t)) \mid \boldsymbol{\theta} \in B(\boldsymbol{\theta}_0, \eta_1), \boldsymbol{\alpha} \in B(\boldsymbol{\alpha}_0, \eta_2), \text{ and } \Lambda \in \Phi\}. \quad (6.34)$$

Suppose that (A2), (A6), (A7), and (A8) are satisfied. Then there exists a constant $C > 0$ *such that*

$$\sup_Q N_{[\,]}(\epsilon, \mathscr{H}_w, L_2(Q)) \leq C(1/\epsilon^{(d_1+d_2)})e^{1/\epsilon} \quad \text{for all} \quad \epsilon > 0,$$

where d_1 *and* d_2 *are the dimension of* $\boldsymbol{\Theta}_1$ *and* $\boldsymbol{\Theta}_2$*, respectively. Hence, for* ϵ *small enough, we have*

$$\sup_Q \log N_{[\,]}(\epsilon, \mathscr{H}_w, L_2(Q)) \leq C(1/\epsilon).$$

Here Q *runs through the class of all probability measures.*

Proof. For $\mathscr{H} = \{l^{ps}(\boldsymbol{\theta}, \Lambda, \aleph(t)) \mid \boldsymbol{\theta} \in B(\boldsymbol{\theta}_0, \eta_1) \text{and } \Lambda \in \Phi\}$ we obtained the bracketing functions in Lemma $B.2.1$:

$$\int_0^{u_V} \log f_{ij}^{*L} dN(v) \leq \sum_{v:dN(v)=1} \log f(v) \leq \int_0^{u_V} \log f_{ij}^{*U} dN(v),$$

where f_{ij}^{*L} and f_{ij}^{*U} are the marginal envelope functions in (6.23) and (6.24), respectively.

By Condition (A6), (A7), and (A8) we can also construct the bracketing functions

for $\mathscr{H}_\alpha = \{H_\alpha^{-1} \mid \boldsymbol{\alpha} \in B(\boldsymbol{\Theta}_2, \eta_2)\}$ where $H_\alpha^{-1} = link^{-1}[\boldsymbol{\alpha}^T g\{\overline{A}(v)\}]$. Then Condition (A8) yields in $H_{\alpha_1}^{-1} - \delta m \le H_{\alpha_2}^{-1} \le g_{\alpha_1} - \delta m$ for any $|\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2| \le \epsilon$. Hence, the bracketing number $N_{[\,]}(\epsilon, \mathscr{H}_\alpha, L_2(P))$ is of order $O(\epsilon^{-d_2})$. Thus the entroy is of smaller order than $\log(1/\epsilon)$. Hence the bracketing entropy integral certainly converges, and the class of functions $\mathscr{H}_\alpha$ is Donsker. The envelopes of $\mathscr{H}$ and $\mathscr{H}_alpha$ are integrable, and $\mathscr{H}_w$ is $\mathscr{H} \cdot \mathscr{H}_\alpha$, so $N_{[\,]}(\epsilon, \mathscr{H}_w, L_2(Q))$ is of order $O(\epsilon^{-(d_1+d_2)} e^{1/\epsilon})$. □

Theorem 5.3.2 is trivially justified by the same argument for Theorem 4.3.2.

Proof of Theorem 5.3.3. Since $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\Lambda})$ are maximum likelihood estimators for $(\boldsymbol{\theta}, \boldsymbol{\alpha}, \Lambda)$, we immediately obtain that $\mathbb{P}_n \left\{ \dot{l}_{\boldsymbol{\theta}}^{wps}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\Lambda}) - \dot{l}_{\Lambda}^{wps}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\Lambda})[h^*] \right\} = 0$. Thus

$$\begin{aligned} &\mathbb{G}_n\{\dot{l}_{\boldsymbol{\theta}}^{wps}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\Lambda}) - \dot{l}_{\Lambda}^{wps}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\Lambda})[h^*]\} + o_p(1) \\ &= -n^{1/2} P\{\dot{l}_{\boldsymbol{\theta}}^{wps}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\Lambda}) - \dot{l}_{\Lambda}^{wps}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\Lambda})[h^*]\}, \end{aligned}$$

where $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$.

Let us consider the following classes of functions when $|\boldsymbol{\theta} - \boldsymbol{\theta}_0| \le \eta$, $|\boldsymbol{\alpha} - \boldsymbol{\alpha}_0| \le \eta$, and $\|\Lambda - \Lambda_0\|_{2,P_V} \le \eta$ where η is near 0:

$$\left\{\dot{l}_{\boldsymbol{\theta}}^{wps}(\boldsymbol{\theta}, \boldsymbol{\alpha}, \Lambda) - \dot{l}_{\boldsymbol{\theta}}^{wps}(\boldsymbol{\theta}_0, \boldsymbol{\alpha}_0, \Lambda_0)\right\} \text{ and } \left\{\dot{l}_{\Lambda}^{wps}(\boldsymbol{\theta}, \boldsymbol{\alpha}, \Lambda)[h^*] - \dot{l}_{\Lambda}^{wps}(\boldsymbol{\theta}_0, \boldsymbol{\alpha}_0, \Lambda_0)[h^*]\right\}.$$

The entropy numbers for the classes are of order $1/\eta$ and this implies that these classes are P-Donsker. Hence, $\dot{l}_{\boldsymbol{\theta}}^{wps}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\Lambda}) - \dot{l}_{\Lambda}^{wps}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\Lambda})[h^*]$ belongs to a P-Donsker class. This leads to

$$\mathbb{G}_n \left\{\dot{l}_{\boldsymbol{\theta}}^{wps}(\boldsymbol{\theta}_0, \boldsymbol{\alpha}_0, \Lambda_0) - \dot{l}_{\Lambda}(\boldsymbol{\theta}_0, \boldsymbol{\alpha}_0, \Lambda_0)[h^*]\right\} + o_p(1) \tag{6.35}$$

$$= -n^{1/2} P \left\{\dot{l}_{\boldsymbol{\theta}}^{wps}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\Lambda}) - \dot{l}_{\Lambda}^{wps}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\Lambda})[h^*]\right\}. \tag{6.36}$$

We perform a Taylor's series expansion of the right side in (6.36) at $(\boldsymbol{\theta}_0, \boldsymbol{\alpha}_0, \Lambda_0)$:

$$
\begin{aligned}
& \mathbb{G}_n\{\dot{l}^{wps}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \boldsymbol{\alpha}_0, \Lambda_0) - \dot{l}^{wps}_{\Lambda}(\boldsymbol{\theta}_0, \boldsymbol{\alpha}_0, \Lambda_0)[h^*]\} + o_p(1) \\
= \; & -n^{1/2}P\{\ddot{l}^{wps}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \boldsymbol{\alpha}_0, \Lambda_0) - \ddot{l}^{wps}_{\Lambda\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \boldsymbol{\alpha}_0, \Lambda_0)[h^*]\}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\
& -n^{1/2}P\{\ddot{l}^{wps}_{\boldsymbol{\theta}\Lambda}(\boldsymbol{\theta}_0, \boldsymbol{\alpha}_0, \Lambda_0)[\widehat{\Lambda} - \Lambda_0] - \ddot{l}^{wps}_{\Lambda\Lambda}(\boldsymbol{\theta}_0, \boldsymbol{\alpha}_0, \Lambda_0)[h^*, \widehat{\Lambda} - \Lambda_0]\} \\
& -n^{1/2}P\{\ddot{l}^{wps}_{\boldsymbol{\theta}\boldsymbol{\alpha}}(\boldsymbol{\theta}_0, \boldsymbol{\alpha}_0, \Lambda_0) - \ddot{l}^{wps}_{\Lambda\boldsymbol{\alpha}}(\boldsymbol{\theta}_0, \boldsymbol{\alpha}_0, \Lambda_0)[h^*]\}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) \\
& +n^{1/2}O(|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0|^2 + |\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0|^2 + \|\widehat{\Lambda} - \Lambda_0\|^2_{2,P_V}). \qquad (6.37)
\end{aligned}
$$

Here $\ddot{l}^{wps}_{\boldsymbol{\theta}\Lambda}(\boldsymbol{\theta}_0, \boldsymbol{\alpha}_0, \Lambda_0)[\widehat{\Lambda} - \Lambda_0]$ is the derivative of $\dot{l}^{wps}_{\boldsymbol{\theta}}$ along the path $\boldsymbol{\theta} = \boldsymbol{\theta}_0$,$\boldsymbol{\alpha} = \boldsymbol{\alpha}_0$, $\Lambda = \Lambda_0 + \epsilon(\widehat{\Lambda} - \Lambda_0)$, and $\ddot{l}^{wps}_{\Lambda\Lambda}(\boldsymbol{\theta}_0, \boldsymbol{\alpha}_0, \Lambda_0)[h^*, \widehat{\Lambda} - \Lambda_0]$ is the derivative of $\dot{l}_{\Lambda}[h^*]$ along the path $\boldsymbol{\theta} = \boldsymbol{\theta}_0$,$\boldsymbol{\alpha} = \boldsymbol{\alpha}_0$, $\Lambda = \Lambda_0 + \epsilon(\widehat{\Lambda} - \Lambda_0)$.

Let $\boldsymbol{D}_w$ be $P\{\ddot{l}^{wps}_{\boldsymbol{\theta}\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \boldsymbol{\alpha}_0, \Lambda_0) - \ddot{l}^{wps}_{\Lambda\boldsymbol{\theta}}(\boldsymbol{\theta}_0, \boldsymbol{\alpha}_0, \Lambda_0)[h^*]\}$ in the first term in right side of (6.37). Then we need to calculate the following second derivatives for $\boldsymbol{D}_w$:

$$
\begin{aligned}
\ddot{l}^{wps}_{\mu\mu} &= \int_0^{\infty} W_1(v)\Big(-\mu^{-2} - \Lambda(v)e^{\boldsymbol{\beta}^T\boldsymbol{X}}\left[3E\{\xi^2 e^{-\mu\xi} \mid \boldsymbol{\aleph}\} - 2E\{\xi \mid \boldsymbol{\aleph}\}E\{\xi e^{-\mu\xi} \mid \boldsymbol{\aleph}\}\right] \\
&\quad +\Lambda(v)^2 e^{2\boldsymbol{\beta}^T\boldsymbol{X}}\left[E\{\xi^2 e^{-2\mu\xi} \mid \boldsymbol{\aleph}\} - E\{\xi e^{-\mu\xi} \mid \boldsymbol{\aleph}\}^2\right] - E\{\xi \mid \boldsymbol{\aleph}\}^2 \\
&\quad +E\{\xi^2 | \boldsymbol{\aleph}\}\Big) dN(v), \\
\ddot{l}^{wps}_{\mu\boldsymbol{\beta}} &= \int_0^{\infty} \boldsymbol{X} W_1(v)\Lambda(v)e^{\boldsymbol{\beta}^T\boldsymbol{X}}\Big(2E(\xi e^{-\mu\xi} \mid \boldsymbol{\aleph}) - E\{e^{-\mu\xi} \mid \boldsymbol{\aleph}\}E\{\xi \mid \boldsymbol{\aleph}\} \\
&\quad +\Lambda(v)e^{\boldsymbol{\beta}^T\boldsymbol{X}}\left[E\{e^{-\mu\xi} \mid \boldsymbol{\aleph}\}E\{\xi e^{-\mu\xi} \mid \boldsymbol{\aleph}\} - E\{\xi e^{-2\mu\xi} \mid \boldsymbol{\aleph}\}\right]\Big) dN(v), \\
\ddot{l}^{wps}_{\boldsymbol{\beta}\boldsymbol{\beta}} &= -\boldsymbol{X}\boldsymbol{X}^T \int_0^{\infty} W_1(v)\Lambda(v)e^{\boldsymbol{\beta}^T\boldsymbol{X}}\Big(\Lambda(v)e^{\boldsymbol{\beta}^T\boldsymbol{X}}\left[E\{e^{-\mu\xi} \mid \boldsymbol{\aleph}\}^2 - E\{e^{-2\mu\xi} \mid \boldsymbol{\aleph}\}\right] \\
&\quad +E\{e^{-\mu\xi} \mid \boldsymbol{\aleph}\}\Big) dN(v),
\end{aligned}
$$

where $W_1(v) = R_j\overline{\pi}_j(\boldsymbol{\alpha})^{-1}$ and $\int_{0\le s\le v} dN(s) = j$.

Focusing on the second term in the right side (6.37), we need to find the $\boldsymbol{h}^*$ to make the second term on the right side of (6.37) be 0. The derivatives of the weighted

pseudo-score functions in (5.9),(5.10), and (5.11) with respect to Λ at direction of $h(v)$ are

$$\begin{aligned}
\ddot{l}^{wps}_{\mu\Lambda}[h_\mu(v)] &= \int_0^\infty W_1(v)h_\mu(v)e^{\boldsymbol{\beta}^T\boldsymbol{X}}\Big(2E\{\xi e^{-\mu\xi} \mid \boldsymbol{\aleph}\} - E\{\xi \mid \boldsymbol{\aleph}\}E\{e^{-\mu\xi}\} + \\
&\quad \Lambda(v)e^{\boldsymbol{\beta}^T\boldsymbol{X}}\left[E\{\xi e^{-\mu\xi} \mid \boldsymbol{\aleph}\}E\{e^{-\mu\xi} \mid \boldsymbol{\aleph}\} - E\{\xi e^{-2\mu\xi}\}\right]\Big)dN(v), \quad (6.38)\\
\ddot{l}^{wps}_{\boldsymbol{\beta}\Lambda}[h_\beta(v)] &= -\boldsymbol{X}\int_0^\infty W_1(v)h_\theta(v)e^{\boldsymbol{\beta}^T\boldsymbol{X}}\Big(\Lambda(v)e^{\boldsymbol{\beta}^T\boldsymbol{X}}\left[E\{e^{-\mu\xi}\}^2 - E\{e^{-2\mu\xi}\}\right] \\
&\quad +E\{e^{-\mu\xi}\}\Big)dN(v), \quad (6.39)\\
\ddot{l}^{wps}_{\Lambda\Lambda}[h(v), h^*(v)] &= -\int_0^\infty W_1(v)h(v)h^*(v)\Big(e^{2\boldsymbol{\beta}^T\boldsymbol{X}}\left[E\{e^{-\mu\xi}\}^2 - E\{e^{-2\mu\xi} \mid \boldsymbol{\aleph}\}\right] \\
&\quad +\Lambda(v)^{-2}\Big)dN(v). \quad (6.40)
\end{aligned}$$

Based on the derivatives in (6.38), (6.39), and (6.40), we obtain $h^*_\mu(v)$ and $h^*_\beta(v)$ such that respectively

$$\begin{aligned}
\sum_{v:dN(v)=1} E\Big(h(v)E\left[g_1\{\boldsymbol{\aleph}\} - h^*_\mu(v)g_3\{\boldsymbol{\aleph}\}\right] \mid V = v\Big) = 0,\\
\sum_{v:dN(v)=1} E\Big(h(v)E\left[g_2\{\boldsymbol{\aleph}\} - h^*_\beta(v)g_3\{\boldsymbol{\aleph}\}\right] \mid V = v\Big) = 0,
\end{aligned}$$

respectively, and

$$\begin{aligned}
g_1\{\boldsymbol{\aleph}\} &= W_1(v)e^{\boldsymbol{\beta}^T\boldsymbol{X}}\Big(2E\{\xi e^{-\mu\xi} \mid \boldsymbol{\aleph}\} - E\{\xi \mid \boldsymbol{\aleph}\}E\{e^{-\mu\xi} \mid \boldsymbol{\aleph}\} \\
&\quad +\Lambda(v)e^{\boldsymbol{\beta}^T\boldsymbol{X}}\left[E\{\xi e^{-\mu\xi} \mid \boldsymbol{\aleph}\}E\{e^{-\mu\xi} \mid \boldsymbol{\aleph}\} - E\{\xi e^{-2\mu\xi} \mid \boldsymbol{\aleph}\}\right]\Big),\\
g_2\{\boldsymbol{\aleph}\} &= W_1(v)\boldsymbol{X}e^{\boldsymbol{\beta}^T\boldsymbol{X}}\Big(\Lambda(v)e^{\boldsymbol{\beta}^T\boldsymbol{X}}\left[E\{e^{-\mu\xi} \mid \boldsymbol{\aleph}\}^2 - E\{e^{-2\mu\xi} \mid \boldsymbol{\aleph}\}\right] + E(e^{-\mu\xi} \mid \boldsymbol{\aleph})\Big),\\
g_3\{\boldsymbol{\aleph}\} &= W_1(v)e^{2\boldsymbol{\beta}^T\boldsymbol{X}}\left[E\{e^{-\mu\xi} \mid \boldsymbol{\aleph}\}^2 - E\{e^{-2\mu\xi} \mid \boldsymbol{\aleph}\}\right] + W_1(v)\Lambda(v)^{-2}.
\end{aligned}$$

Then

$$
\begin{aligned}
h_1(v) &= E\big\{W_1(v)e^{\boldsymbol{\beta}^T\boldsymbol{X}}\big(E\{\xi \mid \aleph\}E\{e^{-\mu\xi} \mid \aleph\} - 2E\{\xi e^{-\mu\xi} \mid \aleph\} \\
&\quad -e^{\boldsymbol{\beta}^T\boldsymbol{X}}\Lambda(v)\big[E\{\xi e^{-\mu\xi} \mid \aleph\}E\{e^{-\mu\xi} \mid \aleph\} - E\{\xi e^{-2\mu\xi} \mid \aleph\}\big]\big) \mid V = v\big\}, \\
h_2(v) &= E\big\{W_1(v)\boldsymbol{X}e^{\boldsymbol{\beta}^T\boldsymbol{X}}\big(\Lambda(v)e^{\boldsymbol{\beta}^T\boldsymbol{X}}\big[E\{e^{-\mu\xi} \mid \aleph\}^2 - E\{e^{-2\mu\xi} \mid \aleph\}\big] + E\{e^{-\mu\xi} \mid \aleph\}\big) \mid V = v\big\}, \\
h_3(v) &= E\big(W_1(v)\Lambda(v)^{-2} + W_1(v)e^{2\boldsymbol{\beta}^T\boldsymbol{X}}\big[E\{e^{-\mu\xi} \mid \aleph\}^2 - E\{e^{-2\mu\xi} \mid \aleph\}\big] \mid V = v\big), \\
h^*_\mu(v) &= h_1(v)/h_3(v), \\
h^*_\beta(v) &= h_2(v)/h_3(v).
\end{aligned}
$$

Regarding the third term in (6.37),

$$
\begin{aligned}
\dot{l}^{wps}_{\mu\alpha}\{\boldsymbol{\theta}, \boldsymbol{\alpha}, \Lambda\} &= \int_{v=0}^{\infty} W_2(v)\big[\mu^{-1} - E\{\kappa(v)\xi \mid \aleph\}\big]dN(v), \\
\dot{l}^{wps}_{\beta\alpha}\{\boldsymbol{\theta}, \boldsymbol{\alpha}, \Lambda\} &= \int_{v=0}^{\infty} \boldsymbol{X}W_2(v)E\{\kappa(v) \mid \aleph\}\,dN(v), \\
\dot{l}^{wps}_{\Lambda\alpha}\{\boldsymbol{\theta}, \boldsymbol{\alpha}, \Lambda\}[h(v)] &= \int_{v=0}^{\infty} W_2(v)h(v)E\{\kappa(v) \mid \aleph\}/\Lambda(v)dN(v),
\end{aligned}
$$

where $W_2(v) = -R_j\overline{\pi}_j\{\boldsymbol{\alpha}\}^{-2}\partial\overline{\pi}_j(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}$ and $\int_{0\le s\le v} dN(s) = j$.

Then we have

$$
\sqrt{n}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) = n^{1/2}P(-\partial^2 \log L/(\partial\boldsymbol{\alpha})^2)^{-1}\mathbb{P}_n S(\boldsymbol{\alpha}_0) + o_p(1), \tag{6.41}
$$

where $S(\boldsymbol{\alpha}) = \partial \log L(\alpha)/\partial\alpha = \int_0^\infty \{R_j - R_{(j-1)}\pi_j(\boldsymbol{\alpha})\}\{\partial \text{logit}\pi_j(\boldsymbol{\alpha})/\partial\alpha\}dN(t)$, and $\mathbb{P}_n = n^{-1}\sum_{i=1}^n$. The asymptotic expression in the right side of (6.41) replaces $\sqrt{n}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)$ in (6.37). Finally, the fourth term on the right side of (6.37) is $o_p(1)$ because of the convergence rate for $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\Lambda})$.

We can asymptotically reexpress $\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)$ from the equation in (6.37) as following and $\boldsymbol{D}_w$ is invertible (this can be proved by the same argument used for the

invertible property of $\boldsymbol{D}$ in Theorem 4.3.3):

$$\begin{aligned}\sqrt{n}\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right) &= -\boldsymbol{D}_w^{-1}\mathbb{G}_n\Big[\dot{l}^{wps}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0,\boldsymbol{\alpha}_0,\Lambda_0)-\dot{l}^{wps}_{\Lambda}(\boldsymbol{\theta}_0,\boldsymbol{\alpha}_0,\Lambda_0)[h^*] \\ &\quad -P\left\{\ddot{l}^{wps}_{\boldsymbol{\theta}\boldsymbol{\alpha}}(\boldsymbol{\theta}_0,\boldsymbol{\alpha}_0,\Lambda_0)-\ddot{l}^{wps}_{\Lambda\boldsymbol{\alpha}}(\boldsymbol{\theta}_0,\boldsymbol{\alpha}_0,\Lambda_0)[h^*]\right\}P\left\{\partial^2\log L/(\partial\boldsymbol{\alpha})^2\right\}^{-1}S(\boldsymbol{\alpha}_0)\Big] \\ &= \mathbb{G}_n\tilde{\psi}_w. \end{aligned} \tag{6.42}$$

Hence we obtain the pseudo-influence function for $\boldsymbol{\theta}$, $\tilde{\psi}_w$ from the equation in (6.42), so we can calculate the information matrix for $\widehat{\boldsymbol{\theta}}$:

$$\begin{aligned} I_w &= P(\tilde{\psi}_w\tilde{\psi}_w^T) = \boldsymbol{D}_w^{-1}P\left\{(\boldsymbol{M}_\theta+\boldsymbol{M}_\alpha)(\boldsymbol{M}_\theta+\boldsymbol{M}_\alpha)^T\right\}\left\{\boldsymbol{D}_w^{-1}\right\}^T \\ &= \boldsymbol{D}_w^{-1}P\left\{(\boldsymbol{M}_\theta\boldsymbol{M}_\theta^T+2\boldsymbol{M}_\theta\boldsymbol{M}_\alpha^T+\boldsymbol{M}_\alpha\boldsymbol{M}_\alpha^T\right\}\left\{\boldsymbol{D}_w^{-1}\right\}^T \end{aligned} \tag{6.43}$$

where $\boldsymbol{M}_\theta = \dot{l}^{wps}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0,\boldsymbol{\alpha}_0,\Lambda_0)-\dot{l}^{wps}_{\Lambda}(\boldsymbol{\theta}_0,\boldsymbol{\alpha}_0,\Lambda_0)[h^*]$, and
$\boldsymbol{M}_\alpha = P\left\{\ddot{l}^{wps}_{\boldsymbol{\theta}\boldsymbol{\alpha}}(\boldsymbol{\theta}_0,\boldsymbol{\alpha}_0,\Lambda_0)-\ddot{l}^{wps}_{\Lambda\boldsymbol{\alpha}}(\boldsymbol{\theta}_0,\boldsymbol{\alpha}_0,\Lambda_0)[h^*]\right\}P\left\{-\partial^2\log L/(\partial\boldsymbol{\alpha})^2\right\}^{-1}S(\boldsymbol{\alpha}_0)$. □

REFERENCE

Abrevaya, J. and Hausman, J. A. (2004), "Response Error in a Transformation Model with an Application to Earnings-Equation Estimation," *Econometrics Journal*, 7, 366–388.

Ayer, M., Brunk, H., Ewing, G., Reid, W., and Silverman, E. (1955), "An Empirical Distribution Function for Sampling with Incomplete Information," *The Annals of Mathematical Statistics*, 26, 641–647.

Bennett, S. (1983a), "Analysis of Survival Data by The Proportional Odds Model," *Statistics in Medicine*, 2, 273–277.

— (1983b), "Log-logistic Regression Models for Survival Data," *Applied Statistics*, 32, 165–171.

Breslow, N. E. and Wellner, J. A. (2006), "Weighted Likelihood for Semiparametric Models and Two-phase Stratified Samples, with Application to Cox Regressioin," *Scandinavian Journal of Statistics*, 34, 86–102.

Buonaccorsi, J. P. (1996), "Measurement Error in the Response in the General Linear Model," *Journal of the American Statistical Association*, 91, 633–642.

Buonaccorsi, J. P. and Tosteson, T. D. (1993), "Correcting for Nonlinear Measurement Errors in the Dependent Variable in the General Linear Model," *Communications in Statistics - Theory and Methods*, 22, 2687–2702.

Carroll, R. J. (2006), *Measurement Error in Nonlinear Models: A Modern Perspective*, Chapman and Hall/CRC.

Carroll, R. J., Spiegelman, C., Lan, K., Bailey, K., and Abbott, R. (1984), "On Errors-in-varialbes for Binary Regression Models," *Biometrika*, 71, 19–26.

Carroll, R. J. and Wand, M. P. (1991), "Semiparametric Estimation in Logistic Measurement Error Models," *Journal of Royal Statistical Society*, 53, 573–585.

Cox, D. R. (1972), "Regressioin Models and Life-Tables," *Journal of Royal Statistical Society*, 34, 187–220.

Dabrowska, D. M. and Doksum, K. A. (1988), "Partial Likelihood in Transformation Models with Censored Data," *Scandinavian Journal of Statistics*, 15, 1–23.

Davis, P. J. (1984), *Methods of Numerical Integration*, Orlando: Academic Press.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, 39, 1–38.

Duncan, B. B., Schmidt, M. I., Pankow, J. S., Ballantyne, C. M., Couper, D., Vigo, A., Hoogeveen, R., Folsom, A. R., and Heiss, G. (2003), "Low-Grade Systemic Inflammation and the Development of Type 2 Diabetes: The Atherosclerosis Risk in Communities Study," *Diabetes*, 52, 1799–1805.

Efron, B. (1967), "The Two Sample Problem with Censored Data," in *Proceedings of the 5th Berkeley Symposium (Vol. 4)*, vol. 4.

Finkelstein, D. M. (1986), "A Proportional Hazards Model for Interval-Censored Failure Time Data," *Biometrics*, 42, 845–854.

Fuller, W. A. (1987), *Measurement Error Models*, John Wiley & Sons, Inc.

Geksus, R. and Groeneboom, P. (1996a), "Asymptotic Optimal Estimation of Smooth Functionals for Interval Censoring, Part 1," Tech. rep., Delft University of Technology.

— (1996b), "Asymptotic Optimal Estimation of Smooth Functionals for Interval Censoring, Part 2," Tech. rep., Delft University of Technology.

— (1999), "Asymptotically Optimal Estimation of Smooth Functionals for Interval Censoring, Case 2," *Annals of Statistics*, 27, 627–674.

Gentleman, R. and Geyer, C. J. (1994), "Maximum Likelihood for Interval Censored Data: Consistency and Computation," *Biometrika*, 81, 618–623.

Groeneboom, P. (1991), "Nonparametric Maximum Likelihood Estimators for Interval Censoring and Deconvolutions," Tech. Rep. 378, Department of Statisticis, Stanford University.

Groeneboom, P. and Wellner, J. A. (1992), *Information Bounds and Nonparametric Maximum Likelihood Estimation*, Basel: Birhauser Verlag.

Guolo, A. (2008), "Robust Techniques for Measurement Error Correction: a Review," *Statistical Methods in Medical Research*, 17, 555–580.

Hansen, B. E. (2008), "Uniform Convergence Rates for Kernel Estimation with Dependent Data," *Econometric Theory*, 24, 726–748.

Hausmana, J., Abrevayab, J., and Scott-Mortonb, F. (1998), "Misclassification of the dependent variable in a discrete-response setting," *Journal of Econometrics*, 87, 239–269.

Hellman, R. (2012), "Glucose meter inaccuracy and the impact on the care of patients," *Diabetes/Metabolism Research and Reviews*, 28, 207–209.

Henschel, V., Heiss, C., and Mansmann, U. (2007), *Iterated Convex Minorant Algorithm for interval censored event data*, http://cran.r-project.org/web/packages/intcox/intcox.pdf.

Hoel, D. G. and Walburg, H. E. (1991), "Statistical Analysis of Survival Experiments," *Journal of the Cancer Institute*, 49, 361–372.

Horvitz, D. G. and Thompson, D. J. (1952), "A Generalization of Sampling without Replacement from a Finite Universe," *Journal of the American Statistical Association*, 47, 663Ű685.

Huang, J. (1996), "Efficient Estimation for the Proportional Hazards Model with Interval Censoring," *Annals of Statistics*, 24, 540–568.

Huang, J. and Rossini, A. J. (1997), "Sieve Estimation for the Poroportional-Odds Failure-Time Regression Model with Interval Censoring," *Journal of the American Statistical Association*, 92, 960–967.

Huang, J. and Wellner, J. A. (1997), "Interval Censored Survival Data: a Review of Recent Progress," in *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis. Lecture notes in Statistics*, eds. Lin, D. Y. and Fleming, T. R., New York: Springer.

Huang, Y. and Wang, C. Y. (2000), "Cox Regression with Accurate Covariates Unascertainable: A Nonparametric-Correction Approach," *Journal of the American Statistical Association*, 95, pp. 1209–1219.

Hudgens, M. G. (2005), "On Nonparametric Maximum Likelihood Estimation with Interval Censoring and Left Truncation," *Journal of Royal Statistical Society. B*, 67, 573–587.

Hudgens, M. G., Maathuis, M. H., and Gilbert, P. B. (2007), "Nonparametric Estimation of the Joint Distribution of a Survival Time Subject to Interval Censoring and a Continuous Mark Variable," *Biometrics*, 63, 372–380.

Hudgens, M. G., Satten, G. A., and Ira M. Longini, J. (2001), "Nonparametric Maximum Likelihood Estimation for Competing Risks Survival Data Subject to Interval Censoring and Truncation," *Biometrics*, 57, 74–80.

Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R., and Herring, A. H. (2005), "Missing-Data Methods for Generalized Linear Models: A Comparative Review," *Journal of the American Statistical Association*, 100, 332–346.

Isomaa, B., Almgren, P., Tuomi, T., Forsén, B., Lahti, K., Nissén, M., Taskinen, M.-R., and Groop, L. (2001), "Cardiovascular Morbidity and Mortality Associated With the Metabolic Syndrome," *Diabetes Care*, 24, 683–689.

Jongbloed, G. (1998), "The Iterative Convex Minorant Algorithm for Nonparametric Estimation," *Journal of Computational and Graphical Statistics*, 7, 310–321.

Kalbfleisch, J. D. and Prentice, R. L. (2002), *The Statistical Analysis of Failure Time Data*, John Wiley & Sons, Inc.

Komárek, A. and Lesaffre, E. (2007), "Bayesian Accelerated Failure Time Model for Correlated Interval-Censored Data with a Normal Mixture as Error Distribution," *Statistica Sinica*, 17, 549–569.

— (2008), "Bayesian Accelerated Failure Time Model With Multivariate Doubly Interval-Censored Data and Flexible Distributional Assumptions," *Journal of the American Statistical Association*, 103, 523–533.

Kosorok, M. R. (2008), *Introduction to Empirical Processes and Semiparametric Inference*, Springer.

Kumar, V., Fausto, N., Abbas, A. K., Cotran, R. S., and Robbins, S. L. (2005), *Robbins and Cotran Pathologic Basis of Disease*, Saunders Elsevier, 7th ed.

Li, J. and Ma, S. (2010), "Interval-Censored Data with Repeated Measurements and a Cured Subgroup," *Journal of Royal Statistical Society*, 59, 693Ű705.

Li, Z., Gilbert, P., and Nan, B. (2008), "Weighted Likelihood Method for Grouped Survival Data in Case-Control Studies with Application to HIV Vaccine Trials," *Biometrics*, 64, 1247–1255.

Lin, D., Oakes, D., and Ying, Z. (1998), "Additive Hazards Regressioin with Current Status Data," *Biometrika*, 85, 289–298.

Lindsey, J. C. and Ryan, L. M. (1998), "Tutorial in Biostatistics: Methods for Interval-Censored Data," *Statistics in Medicine*, 17, 219–238.

Lipsitz, S. R., Ibrahim, J. G., and Zhao, L. P. (1999), "A Weighted Estimating Equation for Missing Covariate Data with Properties Similar to Maximum Likelihood," *Journal of the American Statistical Association*, 94, 1147–11160.

Lu, M., Zhang, Y., and Huang, J. (2007), "Estimation of the Mean Function with Panel Count Data Using Monotone Polynomial Splines," *Biometrika*, 94, 705–718.

— (2009), "Semiparametric Estimation Methods for Panel Count Data Using Monotone B-Splines," *Journal of the American Statistical Association*, 104, 1060–1070.

Ma, S. (2009), "Cure Model with Current Status Data," *Statistica Sinica*, 19, 233–249.

— (2010), "Mixed Case Interval Censored Data with a Cured Subgroup," *Statistica Sinica*, 20, 1165–1181.

Ma, S. and Kosorok, M. R. (2005), "Penalized Log-Likelihood Estimation for Partly Linear Transformation Models with Current Status Data," *Annals of Statistics*, 33, 2256–2290.

Miyazaki, M., Kubo, M., Kiyohara, Y., Okubo, K., andů K. Fujisawa andů Y. Hata, H. N., Tokunaga, S., Iida, M., Nose, Y., and Ishibashi, T. (2004), "Comparison of Diagnostic Methods for Diabetes Mellitus Based on Prevalence of Retinopathy in a Japanese Population: the Hisayama Study," *Diabetologia*, 47, 1411–1415.

Mokdad, A. H., Ford, E. S., Bowman, B. A., Dietz, W. H., Vinicor, F., Bales, V. S., and Marks, J. S. (2003), "PRevalence of obesity, diabetes, and obesity-related health risk factors, 2001," *Journal of American Medical Association*, 289, 76–79.

Murphy, S. A. and van der Vaart, A. W. (2000), "On Profile Likelihood," *Journal of the American Statistical Association*, 95, 449–465.

Neuhaus, J. M. (2002), "Analysis of Clustered and Longitudinal Binary Data Subject to Response Misclassification," *Biometrics*, 58, 675–683.

Oppenheim, J. E., Herr, P. R., and Carr, D. J. (1994), "Cholesterol Measurement: Test Accuracy and Factors that Influence Cholesterol Levels," Tech. rep., United States General Accounting Office.

Pan, W. (1999), "Extending the Iterative Convex Minorant Algorithm to the Cox Model for Interval Censored Data," *Journal of Computational and Graphical Statistics*, 8, 109–120.

Paulino, C. D., Soares, P., and Neuhaus, J. (2003), "Binomial Regression with Misclassification," *Biometrics*, 59, 670–675.

Pepe, M. S. (1992), "Inference Using Surrogate Outcome Data and a Validation Sample," *Biometrika*, 79, 355–365.

Pepe, M. S. and Fleming, T. R. (1991), "A Nonparametric Method for Dealing with Errors in Mismeasured Covariate Data," *Journal of the American Statistical Association*, 86, 108–113.

Pepe, M. S., Self, S. G., and Prentice, R. L. (1989), "Further Results on Covariate Measurement Errors in Cohort Studies," *Statistics in Medicine*, 8, 1167–1178.

Peto, R. (1973), "Experimental Survival Curves for Interval-Censored Data," *Applied Statistics*, 22, 86–91.

Pettitt, A. N. (1984), "Proportional Odds Models for Survival Data," *Applied Statistics*, 33, 169–175.

Prentice, R. L. (1989), "Surrogate endpoints in clinical trials: Definition and operational criteria," *Statistics in Medicine*, 8, 431–440.

Prescott, G. J. and Garthwaite, P. H. (2002), "A simple Bayesian Analysis of Misclassified Binary Data with a Validation Substudy." *Biometrics*, 58, 454–458.

Robbins, J. M. and Ritov, Y. (1997), "Toward a Curse of Dimensionality Appropriate (CODA) Asymptotic Theory for Semiparametric Models," *Statistics in Medicine*, 16, 285–319.

Robertson, T., Wright, F., and Dykstra, R. (1988), *Order Restricted Statistical Inference*, Wiley.

Robins, J. M. and Rotnitzky, A. (1995), "Semiparametric Efficiency in Multivariate Regression Models With Missing Data," *Journal of the American Statistical Association*, 90, 122–129.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994), "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed," *Journal of the American Statistical Association*, 89, 846–866.

— (1995), "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data," *Journal of the American Statistical Association*, 90, 106–121.

Rossini, A. J. and Tsiatis, A. A. (1996), "A Semiparametric Proportional Odds Regression Model for the Analysis of Current Status Data," *Journal of the American Statistical Association*, 91, 713–721.

Rotnitzky, A. and Robins, J. M. (1995), "Semiparametric Regressioin Estimation in the Presence of Dependent Censoring," *Biometrika*, 82, 805–820.

Rotnitzky, A., Robinsb, J. M., and Scharfsteinc, D. O. (1998), "Semiparametric Regressioin for Repeated Outcomes With Nonignorable Nonresponse," *Journal of the American Statistical Association*, 93, 1321–1339.

Rule, A. D., Larson, T. S., Bergstralh, E. J., Slezak, J. M., Jacobsen, S. J., and Cosio, F. G. (2004), "Using Serum Creatinine To Estimate Glomerular Filtration Rate: Accuracy in Good Health and in Chronic Kidney Disease," *Annals of Internal Medicine*, 141, 929–937.

Schafer, D. W. (1987), "Covariate Measurement Error in Generalized Linear Models," *Biometrika*, 74, 385–391.

Schick, A. and Yu, Q. (2000), "Consistency of the GMLE with Mixed Case Interval-Censored Data," *Scandinavian Journal of Statistics*, 27, 45–55.

Schrot, R. J., Patel, K. T., and Foulis, P. (2007), "Evaluation of Inaccuracies in the Measurement of Glycemia in the Laboratory, by Glucose Meters, and Through Measurement of Hemoglobin A1c," *Clinical Diabetes*, 25, 43–49.

Schwartz, J., Reichberg, S., and Gambino, R. (2005), "Glucose Testing Variability and the Need for an ExpertOversight Committee [article online:http://www.cap.org/]," CAP Today.

Sen, B. and Banerjee, M. (2007), "A Pseudolikelihood Method for Analyzing Interval Censored Data," *Biometrika*, 94, 71–86.

Shen, X. (2000), "Linear Regression with Current Status Data," *Journal of the American Statistical Association*, 95, 842–852.

Sun, J. (2006), *The Statistical Analysis of Interval-Censored Failure Time Data*, Springer.

Sun, J. and Kalbfleisch, J. D. (1995), "Estimation of the Mean Function of Point Processes Based on Panel Count Data," *Statistica Sinica*, 5, 279–290.

Sun, J. and Wei, L. J. (2000), "Regressioin Analysis of Panel Count Data with Covariate-Dependent Observation and Censoring Times," *Journal of Royal Statistical Society. Series B*, 62, 293–302.

The Expert Committee on the Diagnosis & Classification of Diabetes Mellitus (2003), "Report of the Expert Committee on the Diagnosis and Classification of Diabetes Mellitus," *Diabetese Care*, 37, S5–S20.

the National Cholesterol Education Program Expert Panel (2001), "Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III)," .

Tian, L. and Cai, T. (2006), "On the Accelerated Failure Time for Current Status and Interval Censored Data," *Biometrika*, 93, 329–342.

Tonyushkina, K. and Nichols, J. H. (2009), "Glucose Meters: A Review of Technical Challenges to Obtaining Accurate Results," *Journal of Diabetes Science and Technology*, 3, 971–980.

Tsiatis, A. A. and Davidian, M. (2001), "A Semiparametric Estimator for the Proportional Hazards Model with Longitudinal Covariates Measured with Error," *Biometrika*, 88, 447–458.

Tsiatis, A. A., DeGruttola, V., and Wulfsohn, M. S. (1995), "Modeling the Relationship of Survival to Longitudinal Data Measured with Error: Applications to Survival and CD4 Counts in Patients with AIDS," *Journal of the American Statistical Association*, 90, 27–37.

Turnbull, B. W. (1976), "The Empirical Eistribution with Arbitrarily Grouped Censored and Truncated Data," *Journal of Royal Statistical Society*, 38, 290–295.

UK Prospective Diabetes Study Group (1998), "Tight Blood Pressure Control and Risk of Macrovascular and Microvascular Complications in Type 2 Diabetes: UKPDS 38," *British Medical Journal*, 317, 703–713.

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

Vasan, R. S. (2006), "Biomarkers of Cardiovascular Disease : Molecular Basis and Practical Considerations," *Journal of the American Heart Association: Circulation*, 113, 2335–2362.

Wellner, J. A. and Zhang, Y. (2000), "Two Estimators of the Mean of a Counting Process with Panel Count Data," *Annals of Statistics*, 28, 779–814.

— (2007), "Two Likelihood-Based Semiparametric Estimation Methods for Panel Count Data with Covariates," *The Annals of Statistics*, 35, 2106–2142.

Wen, C.-C. (2012), "Cox regression for Mixed Case Interval-Censored Data with Covariate Errors," *Lifetime Data Analysis*, 18, 321–338.

Wu, C. J. (1983), "On the Convergence Properties of the EM Algorithm," *Annals of Statistics*, 11, 95–103.

Xue, H., Lam, K. F., and Li, G. (2004), "Sieve Maximum Likelihood Estimator for Semiparametric Regression Models with Current Status Data," *Journal of the American Statistical Association*, 99, 346–356.

Zeng, D., Cai, J., and Shen, Y. (2006), "Semiparametric Additive Risks Model for Interval-Censored Data," *Statistica Sinica*, 16, 287–302.

Zeng, D., Yin, G., and Ibrahim, J. G. (2005), "Inference for a Class of Transformed Hazards Models," *Journal of the American Statistical Association*, 100, 1000–1008.

Zhan, Y. and Wellner, J. A. (1995), "Double Censoring: Characterization and Computation of the Nonparametric Maximum Likelihood Estimator," Tech. rep., University of Washington.

Zhang, Y. and Jamshidian, M. (2004), "On Algorithm for NPMLE of the failure function with Censored Data," *Journal of Computational and Graphical Statistics*, 13, 123–140.

Zhang, Y. L. and Newton, M. A. (1997), "On Calculating the Nonparametric Maximum Likelihood Estimator of a Distribution Given Interval Censored Data," Tech. Rep. 116, Department of Biostatistics, University of Wisconsin in Madison.

Zhang, Z. and Sun, J. (2010), "Interval Censoring," *Statistical Methods in Medical Research*, 19, 53–70.

Zhao, L. P., Lipsitz, S. R., and Lew, D. (1996), "Regression Analysis with Missing Covariate Data Using Estimating Equations," *Biometrics*, 52, 1165–1182.