ROBUST ESTIMATION FOR LONGITUDINAL DATA WITH INFORMATIVE OBSERVATION TIMES

LIU KIN YAT

M.Phil

The Hong Kong Polytechnic University

2015

The Hong Kong Polytechnic University

Department of Applied Mathematics

Robust Estimation for Longitudinal Data with Informative Observation Times

LIU, KIN YAT

A thesis submitted in partial fulfillment of
the requirements for the degree of Master of Philosophy

July 2014

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree of diploma, except where du acknowledgement has been make in the text.

__________________ (Signed)

LIU, Kin Yat (Name of student)

Abstract

In this thesis, we focus on regression analysis of longitudinal data that often occur in medical follow-up studies and observational investigations. The analysis of these data involves two processes. One is the underlying recurrent event process of interest and the other is the observation process that controls observation times. Most of the existing methods, however, rely on some restrictive models or assumptions such as the Poisson assumption. For this, we propose a more general and robust estimation approach for regression analysis of longitudinal data with related observation times. The asymptotic properties of the proposed estimators are established and numerical studies indicate that the proposed method works well for practical situations.

Keywords: Estimating equation; Informative observation process; Longitudinal data; Model checking; Robust estimation.

Acknowledgements

I would like to express my sincerest gratitude to my supervisor, Dr. Zhao Xingqiu for her guidance and patience during my M.Phil graduate studies. I am deeply grateful to her for introducing me to this research area, and helping me to proceed. It was very fortunate to have her as my supervisor. Dr. Zhao taught me not only the topic of robustness in mathematical statistics but also the importance of being robust as a researcher.

I also owe my deepest gratitude to my co-supervisor, Professor Chen Xiaojun, for her support and encouragement. Without her, this thesis would not even start.

I am grateful to my families for their love and support. My wife, Ruby Liu, and my two amazing sons, Abraham, and Noah are always fuel to my life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This dissertation discusses the statistical analysis of longitudinal data with informative observation times. First, we will propose a general model for longitudinal data with informative observation times, provide a robust estimation approach for regression parameters, and discuss the properties of the estimators. Next, we will study the diagnostic tests for the proposed model. In the third part, finite sample properties of the proposed estimators are studied via Monte Carlo experiments. Forth, we analyze the bladder cancer study data conducted by the Veterans Administration Cooperative Urological Research Group (VACURG) with the proposed model and apply the diagnostic tests to assess the appropriateness of the model. Finally, conclusions and related future research are presented.

1.1 Motivation

In medical follow-up studies and social-demographical studies, longitudinal data arise naturally. The most obvious feature of longitudinal data is the repeated observations on some variables. In addition, observation times and frequencies among subjects are different. In the bladder cancer study, Sun and Wei (2000) and Zhang (2002) discussed a set of longitudinal data arising from a bladder cancer follow-up study conducted by the Veterans Administration Cooperative Urological Research Group (VACURG).

In the bladder cancer study, 118 patients with superficial bladder tumors were selected. After their tumors were removed, each patient was allocated randomly to one of the three treatments. There were 48 patients allocated to placebo, 38 to thiotepa, and 32 to pyridoxine. In the study, the patients visited clinical centers on a regular basis, and if new tumors were found, the number of bladder tumors was recorded and then the tumors were removed. New tumors had been found subsequently among many patients since the removal of tumors in the beginning of the study. Visiting times were not the same of all patients and the censoring times vary among patients. The clinical visit times, the number of recurrent tumors between clinical visits, the number of initial tumors, the size of the largest initial tumor, and the type of treatment were recorded for each patient in the study.

An objective of the study is to assess the treatment effect on the tumor recurrence rate. Sun and Wei (2000); Wellner and Zhang (2000); Zhang (2002); Wellner and Zhang (2007) among others have analyzed the data under the assumption that the observation times were noninformative. Patients in thiotepa group, however, tend to have clinical visit more often than those in the placebo group. One possibility is that tumor recurrence rate in thiotepa group is relatively higher than other groups and hence, patients needed to consult doctors more often. The correlation between the tumor recurrence rate and the observation time should not be ignored. Ignorance of the dependence may need to estimation biased. Analysis of the bladder cancer study data based on the assumption that the observation times follow a Poisson process were provided in Hu et al. (2003), Sun et al. (2005), Li et al. (2010), and Zhao and Tong (2011). Such assumption may be appropriate for some applications. Nevertheless, there is no theoretical justification of the appropriateness of such assumption. Hence, the assumption may not be realistic in other situations.

1.2 Literature Review

The analysis of longitudinal data has recently attracted considerable attention. These data frequently occur in medical follow-up studies and observational investigations. For the analysis of longitudinal data, a number of methods have been developed, mostly under the assumption that the longitudinal response process and the observation process are independent completely, or conditionally independent given covariates. For example, Diggle et al. (1994) presented an excellent summary on some commonly used methods such as the estimating equation and random-effect model approaches. Lin and Ying (2001) and Welsh et al. (2002) discussed general semiparametric regression analysis of longitudinal data when both observation times and the censoring times could depend on covariates.

A common situation where observation times are informative is that those observation times are either subject-dependent or response variable-dependent. In a bladder cancer study, Sun and Wei (2000) and Zhang (2002) discussed a set of longitudinal data arising from a bladder cancer follow-up study conducted by the Veterans Administration Cooperative Urological Research Group; in this study, some patients had significantly more clinical visits than others and thus the occurrence of bladder tumors of a patient and the visit times may be related. Lipsitz et al. (2002) presented a set of longitudinal data from a study of children with acute lymphoblastic leukemia that involved correlated response and observation processes. The same could be true for other medical follow-up studies, but there is limited research on the analysis of longitudinal data when the longitudinal response process of interest may be correlated with the observation process given the covariates. That is, the observation times may be informative. Sun et al. (2005) studied semiparametric models that allow observation times to be correlated with the longitudinal process; Sun et al. (2007) proposed a joint model for the longitudinal process and the observation process, where both processes may be correlated through a shared latent variable or frailty, and used the estimating equation approach to estimate the regression parameters; Liang et al. (2009) discussed a joint model through two random effects, where the relationship between the random effects is specified and a parametric distribution assumption for a random effect is required. A common and key assumption of these methods is that the observation process is a Poisson process.

The aim of this thesis is to consider more general joint models for longitudinal data with dependent observation times, to develop an estimating equation approach for estimation of regression parameters, and to establish the asymptotic properties of the estimators.

1.3 Outline of Thesis

The thesis is organized as follows. In Chapter 2, we will begin with introducing the notation and assumptions and the models. A robust estimation procedure is presented for the parameters of interest and the asymptotic properties of the resulting estimators are established. In Chapter 3, a model checking procedure is presented. Chapter 4 reports some simulation results obtained for assessing the finite sample properties of the proposed estimates. The bladder cancer study data is analyzed in Chapter 5. Chapter 6 concludes with some discussion and remarks.

Chapter 2

Longitudinal Data Analysis with Informative Observation Times

In this chapter, a semiparametric joint model of longitudinal data with informative observation times is developed. Next, estimators of regression parameters in the model are proposed. The asymptotic properties of the proposed estimators, including the consistency, rate of convergence, and asymptotic normality are then presented.

2.1 Introduction

Studies based on longitudinal data analysis often assume that the observation process is independent of the longitudinal outcome process. The observation process, however, may be correlated with the longitudinal outcomes in practice. Hu et al. (2003), Sun et al. (2005), Li et al. (2010), and Zhao and Tong (2011) among other proposed models that consider the correlation between the longitudinal outcomes and the observation process. Most of these models, however, assume that the observation process is a Poisson process.

In this chapter, we develop a joint model that, given the covariates, the longitudinal outcome and the observation process can be correlated and their relationship is specified by a link function and a latent variable, while the link function and the distributional form of the latent variable are unspecified.

2.2 Statistical Models

Consider a longitudinal study that consists of n independent subjects and let $Y_i(\mathrm{t})$ denote the longitudinal response variable of interest before or at time t for subject i. Suppose that for each subject, there exists a p-dimensional vector of covariates denoted by X_i. Given X_i and an unobserved positive random variable Z_i that is independent of X_i, the mean function of $Y_i(t)$ has the form

$$E\{Y_i(t)|X_i, Z_i\} = \mu_0(t) + X_i'\beta + g(Z_i) \tag{2.1}$$

Here, $\mu_0(t)$ is a completely unknown continuous baseline mean function, β is a vector of unknown regression parameters, and $g(\cdot)$ is a completely unspecified link function.

For subject i, suppose that $Y_i(\cdot)$ is observed only at finite time points $T_{i1} < \cdots < T_{iK_i}$, where K_i denotes the potential number of observation times, $i = 1, \ldots, n$. That is, only the values of $Y_i(t)$ at these observation times are known and we have panel count data on the $Y_i(t)$'s. Let C_i denote the follow-up time associated with subject i and thus $Y_i(t)$ is observed only at these T_{ij}'s with $T_{ij} \leq C_i$, $i = 1, \ldots, n$. Define $\tilde{O}_i(t) = O_i(\min(t, C_i))$, where $O_i(t) = \sum_{j=1}^{K_i} I(T_{ij} \leq t)$, $i = 1, \ldots, n$. Then $\tilde{O}_i(t)$ is a point process characterizing the ith subject's observation process and jumps only at the observation times.

For the observation process, we will assume that $O_i(t)$ satisfies the following rate function model

$$E\{dO_i(t)|X_i, Z_i\} = Z_i h(X_i) d\Lambda_0(t), \tag{2.2}$$

where $h(\cdot)$ is a completely unspecified positive function as $g(\cdot)$ and $\Lambda_0(\cdot)$ is a completely unknown continuous baseline function. One may consider model (2.2) as the generalization of a non-homogeneous Poisson process with the intensity function

$$\lambda(t|X_i, Z_i) = Z_i h(X_i) d\Lambda_0(t), \tag{2.3}$$

where $h(X_i) = \exp(X_i'\gamma)$ and $d\Lambda_0(t) = \lambda_0(t)$ which is a baseline intensity function.

Under model (2.2), one does not need the Poisson assumption anymore. In the following, $Y_i(t)$ and $O_i(t)$ are assumed to be independent given (X_i, Z_i). Also C_i is independent of $\{Y_i, O_i, X_i, Z_i\}$ and $\{Y_i(t), O_i(t), C_i, X_i, 0 \leq t \leq \tau\}_{i=1}^{n}$ are independent and identically distributed, where τ denotes the length of the study. Here, the main goal is to estimate regression parameter β.

2.3 Inference Procedure

To estimate β, note that if the latent variables Z_i's are known, model (2.1) would become the usual linear mean model. Unfortunately, the Z_i's are unknown in practice. One natural way for this is to estimate the Z_i's first and then treat them as known. In the following, we take a different approach motivated by that proposed in Sun and Wei (2000) among others.

Specifically, define

$$\bar{Y}_i = \sum_{j=1}^{m_i} Y_i(T_{ij}) I(T_{ij} \leq \tau) = \int_0^{\tau} Y_i(t) d\tilde{O}_i(t),$$

where $m_i = \tilde{O}_i(C_i)$, the total number of observations on subject $i, i = 1, \dots, n$. Then, we have the following results.

Theorem 2.1

$$E(\bar{Y}_i|X_i) = E(Z_i)E\big(\Lambda_0(C_i)\big)h(X_i)(X_i'\beta)$$

$$+ h(X_i)\int_0^\tau [E(Z_i)\mu_0(t) + E\{g(Z_i)Z_i\}] \Pr(C_i \geq t)\, d\Lambda_0(t)$$

and

$$E(m_i|X_i) = E(Z_i)E\{\Lambda_0(C_i)\}h(X_i).$$

Proof of Theorem 2.1 The approach is first finding $E(\bar{Y}_i|X_i, Z_i)$, and then applying the law of total expectation to obtain $E(\bar{Y}_i|X_i)$. Similarly, we first obtain $E(m_i|X_i, Z_i)$, and apply the law of total expectation to obtain $E(m_i|X_i)$.

$$E(\bar{Y}_i|X_i, Z_i) = E\left\{\int_0^\tau Y_i(t)d\tilde{O}_i(t)\,|X_i, Z_i\right\}$$

$$= E\left\{\int_0^\tau Y_i(t)I(t \leq C_i)dO_i(t)\,|X_i, Z_i\right\}$$

$$= \int_0^\tau E\{Y_i(t)I(t \leq C_i)dO_i(t)|X_i, Z_i\}$$

$$= \int_0^\tau E\{Y_i(t)|X_i, Z_i\}E\{I(t \leq C_i)|X_i, Z_i\}E\{dO_i(t)|X_i, Z_i\}$$

$$= \int_0^\tau \big(\mu_0(t) + X_i'\beta + g(Z_i)\big)\Pr(t \leq C_i)\, Z_i h(X_i)d\Lambda_0(t)$$

$$= \int_0^\tau \mu_0(t) \Pr(t \leq C_i) Z_i h(X_i)\, d\Lambda_0(t) + \int_0^\tau X_i'\beta \Pr(t \leq C_i)\, Z_i h(X_i) d\Lambda_0(t)$$

$$+ \int_0^\tau g(Z_i) \Pr(t \leq C_i)\, Z_i h(X_i) d\Lambda_0(t)$$

$$= Z_i h(X_i) \int_0^\tau \mu_0(t) \Pr(t \leq C_i)\, d\Lambda_0(t) + X_i'\beta Z_i h(X_i) \int_0^\tau \Pr(t \leq C_i)\, d\Lambda_0(t)$$

$$+ g(Z_i) Z_i h(X_i) \int_0^\tau \Pr(t \leq C_i)\, d\Lambda_0(t).$$

Next, applying the law of total expectation, we have

$$E(\bar{Y}_i | X_i) = E\{E(\bar{Y}_i | X_i, Z_i) | X_i\}$$

$$= E\left\{ Z_i h(X_i) \int_0^\tau \mu_0(t) \Pr(t \leq C_i)\, d\Lambda_0(t) + X_i'\beta Z_i h(X_i) \int_0^\tau \Pr(t \leq C_i)\, d\Lambda_0(t) \right.$$

$$\left. + g(Z_i) Z_i h(X_i) \int_0^\tau \Pr(t \leq C_i)\, d\Lambda_0(t) \right| X_i \right\}$$

$$= E(Z_i) h(X_i) \int_0^\tau \mu_0(t) \Pr(t \leq C_i)\, d\Lambda_0(t) + X_i'\beta E(Z_i) h(X_i) \int_0^\tau \Pr(t \leq C_i)\, d\Lambda_0(t)$$

$$+ E\{g(Z_i) Z_i\} h(X_i) \int_0^\tau \Pr(t \leq C_i)\, d\Lambda_0(t)$$

$$= h(X_i) \int_0^\tau [E(Z_i)\mu_0(t) + E\{g(Z_i) Z_i\}] \Pr(t \leq C_i)\, d\Lambda_0(t)$$

$$+ X_i'\beta E(Z_i) h(X_i) E\big(\Lambda_0(C_i)\big).$$

For the second equation, first note that $m_i = \tilde{O}_i(C_i) = O_i(\min(C_i, C_i)) = O_i(C_i) = \sum_{j=1}^{K_i} I\big(T_{ij} \leq C_i\big)$. Also, $E\{dO_i(t) | X_i, Z_i\} = Z_i h(X_i) d\Lambda_0(t)$ and $m_i = O_i(C_i) = \int_0^{C_i} dO_i(t)$. Hence, we have

$$
\begin{aligned}
E(m_i|X_i, Z_i) &= E\left\{\int_0^{C_i} dO_i(t)\middle| X_i, Z_i\right\} \\
&= \int_0^{C_i} E\{dO_i(t)|X_i, Z_i\} \\
&= \int_0^{C_i} Z_i h(X_i) d\Lambda_0(t).
\end{aligned}
$$

Applying the law of total expectation, we have

$$
\begin{aligned}
E(m_i|X_i) &= E\{E(m_i|X_i, Z_i)|X_i\} \\
&= E\left\{\int_0^{C_i} Z_i h(X_i) d\Lambda_0(t)\middle| X_i\right\} \\
&= h(X_i) E\left\{Z_i \int_0^{C_i} d\Lambda_0(t)\middle| X_i\right\} \\
&= h(X_i) E\{Z_i \Lambda_0(C_i)\} \\
&= h(X_i) E\{Z_i\} E\{\Lambda_0(C_i)\}
\end{aligned}
$$

Hence,

$$
h(X_i) = \frac{E(m_i|X_i)}{E\{Z_i\}E\{\Lambda_0(C_i)\}}
$$

■

Theorem 2.2

$$
E(\bar{Y}_i|X_i) = E(m_i|X_i)(X_i'\beta + \theta),
$$

where

$$\theta = \frac{\int_0^\tau [E(Z_i)\mu_0(t) + E\{g(Z_i)Z_i\}]P(C_i \geq t)d\Lambda_0(t)}{E(Z_i)E\big(\Lambda_0(C_i)\big)},$$

is an unknown parameter.

Proof of Theorem 2.2

$$E(\bar{Y}_i|X_i) = E(Z_i)E\big(\Lambda_0(C_i)\big)h(X_i)(X_i'\beta) + h(X_i)\int_0^\tau [E(Z_i)\mu_0(t) + E\{g(Z_i)Z_i\}]\,\Pr(C_i \geq t)\,d\Lambda_0(t)$$

$$= E(m_i|X_i)(X_i'\beta) + \frac{E(m_i|X_i)}{E\{Z_i\}E\{\Lambda_0(C_i)\}}\int_0^\tau [E(Z_i)\mu_0(t) + E\{g(Z_i)Z_i\}]\,\Pr(C_i \geq t)\,d\Lambda_0(t)$$

$$= E(m_i|X_i)\left[X_i'\beta + \frac{\int_0^\tau [E(Z_i)\mu_0(t) + E\{g(Z_i)Z_i\}]\,\Pr(C_i \geq t)\,d\Lambda_0(t)}{E\{Z_i\}E\{\Lambda_0(C_i)\}}\right]$$

$$= E(m_i|X_i)(X_i'\beta + \theta)$$

■

For estimation of β, motivated by the equation above, we propose to use the following class of estimating functions

$$U(\beta_1) = \sum_{i=1}^{n} W_i X_{1i}\{\bar{Y}_i - m_i X_{1i}'\beta_1\} = 0, \qquad (2.4)$$

where the W_i's are some weights that could depend on X_i, $X_{1i}' = (X_i', 1)$ and $\beta_1' = (\beta', \theta)$.

Let $\hat{\beta}_1 = \big(\hat{\beta}', \hat{\theta}\big)'$ denote the solution to equation (2.4). Then we have

$$\hat{\beta}_1 = \left[\sum_{i=1}^{n} W_i m_i X_{1i} X'_{1i}\right]^{-1} \sum_{i=1}^{n} W_i X_{1i} \bar{Y}_i.$$

2.4 Asymptotic Theory

In this section, we will sketch the proofs for the consistency and asymptotic normality of the proposed estimate $\hat{\boldsymbol{\beta}}_{\mathbf{1}}$. For this, we will employ the notation defined in the previous sections and assume that $\mathbf{Pr}(\boldsymbol{C_i} \geq \boldsymbol{\tau}) > 0$. We also define $\boldsymbol{\Gamma} = \boldsymbol{E}\{\boldsymbol{W_i m_i X_{1i} X'_{1i}}\}$ and assume that $\boldsymbol{\Gamma}$ is positive definite.

First we will consider the consistency of $\hat{\beta}_1$. We first prove two lemmas.

Lemma 1

$$\frac{U(\beta_{10})}{n} = \frac{\sum \phi_i}{n} \xrightarrow{p} 0$$

where $\phi_i = W_i X_{1i}\{\bar{Y}_i - m_i \beta'_{10} X_{1i}\}$.

Lemma 2

$$\frac{1}{n}\frac{\partial}{\partial \beta_1} U(\beta_1) = -\frac{1}{n}\sum_{i=1}^{n} W_i m_i X_{1i} X'_{1i}$$

converges uniformly to a negative matrix $-E\{W_i m_i X_{1i} X_{1i}'\}$ over β_1 for any value of β_{10}.

Proof of lemma 2 $W_i m_i X_{1i} X_{1i}'$ are independent for $i = 1, \ldots, n$. By the law of large number, $\frac{1}{n}\sum_{i=1}^{n} W_i m_i X_{1i} X_{1i}'$ converges to the expectation $E\{W_i m_i X_{1i} X_{1i}'\}$.

■

Corollary 1

The solution $\hat{\beta}_1$ of the estimating equation $U(\beta_1) = 0$ is unique and consistent.

Now we turn to prove the asymptotic normality of the proposed estimator $\hat{\beta}_1$.

Theorem 2.3 (Asymptotic Normality of $\hat{\boldsymbol{\beta}}_1$)

$\sqrt{n}(\hat{\beta}_{1n} - \beta_{10})$ has an asymptotically normal distribution with mean zero and covariance matrix $\Gamma^{-1}\Sigma(\Gamma^{-1})'$.

Proof of theorem 2.3

Obtaining Taylor series expansion of $U(\beta)$ at β_{10}, we have

$$U(\beta_1) = U(\,\beta_{10}\,) + \frac{\partial U(\beta_1)}{\partial \beta_1}(\beta_1 - \beta_{10}) + o_p(1)$$

Since $U(\hat{\beta}_{1n}) = 0$, we have

$$0 = U(\hat{\beta}_{1n}) = U(\,\beta_{10}\,) + \frac{\partial U(\hat{\beta}_{1n})}{\partial \beta_1}(\hat{\beta}_{1n} - \beta_{10}) + o_p(1)$$

$$\frac{\partial U(\hat{\beta}_{1n})}{\partial \beta_1}(\hat{\beta}_{1n} - \beta_{10}) = -U(\beta_{10}) + o_p(1)$$

$$\frac{1}{n}\frac{\partial U(\hat{\beta}_{1n})}{\partial \beta_1}\sqrt{n}(\hat{\beta}_{1n} - \beta_{10}) = -\frac{U(\beta_{10})}{\sqrt{n}} + o_p(1)$$

$$\Gamma\sqrt{n}(\hat{\beta}_{1n} - \beta_{10}) = \frac{U(\beta_{10})}{\sqrt{n}} + o_p(1)$$

$$\sqrt{n}(\hat{\beta}_{1n} - \beta_{10}) = \frac{\Gamma^{-1}}{\sqrt{n}}U(\beta_{10}) + o_p(1)$$

$$\sqrt{n}(\hat{\beta}_{1n} - \beta_{10}) = \frac{\Gamma^{-1}}{\sqrt{n}}\sum_{i=1}^{n}\phi_i + o_p(1)$$

where $\phi_i = W_i X_{1i}\{\bar{Y}_i - m_i\beta_{10}' X_{1i}\}$.

It thus follows that $\sqrt{n}(\hat{\beta}_{1n} - \beta_{10})$ has an asymptotically normal distribution with mean zero and covariance matrix $\Gamma^{-1}\Sigma(\Gamma^{-1})'$ that can be consistently estimated by $\hat{\Gamma}^{-1}\hat{\Sigma}(\hat{\Gamma}^{-1})'$ where $\Sigma = E\{\phi_i\phi_i'\}$ and $\hat{\Gamma}$ and $\hat{\Sigma}$ are given as

$$\hat{\Gamma} = \frac{1}{n}\sum_{i=1}^{n}\{W_i m_i X_{1i} X_{1i}'\}$$

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\hat{\phi}_i\hat{\phi}_i'$$

with

$$\hat{\phi}_i = W_i X_{1i}\{\bar{Y}_i - m_i\hat{\beta}_1' X_{1i}\}.$$

■

The finite sample properties of the proposed estimator will be studied in chapter 4 via a simulation study.

Chapter 3

Model Diagnostics

In practice, in addition to the estimation of β, one may also be interested in checking the adequacy of models given the observed data. To develop a procedure for this, define

$$\mathcal{A}(t) = \frac{\int_0^t [E(Z_i)\mu_0(u) + E\{g(Z_i)Z_i\}]P(C_i \geq u)d\Lambda_0(u)}{E(Z_i)E\big(\Lambda_0(C_i)\big)}.$$

Since $E\left(\int_0^t \{Y_i(u) - \beta_0' X_i\}d\tilde{O}_i(u)\,|X_i\right) = E(m_i|X_i)\mathcal{A}(t)$, we can estimate $\mathcal{A}(t)$ by

$$\hat{\mathcal{A}}(t) = \frac{\sum_{i=1}^n \int_0^t \{Y_i(u) - \hat{\beta}' X_i\}d\tilde{O}_i(u)}{\sum_{i=1}^n m_i}$$

Furthermore, for each i, $i = 1, \dots, n$, define the residual

$$\hat{R}_i(t) = \int_0^t \{Y_i(u) - \hat{\beta}' X_i\}d\tilde{O}_i(u) - m_i\hat{\mathcal{A}}(t).$$

3.1 Function Form of Covariates

To check the functional form for the jth component of X in (2.1) formally, we consider the process

$$\mathcal{F}_j(x) = \frac{1}{\sqrt{n}}\sum_{i=1}^n I\big(X_{ji} \leq x\big)\tilde{R}_i,$$

where $\tilde{R}_i = \hat{R}_i(\tau)$.

Let

$$S_0 = \frac{1}{n}\sum_{i=1}^{n} m_i,$$

$$S_j(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_{ji} \le x) m_i,$$

and

$$B_j(\tau, x) = \frac{1}{n}\sum_{i=1}^{n}\left\{I(X_{ji} \le x) - \frac{S_j(x)}{S_0}\right\} X_i \tilde{O}_i(\tau).$$

To apply the statistics $\mathcal{F}_j(x)$, we rely on the following theorem.

Theorem 3.1 (Approximation of the Null Distribution of $\boldsymbol{\mathcal{F}_j(x)}$)

The null distribution of $\mathcal{F}_j(x)$ can be approximated the zero-mean Gaussian process

$$\hat{\mathcal{F}}_j(x) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{I(X_{ji} \le x) - \frac{S_j(x)}{S_0}\right\}\tilde{R}_i\, G_i - B_j(\tau, x)' \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{d}_i\, G_i,$$

where $\hat{d}_i$ is the vector $\hat{\Gamma}^{-1}\hat{\phi}_i$ without the last entry and $(G_1, \dots, G_n)$ are independent standard normal variables independent of the data.

Proof of Theorem 3.1

Assume that the limits of S_0, $S_j(x)$, and $B_j(\tau, x)$ exist and are denoted by s_0, s_j, and $b_j(\tau, x)$, respectively. Define

$$R_i = \int_0^{\tau}\{Y_i(u) - X_i'\beta\} d\tilde{O}_i(u) - m_i \mathcal{A}(\tau).$$

To prove the weak convergence of $\phi(t, x)$, first using Lemma A.1 of Lin and Ying (2001) and functional version of the Taylor expansion, we have

$$\mathcal{F}_j(x) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{I\big(X_{ji} \le x\big) - \frac{s_j(x)}{s_0}\right\}R_i - b_j(\tau, x)'\sqrt{n}\big(\hat{\beta} - \beta_0\big) + o_p(1).$$

The tightness of the first term on the right-hand side of the above follows directly from the arguments in Appendix A.5 of Lin et al. (2000). The second term is also tight because $\sqrt{n}\big(\hat{\beta} - \beta_0\big)$ converge in distribution and $b(\tau, x)$ is a deterministic function. Thus, $\mathcal{F}_j(x)$ is tight. Let d_i be the vector $\Gamma^{-1}\phi_i$ without the last entry. Then, we can further write $\mathcal{F}_j(x)$ as

$$\mathcal{F}_j(x) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{I\big(X_{ji} \le x\big) - \frac{s_j(x)}{s_0}\right\}R_i - b_j(\tau, x)'\frac{1}{\sqrt{n}}\sum_{i=1}^{n} d_i + o_p(1).$$

It thus follows from the multivariate central limit theorem and the tightness of $\mathcal{F}_j(x)$ that $\mathcal{F}_j(x)$ converges weakly to a zero-mean Gaussian process that can be approximated by the zero-mean Gaussian process

$$\hat{\mathcal{F}}_j(x) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{I\big(X_{ji} \le x\big) - \frac{S_j(x)}{S_0}\right\}\tilde{R}_i\, G_i - B_j(\tau, x)'\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \hat{d}_i\, G_i.$$

■

Thus, using the simulation approach presented in Lin et at. (2000), the null distribution of $\mathcal{F}_j(x)$ can be approximated by that of $\hat{\mathcal{F}}_j(x)$. In other words, one can approximate the distribution of $\mathcal{F}_j(x)$ by the empirical distribution of a large number of realizations of $\hat{\mathcal{F}}_j(x)$ given by repeatedly generating the standard normal random sample $(G_1, \dots, G_n)$ given the observed data. To assess the functional form of the jth component of covariates, one can plot a few realizations, say 20, from $\mathcal{F}_j(x)$ along with the observed $\mathcal{F}_j(x)$ to see if they can be regarded as arising from the same population. More formally, we can apply the supremum test statistic $\sup_x|\mathcal{F}_j(x)|$, where the p-value can be obtained by comparing the observed value of $\sup_x|\mathcal{F}_j(x)|$ to a large number of realizations of $\sup_x|\hat{\mathcal{F}}_j(x)|$.

3.2 Goodness-of-Fit Test

To test the goodness-of-fit of models (2.1) and (2.2), we apply the statistic

$$\phi(t,x) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} I(X_i \le x)\hat{R}_i(t),$$

where the event $I(X_i \le x)$ means that each of the components of X_i is not larger than the corresponding component of x. Note that $\phi(t,x)$ is the cumulative sum of $\hat{R}_i(t)$ over the values of X_i's. Similar to $\mathcal{F}_j(x)$ and $\mathcal{F}_g(x)$, the null distribution of $\phi(t,x)$ can be approximated by the zero-mean Gaussian process

$$\hat{\phi}(t,x) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{I(X_i \le x) - \frac{S^{(1)}(x)}{S_0}\right\}\hat{R}_i(t)\,G_i - B(t,x)'\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{d}_i\,G_i$$

where

$$S_0 = \frac{1}{n}\sum_{i=1}^{n} m_i,$$

$$S^{(1)}(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x)m_i,$$

and

$$B(t,x) = \frac{1}{n}\sum_{i=1}^{n}\left\{I(X_i \leq x) - \frac{S^{(1)}(x)}{S_0}\right\} m_i X_i.$$

Similar to $\mathcal{F}_j(x)$, one can approximate the distribution of $\phi(t,x)$ by the empirical distribution of a large number of realizations of $\hat{\phi}(t,x)$ given by repeatedly generating the standard normal random sample $(G_1, \dots, G_n)$ given the observed data. Thus for checking the overall fit of models (2.1) and (2.2) based on $\phi(t,x)$, the p-value of the omnibus test can be obtained by comparing the observed value of $\sup_{t,x}|\phi(t,x)|$ to a large number of realization of $\sup_{t,x}\left|\hat{\phi}(t,x)\right|$.

Chapter 4

Simulation

We conducted three simulation studies to assess the performances of the proposed inference procedure and the diagnostic tests. The purpose of the first one was to evaluate the finite sample properties of the proposed estimator, while in the second study, we compared the proposed estimator to that given in Zhao et al. (2012). In the third simulation study, we evaluate the sizes and powers of the diagnostic tests when the sample size is finite.

4.1 Finite Sample Properties of $\hat{\boldsymbol{\beta}}$

For the first study, we considered the situation where there exist two covariates, X_{i1}'s and X_{i2}'s which are the Bernoulli distribution with success probability 0.5 and uniform distribution over interval $(0, 1)$ respectively. The latent variable Z_i's were generated from the gamma distribution with shape parameter 10 and scale parameter 10. Also, we consider two cases of $g(Z_i)$. The first one is $g(Z_i) = \rho(Z_i - E[Z_i])/\sqrt{Var[Z_i]}$ which is a linear function, while the second is $g(Z_i) = \rho(\ln Z_i - E[\ln Z_i])/\sqrt{Var[\ln Z_i]}$ which is a nonlinear function. Here ρ characterizes the relationship between the observation process and the longitudinal response process. When $\rho > 0$, the two processes are positively correlated; when $\rho = 0$, the two processes have no correlation given the covariates; when $\rho < 0$, the two processes are negatively correlated. Here, three situations with $\rho = -0.5$, 0, and 0.5 were considered. The follow-up time C_i is generated from the uniform distribution over $[\tau/2, \tau]$ with $\tau = 18$.

With respect to the observation process $O_i(t)$, three set-ups were considered as follows:

a. For the first case, given X_i, Z_i, and C_i, the number of observation times m_i was assumed to follow the Poisson distribution with mean

$$\Lambda(C_i|X_i, Z_i) = Z_i\Lambda_0(C_i)\exp(X_i'\gamma) = \frac{Z_iC_i\exp(X_i'\gamma)}{\tau},$$

$i = 1, 2, \dots, n$, where $\gamma = (1,1)'$ was considered. The observation times $(T_{i1}, \dots, T_{im})$ were taken to be the order statistics of a random sample of size m_i from the uniform distribution over $(0, C_i)$.

b. For the second case, given X_i, Z_i, and C_i, the number of observation times m_i was assumed to follow the Poisson distribution with mean

$$\Lambda(C_i|X_i, Z_i) = Z_i\Lambda_0(C_i)\exp(X_i'\gamma) = \frac{Z_iC_i\left(\frac{C_i}{2}+1\right)\exp(X_i'\gamma)}{\tau\left(\frac{\tau}{2}+1\right)},$$

$i = 1, 2, \dots, n$, where $\gamma = (1,1)'$ was considered. The observation times $(T_{i1}, \dots, T_{im})$ were taken to be the order statistics of a random sample of size m_i from the cumulative function

$$\frac{\frac{t^2}{2}+t}{\frac{C_i^2}{2}+C_i}I(0 \le t \le C_i).$$

c. For the third case, given X_i, Z_i, and C_i, the interarrival times were assumed to follow Weibull distribution with shape parameter 0.5 and scale parameter

$\frac{Z_i C_i \exp(X_i'\gamma)}{\tau^2}$. The m_i was the number of observation times which were less than C_i.

For the response variable, it was assumed that

$$Y_i(t) = \mu_0(t) + X_i'\beta + g(Z_i) + \varepsilon_i,$$

where $\mu_0(t) = \sin(t)$ and ε_i follows normal distribution with mean 0 and variance 0.1. We took $\gamma = (1,1)'$ and $\beta = (1,1)'$, representing effects of the covariates on the observation scheme and the response variable. For W_i, we consider $W_i = 1$. For each setting, we considered $n = 100$ and 200. All the results reported here were based on $1{,}000$ Monte Carlo replications.

Tables 4.1 – 4.6 present the simulation results obtained on estimation of β. All tables include the estimated bias (BIAS) given by the average of proposed estimates of β minus the true value, the sample standard error (SSE) of the proposed estimates, the mean of the estimated standard error (ESE), and the empirical 95% coverage probabilities (CP). These results indicate that the proposed estimate seems to be unbiased and the proposed variance estimation procedure provides reasonable estimates. Also the results on the empirical coverage probabilities indicate that the normal approximation seems to be appropriate.

Table 4.1. Simulation results of β_1 and β_2 with a homogeneous Poisson process and linear $g(Z)$.

	$n = 100$		$n = 200$	
	β_1	β_2	β_1	β_2
$\rho = 0.5$				
Bias	0.0000	0.0143	−0.0010	−0.0003
SSE	0.1246	0.2544	0.0903	0.1664
ESE	0.1214	0.2235	0.0877	0.1621
CP	0.9440	0.9080	0.9480	0.9340
$\rho = 0$				
Bias	0.0002	0.0003	−0.0002	0.0005
SSE	0.0156	0.0290	0.0113	0.0193
ESE	0.0155	0.0271	0.0110	0.0193
CP	0.9490	0.9260	0.9380	0.9470
$\rho = -0.5$				
Bias	−0.0011	−0.0018	0.0007	0.0039
SSE	0.1218	0.2417	0.0885	0.1717
ESE	0.1206	0.2219	0.0883	0.1633
CP	0.9480	0.9170	0.9480	0.9380

Table 4.2. Simulation results of β_1 and β_2 with a homogeneous Poisson process and nonlinear $g(Z)$.

	$n = 100$		$n = 200$	
	β_1	β_2	β_1	β_2
$\rho = 0.5$				
Bias	-0.0011	-0.0083	-0.0032	-0.0057
SSE	0.1183	0.2181	0.0781	0.1411
ESE	0.1130	0.2093	0.0751	0.1402
CP	0.9390	0.9400	0.9300	0.9410
$\rho = 0$				
Bias	-0.0002	-0.0015	-0.0002	0.0011
SSE	0.0161	0.0292	0.0111	0.0198
ESE	0.0155	0.0269	0.0110	0.0192
CP	0.9350	0.9180	0.9510	0.9480
$\rho = -0.5$				
Bias	0.0006	0.0047	0.0031	0.0078
SSE	0.1102	0.2032	0.0779	0.1394
ESE	0.1050	0.1940	0.0745	0.1392
CP	0.9320	0.9300	0.9370	0.9440

Table 4.3. Simulation results of β_1 and β_2 with a non-homogeneous Poisson process and linear $g(Z)$.

	$n = 100$		$n = 200$	
	β_1	β_2	β_1	β_2
$\rho = 0.5$				
Bias	-0.0081	0.0051	-0.0043	-0.0067
SSE	0.1366	0.2441	0.0939	0.1807
ESE	0.1245	0.2275	0.0903	0.1666
CP	0.9230	0.9410	0.9390	0.9220
$\rho = 0$				
Bias	0.0003	0.0014	-0.0002	0.0007
SSE	0.0150	0.0264	0.0107	0.0189
ESE	0.0149	0.0254	0.0106	0.0181
CP	0.9490	0.9450	0.9490	0.9420
$\rho = -0.5$				
Bias	0.0021	0.0128	0.0017	0.0124
SSE	0.1335	0.2513	0.0938	0.1714
ESE	0.1252	0.2255	0.0916	0.1689
CP	0.9290	0.9190	0.9370	0.9490

Table 4.4. Simulation results of β_1 and β_2 with a non-homogeneous Poisson process and nonlinear $g(Z)$.

	$n = 100$		$n = 200$	
	β_1	β_2	β_1	β_2
$\rho = 0.5$				
Bias	0.0065	-0.0070	-0.0002	-0.0053
SSE	0.1207	0.2327	0.0776	0.1532
ESE	0.1174	0.2146	0.0776	0.1447
CP	0.9380	0.9210	0.9510	0.9400
$\rho = 0$				
Bias	0.0002	0.0005	-0.0003	-0.0004
SSE	0.0161	0.0268	0.0109	0.0183
ESE	0.0149	0.0253	0.0106	0.0181
CP	0.9240	0.9340	0.9370	0.9400
$\rho = -0.5$				
Bias	-0.0057	0.0075	0.0036	0.0040
SSE	0.1115	0.2141	0.0814	0.1572
ESE	0.1087	0.1986	0.0816	0.1514
CP	0.9340	0.9280	0.9520	0.9400

Table 4.5. Simulation results of β_1 and β_2 with a non-Poisson process and linear $g(Z)$.

	$n = 100$		$n = 200$	
	β_1	β_2	β_1	β_2
$\rho = 0.5$				
Bias	-0.0027	-0.0002	-0.0029	-0.0051
SSE	0.1304	0.2491	0.0946	0.1802
ESE	0.1252	0.2289	0.0914	0.1674
CP	0.9420	0.9230	0.9430	0.9370
$\rho = 0$				
Bias	0.0006	0.0010	0.0003	0.0007
SSE	0.0138	0.0246	0.0096	0.0176
ESE	0.0136	0.0240	0.0096	0.0170
CP	0.9420	0.9230	0.9490	0.9310
$\rho = -0.5$				
Bias	0.0023	0.0085	0.0050	0.0039
SSE	0.1279	0.2475	0.0923	0.1766
ESE	0.1257	0.2292	0.0910	0.1686
CP	0.9450	0.9240	0.9500	0.9340

Table 4.6. Simulation results of β_1 and β_2 with a non-Poisson process and nonlinear $g(Z)$.

	$n = 100$		$n = 200$	
	β_1	β_2	β_1	β_2
$\rho = 0.5$				
Bias	-0.0027	-0.0151	-0.0028	-0.0078
SSE	0.1080	0.1984	0.0848	0.1561
ESE	0.1039	0.1912	0.0829	0.1541
CP	0.9400	0.9320	0.9440	0.9380
$\rho = 0$				
Bias	0.0001	-0.0017	-0.0001	-0.0001
SSE	0.0139	0.0252	0.0100	0.0175
ESE	0.0136	0.0240	0.0096	0.0170
CP	0.9470	0.9320	0.9460	0.9470
$\rho = -0.5$				
Bias	0.0026	0.0105	-0.0023	-0.0018
SSE	0.0975	0.1807	0.0751	0.1424
ESE	0.0939	0.1724	0.0774	0.1447
CP	0.9300	0.9250	0.9530	0.9430

4.2 Robustness

To further investigate the robustness of the proposed estimate and also why one may need to use the proposed estimate instead of the estimates developed under restricted models such as that given in Zhao et al. (2012), we perform a simulation study to compare the estimates given in Zhao et al. (2012).

For this second study, we considered the situation where there exist one covariate, X_i's which is the Bernoulli distribution with success probability 0.5. The latent variable Z_i's were generated from the gamma distribution with shape parameter 5 and scale parameter 5. Also, we consider two cases of $g(Z_i)$. The first one is $g(Z_i) = \rho(Z_i - E[Z_i])/\sqrt{Var[Z_i]}$ which is a linear function, while the second is $g(Z_i) = \rho(\ln Z_i - E[\ln Z_i])/\sqrt{Var[\ln Z_i]}$ which is a nonlinear function. Here ρ characterizes the relationship between the observation process and the longitudinal response process. When $\rho > 0$, the two processes are positively correlated; when $\rho = 0$, the two processes have no correlation given the covariates; when $\rho < 0$, the two processes are negatively correlated. Here, three situations with $\rho = -0.5$, 0, and 0.5 were considered.

With respect to the observation process $O_i(t)$, three set-ups were considered as in the first study, For the response variable, it was assumed that

$$Y_i(t) = \mu_0(t) + X_i\beta + g(Z_i) + \varepsilon_i,$$

where $\mu_0(t) = \sin(t)$ or $\ln(1 + t)$ and ε_i follows Normal distribution with mean 0 and variance 0.1. We took $\gamma = 1$ and $\beta = 1$, representing effect of the covariate on the observation scheme and the response variable. We also $W_i = 1$ for simplicity. For each setting, we considered $n = 100$ and 200. All the resuls reported here were based on

1000 Monte Carlo replications. Here, we use ZTS to denote the estimate presented in Zhao et al. (2012). Table 4.7 gives the estimated bias (BIAS) which is the average of the estimates minus the true value, and the sample standard error (SSE) of the estimates, which is obtained for the estimate of β proposed here and given in Zhao et al. (2012) based on the simulated data. For $O_i(t)$ following Poisson processes, we also considered the relative efficiency (RE) which is the ratio of sample variance of ZTS to that of the proposed estimator. Comparison of the relative efficiency is ignored for non-Poisson process because ZTS's estimator is clearly biased and it is not meaningful to compare the efficiency.

Simulation studies suggest that when the observation process $O_i(t)$ does not follow a Poisson Process, the estimate proposed here seems to be unbiased, while the estimate given in Zhao et al. (2012) is clearly biased. On the other hand, ZTS's estimators perform better in terms of efficiency in general. It is reasonable because the proposed method has relaxed on the Poisson assumption about the observation process. This can be further confirmed in the simulation studies where $\rho = 0$. In this case, the efficient of the proposed method is better. Meanwhile, the BIAS of the proposed estimator are relatively stable compared to ZTS's. In other words, in general, the proposed estimation procedure seems to be more robust.

Table 4.7. Estimation results of β based on the proposed method and ZTS.

		$n = 100$			$n = 200$		
	$\mu_0(t)$	ZTS	Proposed	RE	ZTS	Proposed	RE
$\rho = 0.5$							
Homogeneous Poisson	$\sin(t)$	0.0029	-0.0020	1.1179	-0.0065	0.0002	1.0815
	$\log(1+t)$	0.0059	-0.0017	2.0788	-0.0035	0.0008	1.9560
Non-homogeneous Poisson	$\sin(t)$	0.0270	0.0020	1.0519	0.0299	0.0054	1.0978
	$\log(1+t)$	0.0301	0.0005	1.6661	0.0345	0.0050	1.7591
Non-Poisson	$\sin(t)$	-0.3469	0.0034	3.0331	-0.3465	0.0034	3.1248
	$\log(1+t)$	-0.3376	-0.0011	2.8982	-0.3491	0.0024	2.7552
$\rho = 0$							
Homogeneous Poisson	$\sin(t)$	-0.0003	0.0007	1.8728	0.0001	0.0015	1.5488
	$\log(1+t)$	-0.0026	0.0010	11.4730	0.0005	-0.0004	10.9011
Non-homogeneous Poisson	$\sin(t)$	0.0333	0.0001	1.5154	0.0322	0.0005	1.5624
	$\log(1+t)$	0.0335	-0.0005	5.1719	0.0308	-0.0007	5.2976
Non-Poisson	$\sin(t)$	-0.3508	0.0005	8.5140	-0.3427	-0.0006	9.2894
	$\log(1+t)$	-0.3501	-0.0003	4.1698	-0.3408	-0.0023	3.8843
$\rho = -0.5$							
Homogeneous Poisson	$\sin(t)$	0.0026	-0.0016	1.0547	-0.0051	-0.0022	1.1340
	$\log(1+t)$	0.0009	-0.0053	1.6433	-0.0037	-0.0013	1.8634
Non-homogeneous Poisson	$\sin(t)$	0.0254	-0.0026	1.1616	0.0275	0.0067	1.1144
	$\log(1+t)$	0.0282	-0.0048	1.6780	0.0292	0.0063	1.4315
Non-Poisson	$\sin(t)$	-0.3551	-0.0008	2.6833	-0.3390	-0.0007	2.6811
	$\log(1+t)$	-0.3349	-0.0057	2.4942	-0.3455	0.0022	2.4725

4.3 Finite Sample Properties of Diagnostic Tests

In the third study, we examined the adequacy of the large-sample approximation to the null distribution of the proposed test statistics of the diagnostic tests for practical sample sizes. For both diagnostic tests, we considered the situation where there exists one covariate, X_{i1}'s which follows the uniform distribution over $\{1, 2, 3, 4, 5\}$. The latent variable Z_i's were generated from the gamma distribution with shape parameter 5 and scale parameter 5 (equivalently, mean 25 and variance 125). Similar to the first and the second study, We consider two cases of $g(Z_i)$. In the first case, we took

$$g(Z_i) = \rho\,(Z_i - 25)/\surd\overline{125}.$$

In the second case, we took

$$g(Z_i) = \rho\,\big(\log Z_i - E(\log Z_i)\big)/\sqrt{var(\log Z_i)}.$$

The follow-up time C_i is generated from the uniform distribution over $[\tau/2, \tau]$ with $\tau = 18$. With respect to the observation process $O_i(t)$, three set-ups were considered as in the first simulation study. For the diagnostic test of the functional form of the covariate, the response variable was assumed to be

$$Y_i(t) = \mu_0(t) + (X_i^{\nu})'\beta + g(Z_i) + \varepsilon_i$$

and for the diagnostic test of the goodness-of-test of models (2.1) and (2.2), the response variable was assumed to be

$$Y_i(t) = \mu_0(t) + (X_i'\beta)^{\nu} + g(Z_i) + \varepsilon_i$$

where $\mu_0(t) = \sin(t)$ and $\varepsilon_i \sim \mathcal{N}(0, 0.25)$. We took $W_i = 1$, representing effect of the covariate on the response variable. For each setting, we considered $n = 100$ and 200, $\rho = 0.5$ and $\nu = 1, 1{,}2, 1.4, \ldots, 2.8$. All the results reported here were based on 1000 Monte Carlo replications. Table 4.8 and 4.9 give the empirical sizes and powers of our tests. The empirical sizes of the tests are close to the nominal ones, 5%. This suggests that the null distributions of the proposed test statistics are well approximated.

Table 4.8: Empirical Sizes and Powers of Diagnostic Test for the Functional Form of a Covariate.

	$n = 100$						$n = 200$					
	Linear $g(Z_i)$			Nonlinear $g(Z_i)$			Linear $g(Z_i)$			Nonlinear $g(Z_i)$		
v	I	II	III	I	II	III	I	II	III	I	II	III
1.0	0.056	0.050	0.047	0.055	0.052	0.047	0.059	0.053	0.044	0.055	0.055	0.052
1.2	0.138	0.132	0.079	0.121	0.137	0.108	0.197	0.181	0.191	0.248	0.237	0.190
1.4	0.299	0.270	0.257	0.269	0.308	0.255	0.511	0.527	0.529	0.660	0.652	0.569
1.6	0.388	0.448	0.338	0.349	0.420	0.334	0.818	0.814	0.764	0.879	0.866	0.764
1.8	0.452	0.525	0.426	0.412	0.512	0.428	0.945	0.955	0.890	0.973	0.967	0.883
2.0	0.542	0.613	0.514	0.508	0.610	0.514	0.991	0.996	0.967	0.995	0.995	0.957
2.2	0.661	0.698	0.620	0.618	0.695	0.611	0.999	1.000	0.994	0.999	1.000	0.984
2.4	0.722	0.801	0.682	0.718	0.810	0.700	0.999	0.999	0.996	1.000	1.000	0.997
2.6	0.812	0.880	0.798	0.815	0.873	0.782	1.000	1.000	1.000	1.000	1.000	1.000
2.8	0.884	0.917	0.835	0.875	0.918	0.844	1.000	1.000	1.000	1.000	1.000	1.000

NOTE: I, II, and III refer to observation processes according to setup 1, 2, and 3 specified in the first simulation study respectively.

Table 4.9: Empirical Sizes and Powers of Diagnostic Test for the Goodness-of-fit.

	$n = 100$						$n = 200$					
	Linear $g(Z_i)$			Nonlinear $g(Z_i)$			Linear $g(Z_i)$			Nonlinear $g(Z_i)$		
v	I	II	III	I	II	III	I	II	III	I	II	III
1.0	0.051	0.054	0.044	0.053	0.052	0.043	0.050	0.052	0.048	0.054	0.053	0.050
1.2	0.145	0.142	0.079	0.136	0.126	0.095	0.179	0.187	0.175	0.227	0.209	0.155
1.4	0.282	0.317	0.244	0.298	0.303	0.266	0.514	0.512	0.557	0.657	0.650	0.540
1.6	0.365	0.392	0.355	0.359	0.400	0.358	0.793	0.808	0.752	0.854	0.871	0.753
1.8	0.430	0.513	0.443	0.415	0.501	0.425	0.947	0.955	0.894	0.965	0.976	0.885
2.0	0.514	0.606	0.496	0.501	0.609	0.505	0.993	0.995	0.965	0.992	0.994	0.961
2.2	0.648	0.695	0.627	0.603	0.698	0.591	1.000	1.000	0.980	0.998	0.999	0.992
2.4	0.755	0.796	0.722	0.713	0.784	0.703	0.999	1.000	0.995	1.000	1.000	0.997
2.6	0.795	0.887	0.776	0.801	0.855	0.790	1.000	1.000	0.999	1.000	1.000	0.999
2.8	0.863	0.917	0.832	0.868	0.905	0.851	1.000	1.000	1.000	1.000	1.000	1.000

NOTE: I, II, and III refer to observation processes according to setup 1, 2, and 3 specified in the first simulation study respectively.

Chapter 5

Application

To illustrate the proposed methodology, we consider a bladder cancer study conducted by the Veterans Administration Cooperative Urological Research Group (Andrews and Herzberg (1985); Byar (1980); Sun and Wei (2000); Wellner and Zhang (2000); Zhang (2002)). In the study, the patients with superficial bladder tumors were randomly assigned to one of three treatment groups – placebo, thiotepa, and pyridoxine. During the study, many patients had multiple recurrences of the bladder tumors and all recurrences between visits were recorded and removed at clinical visits; the number of visits and visit time points varied greatly from patient to patient. At the beginning of the study, for each patient, two important baseline covariates were reported; the number of initial tumors and the size of the largest initial tumor. Following Sun and Wei (2000), we restrict our attention to the patients in the placebo (47) and the thiotepa (38) groups.

5.1. Estimation

For the analysis, we took $Y_i(t)$ to be the logarithm of the number of observed tumors at time t, plus 1 to avoid 0, $i = 1,\ldots,118$. We set the first component of X_i to 1 if the ith patient was given the pyridoxin treatment and 0 otherwise, the second component of X_i to 1 if the ith patient was given the thiotepa treatment and 0 otherwise and the third and the forth components of X_i to the number of initial tumors and the size of the largest initial tumor of the ith patient, respectively, $i = 1,\ldots,118$. The longitudinal process of the bladder tumors $Y_i(t)$ and the clinical visit process were described by models (2.1) and (2.2).

The proposed application of the estimation procedure with $W_i = 1$ gave $\hat{\beta} = (-0.0085, -0.3232, 0.0729, -0.0102)'$ with estimated standard errors (0.1437, 0.0928, 0.0269, 0.0267), and thus p-values (0.9530, 0.0005, 0.0069, 0.7016), respectively. These results indicate that the thiotepa treatment significantly reduced the occurrence rate of the bladder tumors and the number of initial tumors has a significant positive effect on the tumor recurrence rate. However, the pyridoxin treatment and the size of the largest initial tumor did not have significant effect on the occurrence rate of the bladder tumors. Sun et al. (2007) applied their method to analyze the same data and obtained that the thiotepa treatment had a significant effect in reducing the recurrence of bladder tumors, but they did not detect the effect of the initial number of bladder tumors on the recurrence rate of the bladder tumor. The reason for this difference between the two application results may be due to the misspecification of the relationship between the longitudinal response process and the observation process in Sun et al. (2007).

5.2 Model Diagnostics

Consider the application of the model-checking procedures given in Chapter 3 to the data. Treating the four covariates separately, we found $\sup_x |\mathcal{F}_1(x)| = 2.650$ with the p-value of 0.441, $\sup_x |\mathcal{F}_2(x)| = 3.210$ with the p-value of 0.121, $\sup_x |\mathcal{F}_3(x)| = 2.451$ with the p-value of 0.725, $\sup_x |\mathcal{F}_4(x)| = 3.039$ with the p-value of 0.349. All four p-values suggest that the linear form of the covariates is approriate. To illustrate the graphical procedure for checking functional form of covariate, the observed process $\mathcal{F}_3(x)$ along with 20 realizations of the process $\hat{\mathcal{F}}_3(x)$

and the observed process $\mathcal{F}_4(x)$along with 20 realizations of the process $\hat{\mathcal{F}}_4(x)$ are depicted in Figures 5.1 and 5.2, respectively. Both graphical and numerical procedures suggest the appropriateness of the linear form of the covariates.

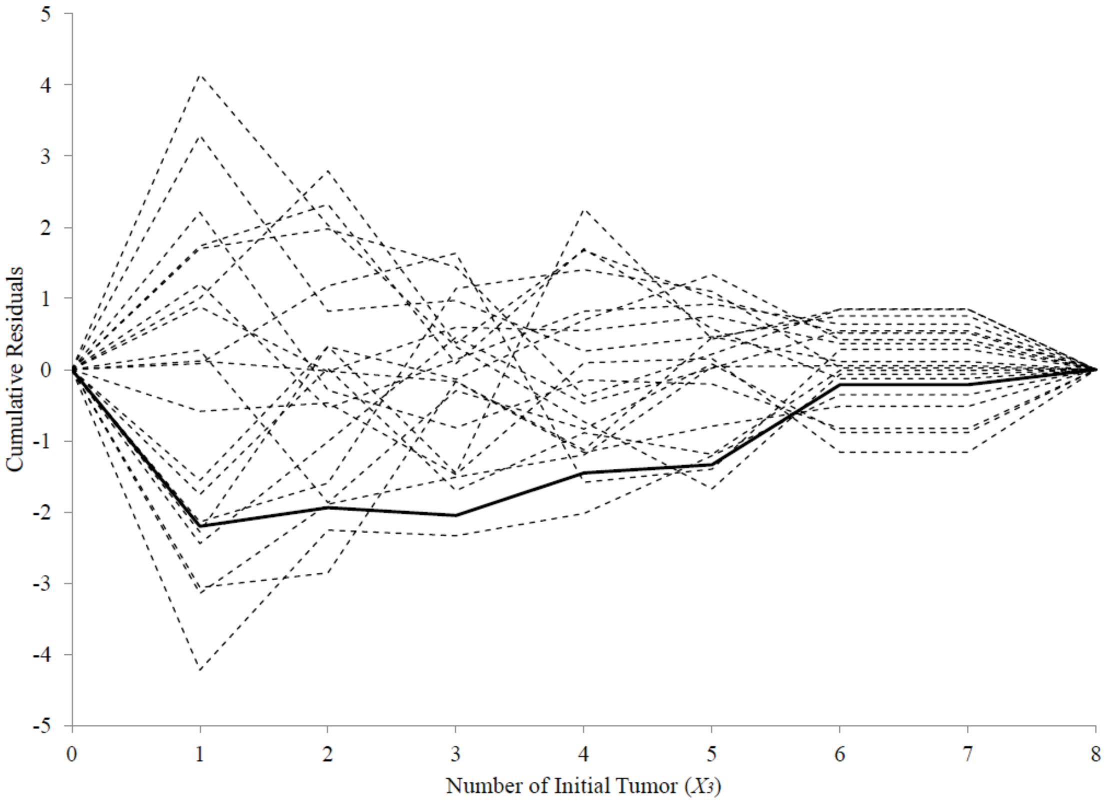

Figure 5.1 The observed process $\mathcal{F}_3(x)$ along with 20 realizations of the process $\hat{\mathcal{F}}_3(x)$

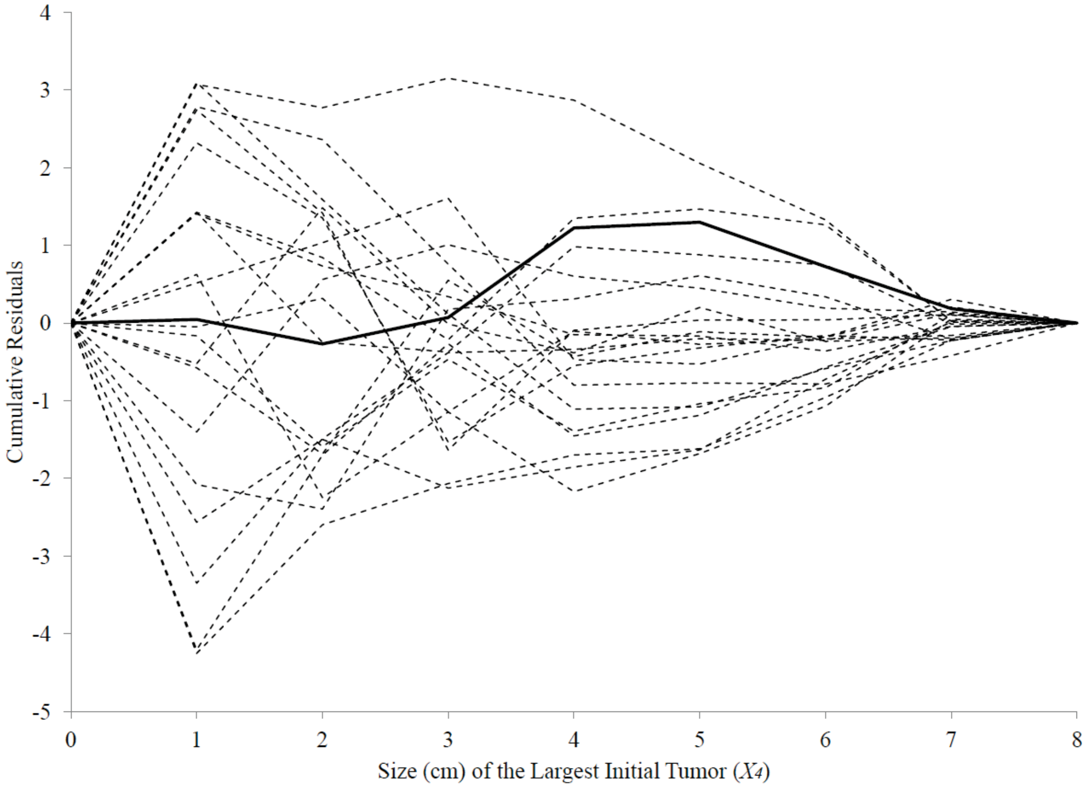

Figure 5.2 The observed process $\mathcal{F}_4(x)$along with 20 realizations of the process $\widehat{\mathcal{F}}_4(x)$

To assess the overall fit of models (2.1) and (2.2), we obtain $\sup_{t,x} |\phi(t,x)| = 3.133$ with the p-value of 0.486. This suggests that these models seem to be appropriate for the bladder cancer data considered here.

Chapter 6

Conclusions

This dissertation investigates regression analysis of longitudinal data when the observation times or process may be related to the underlying recurrent event process of interest. For the problem, some general and robust models were presented and an estimating equation-based estimation procedure was developed. The proposed estimate is consistent and asymptotically normally distributed. Simulation studies indicate that the estimation procedure works well for practical situations. Also a goodness-of-fit procedure was given for the proposed models.

One main advantage of the proposed inference procedure is that it allows the correlation between the recurrent event process of interest and the observation process in a general format. This is very important since the format of the relationship between the two processes is generally unknown in practice and could be very complicated and thus a flexible model may be more preferred. Also the proposed approach does not require the Poisson assumption, which plays some main roles in existing procedures but can be questionable in many situations. Compared to the estimation procedure presented by Zhao et al. (2012), our method is more robust.

In the estimation equation approach, the weight is an important element that improves the efficiency of estimation.

References

Andrews, D. F. and Herzberg, A. M. (1985). *Data: A Collection of Problems from Many Fields for the Student and Research Worker*. Springer-Verlag, New York

Byar, D. P. (1980). The Veterans Administration study of chemoprophylaxis for recurrent stage I bladder tumors: comparisons of placebo, pyridoxine, and topical thiotepa. *In Bladder tumors and other topics in urological oncology* (Edited by M. Pavane-Macaluso, P. H. Smith and F. Edsmyr), 363-370. Plenum, New York.

Diggle, P. J., Liang, K. Y., and Zeger, S. L. (1994). *The Analysis of Longitudinal Data*. Oxford University Press, Oxford.

Liang Y., Lu, W., and Ying, Z. (2009). Joint modeling and analysis of longitudinal data with informative observation times. *Biometrics* 65, 377-384.

Lin, D. Y., Wei, L. J., Yang, I. and Ying, Z. (2000). Semiparametric regression for the mean and rate function of recurrent events. *Journal of the Royal Statistical Society B* 69, 711-730.

Lin, D. Y. and Ying, Z. (2001). Semiparametric and nonparametric regression analysis of longitudinal data. *J. Amer. Statist. Assoc*. 96, 103-126.

Lipsitz, S. R., Fitzmaurice, G. M., Ibrahim, J. G., Gelber, R., and Lipshultz, S. (2002). Parameter estimation in longitudinal studies with outcome-dependent follow-up. *Biometrics* 58, 621-630.

Liu, L., Huang, X. and O'Quigley, J. (2008). Analysis of longitudinal data in the presence of information observational times and a dependent terminal event, with application to medical cost data. *Biometrics* 64, 950-958.

Sun, J., Park, D., Sun, L. and Zhao, X. (2005). Semiparametric regression analysis of longitudinal data with informative observation times. *Journal of the American Statistical Association* 100, 882-889.

Sun, J., Sun, L. and Liu, D. (2007). Regression analysis of longitudinal data in the presence of informative observation and censoring times. *J. Amer. Statist. Assoc.* 102, 1397-1406.

Sun, J. and Wei, L. J. (2000). Regression analysis of panel count data with covariate-dependent observation and censoring times. *J. R. Statist. Soc.* B 62, 293-302.

Wellner, J. A. and Zhang, Y. (2000). Two estimators of the mean of a counting process with panel count data. *Ann. Statist.* 28, 779-814.

Wang, M. C., Qin, J., and Chiang, C. T. (2001). Analyzing recurrent event data with informative censoring. *J. Amer. Statist. Assoc.* 96, 1057-1065.

Welsh, A. H., Lin, X., and Carroll, R. J. (2002). Marginal longitudinal nonparametric regression: locality and efficiency of spline and kernel Methods. *J. Amer. Statist. Assoc.* 97, 482-493.

Zhang, Y. (2002). A semiparametric pseudolikelihood estimation method for panel count data. *Biometrika* 89, 39-48.

Zhao, X., Tong, X., and Sun, L. (2012). Joint analysis of longitudinal data with dependent observation times. *Statistica Sinica* 22, 317-336