SEMIPARAMETRIC AND NONPARAMETRIC METHODS FOR THE ANALYSIS OF PANEL COUNT DATA

A Dissertation

presented to

the Faculty of the Graduate School

the University of Missouri

In Partial Fulfillment

Of the Requirements for the Degree

Doctor of Philosophy

---

by

YANG LI

Dr. (Tony) Jianguo Sun, Dissertation Supervisor

MAY 2013

The undersigned, appointed by the Dean of the Graduate School, have examined the dissertation entitled:

SEMIPARAMETRIC AND NONPARAMETRIC METHODS FOR
THE ANALYSIS OF PANEL COUNT DATA

presented by Yang Li,

a candidate for the degree of Doctor of Philosophy and hereby certify that, in their opinion, it is worthy of acceptance.

Dr. (Tony) Jianguo Sun

Dr. Nancy Flournoy

Dr. Jing Qiu

Dr. Subharup Guha

Dr. Aigen Li

Dedicated to my parents and my family:

Yuling Wei and Xiaoguang Li,

YuChoong Soo and Jesse Li Soo

ACKNOWLEDGMENTS

I would like to acknowledge many people for helping me during my doctoral work. First and foremost I would like to express my deepest gratitude to my esteemed advisor Dr. (Tony) Jianguo Sun for his insightful inspiration, critical encouragement and generous support, and for guiding me to make a difference beyond myself. My appreciation to him is beyond all words and speeches.

I extend my gratitude to my advisory committee members, Drs. Nancy Flournoy, Jing Qiu, Subharup Guha and Aigen Li, for their academic support and invaluable advice on my work. I owe a special note of gratitude to Dr. Hui Zhao for her generous help and great collaboration through my research.

I also owe a debt of gratitude to all faculty in the Department of Statistics who have helped and encouraged me in various ways during my course of studies. I would like to express my appreciation to Drs. Min Yang, Paul Speckman, Dongchu Sun, Chong He, Fei Liu, Christie Spinka, Chris Wikle and Sakis Micheas for all that they have taught me. I am very grateful to Dr. Larry Ries for helping and supporting me with my teaching assignments.

I would like to take the opportunity to extend my thanks to our department for offering me an excellent environment to study here. I especially appreciate Mses. Kathleen Maurer, Tracy Pickens and Judy Dooley for their plenty of help. I also appreciate all my classmates and friends here, especially Ni Li, Yajun Liu, Na Hu, Tianhua Wang and Haiying Wang for their help and friendship.

Finally, I am particularly grateful to my supportive parents whom I owe everything, and YuChoong Soo for accompanying me all the time through and forward.

TABLE OF CONTENTS

SEMIPARAMETRIC AND NONPARAMETRIC METHODS FOR THE ANALYSIS OF PANEL COUNT DATA

YANG LI

Dr. (Tony) Jianguo Sun, Dissertation Supervisor

ABSTRACT

Panel count data are one type of event-history data concerning recurrent events. Ideally for an event-history study, subjects should be monitored continuously, so for the events that may happen recurrently over time, the exact time of each event occurrence is recordable. Data obtained in such cases are commonly referred to as recurrent event data (Cook and Lawless, 2007). In reality, however, subjects may only be observed at their clinical visits or discrete times. As a result, instead of observing the exact event times, one only knows the numbers of events that happen between the observation times. Such interval-censored recurrent event data are usually referred to as panel count data (Kalbfleisch and Lawless, 1985; Sun and Kalbfleisch, 1995; Thall and Lachin, 1988).

The primary interest with panel count data is about the underlying recurrent event process. Meanwhile for the analysis, one needs to consider the times when the observations occur, which can be regarded as realizations of an observation process with follow-up times. This dissertation consists of four parts. In the first part, we will con-

sider regression analysis of panel count data with dependent observation processes while the follow-up times may be subject to a terminal event like death. A semiparametric transformation model is presented for the mean function of the underlying recurrent event process among survivals. To estimate the regression parameters, an estimating equation approach is proposed and the inverse survival probability weighting technique is used. In addition, the asymptotic distribution of the proposed estimate is derived and a model checking procedure is presented. Simulation studies are conducted to evaluate finite sample properties of the proposed approach, and the approach is applied to a bladder cancer study.

The second part will focus on regression analysis of multivariate panel count data in the presence of a terminal event. Both the observation process and the terminal event may be correlated with recurrent event processes of interest. We present a class of semiparametric additive models for the mean functions of the underlying recurrent event processes. For the estimation of the regression parameters, an estimating equation based inference procedure is developed. The asymptotic properties of the proposed estimators are established and a model-checking procedure is derived for practical situations.

The third part will discuss nonparametric comparison based on panel count data. Most approaches that have been developed in the literature require an equal observation process for all subjects. However, such an assumption may not hold in reality. A new class of test procedures are proposed that allow unequal observation processes for the subjects from different treatment groups, and both univariate and multivariate panel count data are considered. The asymptotic normality of the proposed test statistics is established and a simulation study is conducted. The approach is applied to a skin

cancer study. Finally, the last part will discuss some directions for future research.

Chapter 1

INTRODUCTION

1.1 Introduction and Examples

1.1.1 Introduction

Panel count data are one type of event-history data or longitudinal data concerning some recurrent events. In panel count data, the observations consist of discrete time points with no information available about the timing of events between observation times (Kalbfleisch and Lawless, 1985). Compared with event-history data with continuous observation paths, which are commonly referred to as recurrent event data, panel count data are interval-censored and can only provide the numbers of events occurring between observation times. In addition, the observation times are usually different from subject to subject.

There are two counting processes associated with panel count data: the observation process and the recurrent event process. The response variable from the recurrent event

process has observations only when the observation process has jumps. As a result, the analysis of panel count data rely on both of these counting processes and their relationships defined under various scenarios.

In many cases, potential observation times are predetermined. If study subjects can follow their schedules throughout the study, the observation processes are independent from the response variable since the preassigned observation times do not carry on or affect anything of the recurrent events that may occur later on. Moreover, the observation processes can also be subject-independent if they all follow the same distribution. In cases when the observation times are not predetermined, one may still have independent observation processes if they are noninformative about either the subjects or the response variable over time.

When the observation processes appear informative, one may suspect they are either subject or response variable dependent (Sun et al., 2005). For example, consider treatment comparisons in clinical trials, some treatments may require the subjects being examined more often than those with other treatments, so that the observation rate of someone may depend on which treatment group one is from. Also, severe disease development may also cause more or fewer clinical visits, so that the recurrent event process and the observation process can be correlated. The analysis for such cases must take into account the information implied by the observation process.

In practice, the observation process can be stopped by death, drop-out, or the end of the study. Depending on whether or not a stopping event also terminates the underlying recurrent event process, there are two scenarios. One is censoring, which only stops the observation process but the recurrent event may still continue after it has occurred. The other one is a terminal event, which terminates both the observation and the

recurrent event processes. Both censoring and a terminal event can be independent or not with the response variable. Inference methods need to be tuned to different practical situations.

1.1.2 Examples

1.1.2.1 The National Cooperative Gallstone Study

The National Cooperative Gallstone Study (NCGS) is a double-blinded, placebo-controlled clinical trial to study the effect of Chenodiol (chenodeoxycholic acid) in dissolving cholesterol gallstones among 916 patients who chose nonsurgical treatments (Schoenfield et al., 1981). Petients were followed for up to two years with each of the three treatments randomly assigned: high dose (750 mg per day), low dose (375 mg per day), or placebo. The primary objective was to assess the treatments effectiveness on reducing the incidence of digestive symptoms associated with gallstone disease. For this, patients were scheduled to return for clinical visits at 1, 2, 3, 6, 9 and 12 months, and the incidences of digestive symptoms were reported. However, the actual visit times varied. Thall and Lachin (1988) analyzed one of the symptoms, nausea, during the first year of follow-up on a subset of 113 NCGS patients in the high-dose and placebo groups. They treated the observation times as fixed at the scheduled times, with randomly missed observations in between. In conclusion, they demonstrated a significant difference between high-dose and placebo, especially during the first six months of follow-up.

1.1.2.2 The Bladder Cancer Study

The Bladder Cancer Study is a well-known example giving rise to panel count data. It was conducted by the Veterans Administration Cooperative Urological Research Group (Sun and Wei, 2000; Ghosh and Lin, 2002; Wellner and Zhang, 2007). In the study, 116 patients with stage I bladder cancer were randomly assigned to placebo, pyridoxine or intravesical thiotepa and followed for recurrences of superficial bladder tumors. All tumors were removed transurethrally at the beginning of the study. At each patient's clinical visit, the bladder tumors that occurred since the last visit were removed after the number was recorded. During the study, each patient visited the clinics periodically, and the actual visit times vary among the subjects. Besides treatment groups, the data also include some other information of the patients on the initial numbers of tumors, sizes of the largest initial tumors and the death times for those who died during the study. The main purpose was to study the treatment effects on reducing the rate of tumor occurrences. Among others, Sun and Wei (2000) demonstrated that the patients in the thiotepa group tended to visit the clinics more often than the patients in the placebo group, and also that thiotepa reduced the recurrences of tumors significantly compared with placebo. With respect to the covariates, they suggested that the number of initial tumors was a significant prognostic factor related to tumor recurrences, but the initial size was not. Zhang (2002) also got the same conclusion using a robust semiparametric pseudolikelihood estimation method, in which the Poison assumption on the recurrent event process could be relaxed. In addition, Huang et al. (2006), Sun et al. (2007), He et al. (2009) and Zhao and Tong (2011) considered that the observation processes may be informative about the occurrences of tumors and all of their work demonstrated a significance effect on thiotepa. With

respect to other covariate effects, however, He et al. (2009) concluded that neither the number nor largest size of initial tumors were significant.

1.1.2.3 The Skin Cancer Study

The skin cancer chemoprevention trial is a double-blinded, placebo-controlled, 5-year randomized Phase III clinical trial conducted by the University of Wisconsin Comprehensive Cancer Center in Madison, Wisconsin (Li et al., 2011). In the study, 291 patients were randomly assigned to the placebo or difluoromethylornithine (DFMO) group, and the objective was mainly on evaluating the overall effectiveness of 0.5g/m2/day PO DFMO in reducing the recurrences of both basal cell carcinoma and squamous cell carcinoma. Subjects were scheduled to be assessed every six months, but the actual observation times varied. Covariates were recorded including treatment type, the number of prior skin cancer reported up to randomization, gender and age at enrollment. Li et al. (2011) analyzed the data and found that the number of prior skin cancers seemed to be positively related to the recurrences of both basal cell carcinoma and squamous cell carcinoma, but the DFMO treatment or other covariates mentioned above did not show significant effects.

For the examples given above, the first two are univariate panel count data, and the last one gives multivariate panel count data. The remainder of this chapter is organized as follows. Section 1.2 introduces semiparametric and nonparametric estimation methods on the mean function of panel count data. Section 1.3 discusses nonparametric comparison procedures with panel count data. The outline of the dissertation is given in Section 1.4.

1.2 Semiparametric and Nonparametric Estimation for Panel Count Data

Consider a longitudinal study concerning some recurrent events. Let $N(t)$ and $O(t)$ denote the underlying recurrent event process and the observation process, respectively, representing the cumulative number of event occurrences and observation times up to time t. Also let C be a censoring or follow-up time and $Z(t)$ be a vector of external covariate process (Kalbfleisch and Prentice, 2002). For the observation process, let K denote the total number of observations, and $\{T_1, \cdots, T_K\}$ be the time points at which $O(t)$ jumps, then $N(t)$ is observed only at these T_j's. Suppose that the study consists of n independent subjects. Then the observed data have the form

$$\{\, O_i(t), Z_i(t), N_i(T_{i,1}), \ldots, N_i(T_{i,K_i});\ 0 \le t, T_{i,K_i} \le C_i,\ i = 1, \ldots, n \,\}.$$

For the recurrent event process, we will use $\Lambda(t)$ to denote the mean function of $N_i(t)$'s, *i.e.*, $\Lambda(t) = E\{N_i(t)\}$, $i = 1, \ldots, n$ for the rest of this section.

1.2.1 Nonparametric Estimation of the Mean Function

Let $s_1 < \cdots < s_m$ denote the ordered different time points of all observation times $\{T_{i,j}\}$. First consider a simple case, where $T_{i,j} = s_j$ and $K_i = m$ for all $i = 1, \ldots, n$. Then Nelson-Aalen estimator can be used for estimating $\Lambda(s_l)$ in form of

$$\hat{\Lambda}(s_l) = \sum_{j=1}^{l} \frac{\sum_{i=1}^{n} I(s_j \le T_{i,K_i})(N_i(s_j) - N_i(s_{j-1}))}{\sum_{i=1}^{n} I(s_j \le T_{i,K_i})}.$$

In general for panel count data, since the observation times differ among subjects,

the Nelson-Aalen estimator cannot be used to estimate $\Lambda(t)$. Thall and Lachin (1988) used data grouping method assuming the rate function $d\Lambda(t)$ being constant between common observation times for all subjects, then $d\Lambda(t)$ can be estimated by

$$d\hat{\Lambda}(t) = \frac{1}{\sum_{i=1}^{n} I(t \leq T_{i,K_i})} \sum_{i=1}^{n} \sum_{j=1}^{K_i} \frac{N_i(t_{i,j}) - N_i(t_{i,j-1})}{t_{i,j} - t_{i,j-1}} I(t_{i,j-1} < t < t_{i,j}).$$

And $\Lambda(t)$ can be estimated by integrating $d\hat{\Lambda}(t)$

$$\hat{\Lambda}(t) = \int_0^t d\hat{\Lambda}(s).$$

However, it is obvious that the assumption of the above method cannot always hold. Instead of estimate $d\Lambda(t)$, a more common practice is to estimate $\Lambda(t)$ directly. A well-known estimator of $\Lambda(t)$ is given by the isotonic regression estimator (IRE) (Sun and Kalbfleisch, 1995; Wellner and Zhang, 2000). Following the notation above, let w_l and $\bar{N}_l$ represent the number and mean value of observations made at s_l, $l = 1, \ldots, m$. The isotonic regression estimator $(\hat{\Lambda}(s_1), \ldots \hat{\Lambda}(s_m))$ is then defined as $(\Lambda(s_1), \ldots, \Lambda(s_m))$ that minimizes the weighted sum of squares

$$\sum_{l=1}^{m} w_l(\bar{N}_l - \Lambda(s_l))^2$$

subject to the order restriction $\Lambda(s_1) \leq \cdots \leq \Lambda(s_m)$. Following the original formula for isotonic regression (Barlow et al., 1972; Robertson et al., 1988), the isotonic estimator for $\Lambda(s_l)$ is given by

$$\hat{\Lambda}(s_l) = max_{r \leq l} min_{u \geq l} \frac{\sum_{v=r}^{u} w_v \bar{n}_v}{\sum_{v=r}^{u} w_v} = min_{u \geq l} max_{r \leq l} \frac{\sum_{v=r}^{u} w_v \bar{n}_v}{\sum_{v=r}^{u} w_v}, \ l = 1, \ldots, m. \quad (1.1)$$

Wellner and Zhang (2000) showed that IRE in (1.1) is the same as the nonparametric maximum pseudo-likelihood estimator (NPMPLE). Assuming that the counting process of $N(t)$ is a non-homogeneous Poisson process and ignoring the dependence of events of the same subject, the pseudo log likelihood function can be written as:

$$l_n^{ps}(\Lambda) = \sum_{i=1}^{n}\sum_{j=1}^{K_i}\{N_i(T_{i,j})log(\Lambda(T_{i,j})) - \Lambda(T_{i,j-1})\}. \tag{1.2}$$

Under the non-homogeneous Poisson assumption, Wellner and Zhang (2000) also proposed a nonparametric maximum likelihood estimator (NPMLE) maximizing the full log-likelihood function of Λ proportional to

$$l_n(\Lambda) = \sum_{i=1}^{n}\sum_{j=1}^{K_i}\{N_i(T_{i,j}) - N_i(T_{i,j-1})\}log\{\Lambda(T_{i,j}) - \Lambda(T_{i,j-1})\} - \sum_{i=1}^{n}\Lambda(T_{i,K_i}). \tag{1.3}$$

Wellner and Zhang (2000) studied the asymptotic properties of both estimators and gave a modified iterative convex minorant (MICM) algorithm for NPMLE. It was demonstrated that NPMLE could be more efficient than IRE or NPMPLE, but NPMLE is computationally more demanding. Hu et al. (2009a) proposed an alternative algorithm which is simpler and faster.

Hu et al. (2009b) suggested a new class of estimates which can be considered as generalizations of IRE, by minimizing the generalized least-squares function involving a known $K_i \times K_i$ symmetric weight matrix $(W(T_{i,j}, T_{i,l}))$:

$$\sum_{i=1}^{n}\sum_{j=1}^{K_i}\sum_{l=1}^{K_i} W(T_{i,j}, T_{i,l})\{N_i(T_{i,j}) - \Lambda(T_{i,j})\}\{N_i(T_{i,l}) - \Lambda(T_{i,l})\} \tag{1.4}$$

subject to the non-decreasing property of $\Lambda(t)$. Compared with NPMLE, Hu et al.

(2009b) showed that the estimator defined above could have close efficiency as NPMLE for Poisson processes and be more efficient for non-Poisson processes.

Some other methods for the mean function estimation problem are given by Zhang and Jamshidian (2003) and Lu et al. (2007). The former modeled the dependence of $\{N_i(T_{i,j}), j = 1, \ldots, K_i\}$ by employing a latent variable, and an EM-algorithm was developed when the latent variable followed a gamma distribution. The latter studied both pseudo-likelihood and likelihood based approaches when the mean function of $\Lambda(t)$ can be approximated by the monotone cubic I-splines.

1.2.2 Semiparametric Regression Analysis

1.2.2.1 Observation times Independent of the Recurrent Responses

For noninformative observation processes, Cheng and Wei (2000) considered a semiparametric model, relating the mean of $N(t)$ proportionally to a function of a time-dependent covariate vector $Z(t)$, given by

$$E\{N_i(t)|Z_i(t)\} = \mu(t)\exp\{\beta' Z_i(t)\},$$

where $\mu(t)$ is an unknown baseline mean function, and the observation process is assumed to be independent with the event process subject to independent censoring. An estimating equation method was employed for the inference and the proposed estimate was shown to be asymptotically normal.

As shown by the Bladder Cancer Study in Section 1.1.2.2, sometimes the observation times may be covariate dependent. Sun and Wei (2000) proposed the following model:

conditioning on the covariate, $\Lambda_i(t) = E\{N_i(t)|Z_i\}$ is of the form

$$\Lambda_i(t) = \Lambda_0(t)\exp(\beta Z_i).$$

Three cases were studied for the observation process and follow-up times, including both of those being covariate independent, and either one of them or both being covariate dependent. The analysis was based on estimating equation methods.

Sometimes it is plausible to assume that the recurrent event process, the observation process and the censoring time are independent given the covariate Z_i. For this, Hu, et al. (2003) proposed the model

$$E\{N_i(t)|Z_i = z_i\} = \Lambda_0(t)\exp(\beta' z_i).$$

Two estimating equation based methods were constructed by conditioning on or modeling the observation process.

Instead of univariate panel count data, one may observe multivariate panel count data. Suppose there are m types of recurrent events of interest and individuals are only observed intermittently. For the analysis, one may use the above models for each of the kth-type event and its observation process, *i.e.*

$$E\{N_{ik}|z_i\} = \mu_k(t)g_N(z_i'\beta_0),$$

and

$$E\{O_{ik}|z_i\} = \nu_k(t)g_O(z_i'\gamma_0), \quad k = 1, \ldots, m;\ i = 1, \ldots, n,$$

where $\mu_k(t)$ and $\nu_k(t)$ are unknown baseline mean functions at time t, and g_N, g_O are

positive functions that are strictly increasing and twice differentiable. He et al. (2008) presented such class of marginal transformation models and developed estimating equation based regression analysis along with the asymptotic properties of the proposed estimators.

Assuming that observation times are independent of the response variable, some other semiparametric regression methods proposed in the literature include Zhang (2002) and Kim, Y. J. (2007). The former considered a proportional rate model on the event process given the covariate Z_i, and proposed a semiparametric pseudolikelihood estimation method that is robust in sense that the estimator converges to its true value whether or not $N(t)$ is a Poisson process given Z_i. The latter dealt with situations when measurement errors may occur with covariates, and the estimation method in Zhang (2002) was combined with a partial likelihood method using auxiliary covariates (Zhou and Pepe, 1995; Zhou and Wang, 2000).

1.2.2.2 Observation times Dependent of the Recurrent Responses

For practical situations, the observation process and the recurrent event process may be dependent. For this, Sun et al. (2007) proposed a semiparametric regression model considering dependent observation times. The dependence structure was modeled via a positive subject-specific shared frailty x_i given by

$$E\{N_i(t)|z_i, x_i\} = x_i^{\alpha}\mu_0(t)\exp\{\beta' z_i\}$$

for the event process and

$$\lambda_i(t) = x_i\lambda_0(t)\exp(\gamma' z_i)$$

for the intensity of the observation process under a nonhomogeneous Poisson assumption, with $\Lambda_0(\tau) = \int_0^t \lambda_0(s)ds = 1$ or $E(X_i) = 1$ assumed for identifiability. Estimating equation approaches were proposed for the estimate of regression parameters and the asymptotic normality of the proposed estimates were also established.

Zhao and Tong (2011) discussed the above models and generalized the model by replacing x_i^α by a completely unspecified function $g(x_i)$ and proposed a joint modeling approach and established the asymptotic normality of the resulting estimates.

Other than the observation process, for some situations one may suspect that the follow-up process may also be correlated with both the event process and the observation process. He et al. (2009) considered such cases with the main interest on the estimation of covariate effects on the event process after adjusting for the possible correlation among the three processes. Given z_i and two latent variables u_i and v_i, the model is:

$$E\{N_i(t)|z_i, u_i, v_i\} = \mu_N(t)\exp(x_i'\beta_1 + u_i\beta_2 + v_i\beta_3)$$

for the event process. For the observation process, the intensity function is given by

$$\lambda_{ih}(t) = \lambda_{0h}(t)\exp(z_i'\alpha_1 + u_i).$$

The hazard function of the follow-up time C_i^* is in form of

$$\lambda_{ic}(t) = \lambda_{0c}(t)\exp(z_i'\gamma_1 + u_i\gamma_2 + v_i).$$

A three-step estimation procedure was developed based on estimating equations for the above regression parameters.

For analyzing panel count data with observation process dependent of the recur-

rent event process, all the above methods were constructed by shared frailty models. Instead, one may wish to use marginal models on the response variable directly, with the correlation structure incorporated. Li et al. (2010) considered a marginal transformation model given by

$$E\{N_i(t)|Z_i(t), \mathcal{F}_{it}\} = g\{\mu_0(t)\exp(\beta' Z_i(t) + \alpha' h(\mathcal{F}_{it}))\},$$

where $\mathcal{F}_{it} = \{O_i(s), 0 \leq s < t\}$ is the history or filtration of the observation process $O(\cdot)$ up to time t^- on subject i, $h(\cdot)$ is a vector of known functions of $\mathcal{F}_{it}$, and $g(\cdot)$ is a known twice continuously differentiable and strictly increasing function. The observation process was modeled under conditional Poisson assumption with its intensity in form of

$$E\{dO_i(t)|Z_i(t)\} = \lambda_0(t)e^{\gamma' Z_i(t)}dt.$$

The regression parameters were estimated by estimating equation methods and they were shown to be consistent and asymptotically normal. Furthermore, both of the marginal models above can be extended to multivariate panel count data analysis with dependent observation processes (Li et al., 2011).

All the methods discussed above considered censoring for the follow-up times. In reality, however, there may also be some events terminating both the observation process and the recurrent event process, like death. For such cases, we will present marginal approaches that model the mean function of recurrent events among survivals in Chapters 2 and 3.

1.3 Nonparametric Comparisons with Panel Count Data

Besides estimation with panel count data, treatment comparison on the mean functions is another objective of the most interests. Consider p populations corresponding to p different treatments regarding the occurrences of some recurrent event. Let τ be the largest follow-up time and $N_i(t)$, K_i, $T_{i,j}$ be defined as in the previous section. Also, let $\Lambda_l(t)$ be the mean function for group l, *i.e.*, $\Lambda_l(t) = E\{N_i(t)\}$ for $i = 1, \ldots, n_l$, where n_l is the sample size in group l. Now our goal is to test the null hypothesis H_0: $\Lambda_1(t) = \cdots = \Lambda_p(t)$.

Thall and Lachin (1988) suggested first grouping the panel count data to K intervals, then using specially defined multivariate Wilcoxon-like rank test within the intervals. However, the test result may depend on how the intervals are divided.

Sun and Kalbfleisch (1993) and Sun and Fang (2003) proposed model-free test procedures for the two-sample ($p = 2$) comparison problem. Let Z_i represent a group indicator valued 0 or 1 for subject i, $i = 1, \ldots, n$, then the test statistic is in form of

$$U_{SF} = \sum_{i=1}^{n} Z_i \sum_{j=1}^{K_i} \{N_i(T_{i,j}) - \hat{\Lambda}(T_{i,j})\},$$

where $\hat{\Lambda}(T_{i,j})$ is the IRE as defined in (1.1). Under some regularity conditions and H_0, $n^{-1/2}U_{SF}$ can be approximated by the normal distribution with mean zero and variance

$$\hat{\sigma}^2_{SF} = \frac{1}{n} \sum_{i=1}^{n} \left[(Z_i - \bar{Z}) \{N_i(T_{i,j}) - \hat{\Lambda}(T_{i,j})\} \right]^2.$$

The above procedure requires that the treatment indicators Z_i's are independent and identically distributed random variables, which may not hold in practice. Park

et al. (2007) proposed a new class of two-sample nonparametric test procedures. Motivated by comparison methods of two survival functions (Pepe and Fleming, 1989; Petroni and Wolfe, 1994), the class of test statistics are

$$U_{PSZ} = \sqrt{\frac{n_1 n_2}{n}} \int_0^\tau W_n(t)\{\hat{\Lambda}_{n_1}(t) - \hat{\Lambda}_{n_2}(t)\} dG_n(t),$$

where $G_n(t) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{K_i} I(t_{ij} \leq t)$ and $\hat{\Lambda}_{n_1}(t)$, $\hat{\Lambda}_{n_2}(t)$ are the IREs for the mean functions of $\Lambda_1(t)$ and $\Lambda_2(t)$ in each individual group.

It could be shown that under H_0, the distribution of U_{PSZ} is asymptotically normal with mean zero and variance

$$\hat{\sigma}^2_{PSZ} = \frac{n_2}{n}\hat{\sigma}^2_1 + \frac{n_1}{n}\hat{\sigma}^2_2$$

with

$$\hat{\sigma}^2_l = \frac{1}{n_l}\sum_{i=1}^{n_l}\left[\sum_{j=1}^{K_i} W_n(T_{i,j})\{N_i(T_{i,j}) - \hat{\Lambda}_{n_l}(T_{i,j})\}\right]^2, \quad l = 1, 2.$$

Different weight functions may be chosen depending on the purpose of the study. For example, $W_n^{(1)}(t) = 1$, $W_n^{(2)}(t) = Y_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(t \leq T_{i,K_i})$ or $W_n^{(3)}(t) = \frac{Y_{n_1}(t)Y_{n_2}(t)}{Y_n(t)}$.

Instead of employing IRE or NPMPLE for the estimation of the mean functions as in the methods discussed above, one can consider using NPMLE for similar test procedures. Motivated by the idea used in Sun and Fang (2003) for two-sample comparisons, Balakrishnan and Zhao (2010) proposed the following test statistic with NPMLE $\hat{\Lambda}$:

$$U_{BZ} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i\left[\sum_{j=1}^{K_i-1}\hat{\Lambda}(T_{i,j})\{\frac{\Delta N_i(T_{i,j+1})}{\Delta\hat{\Lambda}(T_{i,j+1})} - \frac{\Delta N_i(T_{i,j})}{\Delta\hat{\Lambda}(T_{i,j})}\} + \hat{\Lambda}(T_{i,K_i})\{1 - \frac{\Delta\hat{\Lambda}(T_{i,K_i})}{\Delta\hat{\Lambda}(T_{i,K_i})}\}\right],$$

where $\Delta\hat{\Lambda}(T_{i,j}) = \hat{\Lambda}(T_{i,j}) - \hat{\Lambda}(T_{i,j-1})$, $\Delta N(T_{i,j}) = N(T_{i,j}) - N(T_{i,j-1})$ and $\hat{\Lambda}(t)$ is the NPMLE of the common mean function $\Lambda(t)$ under H_0.

Under some regularity conditions, U_{BZ} has an asymptotic normal distribution with mean vector 0 and covariance

$$\sigma^2_{BZ} = E\Bigg[(Z - E(Z))\Bigg\{\sum_{j=1}^{K-1}\Lambda_0(T_{1,j})\left(\frac{\Delta N(T_{1,j+1})}{\Delta\Lambda_0(T_{1,j+1})} - \frac{\Delta N(T_{1,j})}{\Delta\Lambda_0(T_{1,j})}\right)$$
$$+\Lambda_0(T_{1,K})\left(1 - \frac{\Delta N(T_{1,K})}{\Delta\Lambda_0(T_{1,K})}\right)\Bigg\}\Bigg]^2,$$

where $\Lambda_0(\cdot)$ is the true value of $\Lambda(\cdot)$. It was shown that σ^2_{BZ} can be estimated consistently by replacing $E(Z)$ and $\Lambda_0(t)$ with $\bar{Z} = \sum_{i=1}^{n} Z_i/n$ and $\hat{\Lambda}(t)$, respectively.

For p-sample ($p > 2$) comparison problems, Balakrishnan and Zhao (2010) further remarked that the above procedure can be extended with the test statistics being a similar form. Let Z_i be a p-dimensional vector of treatment indicators, with the lth element equal to 1 if subject i is from group l and 0 elsewhere. Then a generalized version of U_{BZ} was proved to follow an asymptotic normal distribution with mean vector $\mathbf{0}$. The covariance matrix estimate was also derived.

There are other procedures employing IRE or NPMPLE for p-sample comparisons, including Zhang (2006), Balakrishnan and Zhao (2009) and Balakrishnan and Zhao (2011). The former extended the two-sample test procedure in Park et al. (2007) with the test statistics in a similar form and a common weight function for all groups. The latter relaxed such an equal-weight requirement and used group-specific weight functions in their proposed test statistics.

One hidden assumption that all the test procedures above have in common is that the observation processes are identical across different treatment groups. However, as

noticed by many authors, the observation processes may differ among different groups of subjects. For this, Zhao and Sun (2011) proposed a class of nonparametric test procedures allowing different observation processes given as follows.

Let $G_n^{(l)}(t) = \frac{1}{n_l}\sum_{i\in S_l}\sum_{j=1}^{K_i^{(l)}} I(T_{i,j}^{(l)} \le t)$ and $G_n(t) = \sum_{l=1}^{k} p_l G_n^{(l)}(t)$ be the empirical observation process from group l and the overall empirical observation process respectively, with $p_l = n_l/n$. Also define

$$\Psi_n^{(l)} = \int_0^\tau W_n(t)\hat{\Lambda}_n^{(l)}(t)dG_n(t)$$

as a summary measure of the event history in group l, where $W_n(t)$'s are bounded weight processes, and

$$\hat{\sigma}_l^2 = \frac{1}{n_l}\sum_{i\in S_l}\left[\sum_{j=1}^{K_i^{(l)}} A_n^{(l)}(T_{i,j}^{(l)})\{N_i^{(l)}(T_{i,j}^{(l)}) - \hat{\Lambda}_n^{(l)}(T_{i,j}^{(l)})\}\right]^2,$$

$$A_n^{(l)}(t) = \sum_{r=1}^{k}\frac{n_r}{n}W_n(t)\frac{G_n^{(r)}(t) - G_n^{(r)}(t-)}{G_n^{(l)}(t) - G_n^{(l)}(t-)}.$$

Then their test statistics are given by

$$U_{ZS} = \sum_{l=1}^{k} c_l(\Psi_n^{(l)} - \bar{\Psi}_n)^2,$$

where $c_l = n_l/\hat{\sigma}_l^2$, $\bar{\Psi}_n = \sum_{l=1}^{k}\alpha_l\Psi_n^{(l)}$, $\alpha_l = c_l/(\sum_{l=1}^{k} c_l)$ for $l = 1, \ldots, p$. Under H_0, U_{ZS} asymptotically follows the central χ^2-distribution with $(p-1)$ degrees of freedom.

All the procedures mentioned above involve the estimation of a common mean function $\Lambda(t)$ under H_0 or group-specific mean function $\Lambda_l(t)$, using either NPMLE, IRE or NPMPLE. Such procedures can perform well when there are enough data

over all observation times, however, if one can only obtain rare observations on some observation times, the performance of the test procedures could be affected because the mean function estimator may not perform well at those observation times. We will discuss this issue in more details in Chapter 4 and propose a new class of test procedures overcoming the problem.

1.4 Outline of the Dissertation

The rest of this dissertation contains four parts about semiparametric and nonparametric methods for the analysis of panel count data from Chapter 2 to Chapter 5.

In Chapter 2, we consider regression analysis of panel count data in the presence of dependent observation processes and a terminal event. A semiparametric transformation model is presented for the mean function of the underlying recurrent event process among survivals. To estimate regression parameters, an estimating equation approach is proposed in which the inverse survival probability weighting technique is used. In addition, the asymptotic distribution of the proposed estimate is derived and a model checking procedure for the mean function model is presented. Simulation studies are conducted and the proposed approach is applied to the bladder cancer study described in Section 1.1.2.2.

Chapter 3 discusses regression analysis of multivariate panel count data in the presence of some terminal event. Furthermore, both the observation process and the terminal event may be correlated with the underlying recurrent event process of interest. A semiparametric additive model for the mean function of the recurrent event

process will be considered and an estimating equation based inference procedure will be developed for the estimation of the regression parameters. In the procedure, the inverse survival probability weighting technique is used and the asymptotic properties of the proposed estimators are established.

Chapter 4 considers nonparametric comparison based on panel count data. Most approaches that have been developed in the literature require an equal observation process for all subjects. However, such assumption may not hold in reality. A new class of test procedures are proposed that allow unequal observation processes for the subjects from different treatment groups, and both univariate and multivariate panel count data are considered. The asymptotic normality of the proposed test statistics is established and a simulation study is conducted to evaluate the finite sample properties of the proposed approach. The simulation results show that the proposed procedures work well for practical situations and especially for sparsely distributed data. They are applied to a set of panel count data from the skin cancer study described in Section 1.1.2.3. The dissertation concludes with Chapter 5, which discusses several directions for future research.

Chapter 2

ANALYZING PANEL COUNT DATA WITH DEPENDENT OBSERVATION PROCESSES AND A TERMINAL EVENT

2.1 Introduction

This chapter discusses semiparametric regression analysis of panel count data, which usually arise in longitudinal follow-up studies that concern some recurrent events and in which each study subject is observed only at discrete time points instead of continuously. In these situations, only the numbers of the events that occur between observation times, not their exact occurrence times, are observed. For example, consider the bladder cancer study discussed in Section 1.1.2.2 (Sun & Wei, 2000; Ghosh & Lin, 2002; Wellner & Zhang, 2007). In the study, the patients visited the clinical centers periodically and some patients died before the end of the follow-up. At each

visit, only the number of the bladder tumors that occurred since the last visit was recorded. That is, only panel count data are available about the tumor occurrence. Other fields that often produce such data include clinical trials, reliability experiments, sociological studies and tumorigenicity experiments.

As mentioned in Section 1.2, many authors have considered the analysis of panel count data. For example, Sun & Kalbfleisch (1995) and Wellner & Zhang (2000) investigated nonparametric estimation of the mean function of the underlying recurrent event process. Sun & Wei (2000), Cheng & Wei (2000), Zhang (2002) and Wellner & Zhang (2007) developed some semiparametric procedures for regression analysis of panel count data under the proportional mean models. More recently, Zhao, Balakrishnan, & Sun (2011) gave a relatively complete review of the literature on panel count data. In all of these methods and most of the existing approaches for panel count data, it was assumed that the censoring or stopping time for the follow-up is independent of the underlying recurrent event process of interest. In other words, there is no terminal event. In many situations, however, the follow-up of study subjects could be stopped by a terminal event, such as death, which precludes further recurrent events. For example, tumors would not develop after death. Furthermore, it is often the case that the terminal event is strongly correlated with recurrent events of interest. For example, a higher rate of recurrent events is often associated with an increased rate of death.

Unlike recurrent event data, which are available if all study subjects are under continuous observation, panel count data also involve an observation process that characterizes the observation times for each subject. In addition to the possible existence of a dependent terminal event, this observation process could be related to the underlying recurrent event process of interest too. As mentioned in Section 1.2.2.2, among others,

Huang, Wang, & Zhang (2006), He, Tong, & Sun (2009) and Li, Sun, & Sun (2010) proposed some semiparametric approaches for regression analysis of the panel count data with dependent observation processes. However, all these authors treated the death as an independent censoring variable or assumed that there does not exist a dependent terminal event. Note that terminal events are quite different from the ordinary censoring. When a terminal event occurs, the recurrent event will be stopped permanently, while with a dependent censoring, the recurrent event may still occur continuously, just cannot be observed. In the case of dependent death, the analysis that treats it as a simple dependent censoring could generally overestimate the occurrence rate of the recurrent events of interest.

In the presence of terminal events, there exists considerable work on regression analysis of recurrent event data and in this case, two approaches are commonly adopted. One is the marginal model approach that usually models the marginal rates of both recurrent and terminal events and leaves the correlation between the recurrent event process and the terminal event arbitrary (Cook & Lawless, 1997; Ghosh & Lin, 2002; Zhao, Zhou, & Sun, 2011). The other is the frailty model approach that often employs a latent variable to account for the correlation between the rates of recurrent and terminal events and assumes that these two event processes are independent given the frailty (Huang & Wang, 2004; Liu Wolfe, & Huang, 2004; Ye, Kalbfleish, & Schaubel, 2007; Zeng & Cai, 2010). However, the problem is much harder for panel count data, and it does not seem to exist an established procedure for panel count data with a dependent terminal event. In the following, a semiparametric marginal model approach will be developed for regression analysis of panel count data in the presence of a dependent terminal event. In addition, the proposed approach will also allow the existence of a

dependent or informative observation process.

The rest of this chapter is organized as follows. We will begin in Section 2.2 with introducing some notation and describing the proposed models that will be used throughout this chapter. In particular, we will present a class of semiparametric transformation models for the underlying recurrent event process of interest, which have great flexibility and allow a variety of patterns for the underlying recurrent event process. In Section 2.3, an estimating equation approach is developed for estimation of regression parameters. The approach leaves the correlation between the recurrent event and the terminal event unspecified and makes use of the inverse probability weighting technique to take into account the fact that the subjects who die cannot experience further occurrence of the events of interest. Section 2.4 gives the asymptotic properties of the proposed estimates and also presents a goodness-of-fit test procedure for checking the adequacy of the proposed models. Some simulation results are given in Section 2.5 and in Section 2.6, we apply the proposed methodology to the bladder cancer study described above. Section 2.7 contains some concluding remarks.

2.2 Notation and Models

Consider a longitudinal study concerning some recurrent events. Let $Y(t)$ denote the underlying point process representing the cumulative number of occurrences of the events of interest up to time t and $N(t)$ the observation process. In the following, we will assume that $N(t)$ is a continuous-time counting process with independent increments and $Y(t)$ is observed only at the time points where $N(t)$ jumps. Also it will be assumed that there exists a vector of external covariate process denoted by $Z(t)$ (Kalbfleisch

& Prentice, 2002) and a terminal event denoted by D that may be related to $Y(t)$. A common example of the terminal event is death and in this case, the correlation between $Y(t)$ and D occurs if, for example, the high recurrence rate of the events such as tumors means the increasing death risk. Also the subjects can not experience further observations and recurrent events after death. In this chapter, we will focus on the actual recurrent event and observation processes $Y^*(t) = Y(t \wedge D)$ and $N^*(t) = N(t \wedge D)$, where $a \wedge b = \min\{a, b\}$. Note that both $N^*(t)$ and $Y^*(t)$ will remain constants after D.

In practice, it is usually the case that there also exists a censoring or follow-up time C. That is, the follow-up is stopped by $T^* = C \wedge D$ and one only observes $\widetilde{Y}(t) = Y^*(t \wedge C)$ and $\widetilde{N}(t) = N^*(t \wedge C)$. Let $\{T_1, \cdots, T_K\}$ denote the time points at which $\widetilde{N}(t)$ jumps. Then $\widetilde{Y}(t)$ is observed only at these T_j's and K denotes the total number of observations. Suppose that the study consists of n independent subjects. Then the observed data have the form

$$\{\, \widetilde{N}_i(t), Z_i(t), T_i^*, I(D_i \le C_i), \widetilde{Y}_i(T_{i,1}), \ldots, \widetilde{Y}_i(T_{i,K_i});\ 0 \le t, T_{i,K_i} \le T_i^*,\ i = 1, \ldots, n \,\}\,.$$

Define $\mathcal{F}_t = \{N(s), 0 \le s < t\}$, the history or filtration of the observation process N up to time t^-, and $\mathcal{Z}(t) = \{Z(s), 0 \le s \le t\}$, the history of the covariate process. In the following, we will assume that given $\mathcal{Z}(t)$, the adjusted observation process $N^*(t)$ follows the proportional rate model

$$E\{dN^*(t)|\mathcal{Z}(t)\} = e^{\gamma_0' Z(t)} d\Lambda_0(t)\,, \tag{2.1}$$

where γ_0 is a vector of unknown parameters and $d\Lambda_0(\cdot)$ is an unspecified baseline rate

function. Also it will be assumed that C is independent of $\{N^*(t), Y^*(t), D\}$ conditional on $\mathcal{Z}(t)$.

To model the covariate effects on the recurrent event process $Y^*(t)$, we will assume that given $\mathcal{Z}(t)$, $\mathcal{F}_t$ and $D \geq t$, the conditional mean function of $Y^*(t)$ has the form

$$E\{Y^*(t)|\mathcal{Z}(t), \mathcal{F}_t, D \geq t\} = g\{\mu_0(t)e^{\beta_0' Z(t)+\alpha_0' h(\mathcal{F}_t)}\}, \tag{2.2}$$

where $g(\cdot)$ is a known twice continuously differentiable and strictly increasing function, $\mu_0(t)$ is an unspecified smooth function of t, α_0 and β_0 are vectors of unknown parameters, and $h(\cdot)$ is a vector of known functions of $\mathcal{F}_t$. Here $g(\cdot)$ can take many forms to account for various types of dependence of $Y^*(t)$ and $(\mathcal{Z}(t), \mathcal{F}_t)$. For example, $g(x) = x$ and $g(x) = \log x$ result in the proportional mean model and the additive mean model, respectively. Besides, we can also take g to be the commonly referred Box-Cox transformation, $g(x) = \frac{(x+1)^a-1}{a}$, where a is a constant. In particular, if $a = 0$, then $g(x) = \log(x+1)$. For the choice of $h(\cdot)$, there are also various forms can be taken. One example is $h(\mathcal{F}_t) = N(t-)$ if it is believed that $Y(t)$ may depend on the total number of visits up to t. It will be assumed that $N_i^*(t)$ and $Y_i^*(t)$ are independent given $\mathcal{Z}_i(t)$, $D_i \geq t$ and $\mathcal{F}_{it}$.

Note that here we focus on the adjusted mean function and the same idea has been used for the analysis of recurrent event data by several authors (e.g., Cook & Lawless, 1997; Ghosh & Lin, 2002). Among others, one advantage is that no assumption is needed for the recurrent event process after the terminal event (Luo & Huang, 2010). In contrast, if one simply treats death as a censoring variable as in most of the existing methods, one could overestimate the mean function and it is obvious that the analysis would not take into account the fact that the subjects who die can not experience any

further recurrent events.

The model (2.2) is commonly referred to as the semiparametric transformation model. It was motivated by Lin, Wei, & Ying (2001) and Sun et al. (2005) and is a generalization of the model proposed in Li, Sun, & Sun (2010). It can be easily seen that this model is quite flexible and allows various types of the dependence of the mean function of $Y^*(t)$ on $Z(t)$ and $N^*(t)$. If there does not exist death or $D = \infty$, it is obvious that $\widehat{E}\{Y^*(t)|\mathcal{Z}(t), \mathcal{F}_t, D \geq t\}$ reduces to $\widehat{E}\{Y^*(t)|\mathcal{Z}(t), \mathcal{F}_t\}$. In the presence of death, one can show that the marginal mean function has the form

$$E\{Y^*(t)|\mathcal{Z}(t), \mathcal{F}_t\} = \int_0^t S(u|Z) E\{dY^*(u)|\mathcal{Z}(u), \mathcal{F}_u, D \geq u\}$$

given $\mathcal{Z}(t)$ and $\mathcal{F}_t$ and after adjusting the fact that the death precludes further recurrent events, where $S(t|Z) = P(D \geq t|\mathcal{Z}(t))$. It is easy to see that in this case, we have

$$\widehat{E}\{Y^*(t)|\mathcal{Z}(t), \mathcal{F}_t, D \geq t\} > \widehat{E}\{Y^*(t)|\mathcal{Z}(t), \mathcal{F}_t\}$$

for t greater than the first observed death time.

In reality, the terminal event time D may also depend on covariates $Z(t)$. For this, we will assume that D follows the proportional hazards model given by

$$\lambda^d(t|Z(t)) = \lambda_0^d(t)\, e^{\delta_0' Z(t)}\,, \tag{2.3}$$

where $\lambda_0^d(t)$ is an unspecified baseline hazard function and δ_0 is a vector of unknown regression parameters. Under the above model, we have $S(t|Z) = \exp\{-\int_0^t e^{\delta_0' Z(s)} d\Delta_0(s)\}$, where $\Delta_0(t) = \int_0^t \lambda_0^d(s) ds$.

2.3 Estimation Procedures

In this section, we will present some inference procedures for the models described in the previous section. Let β_0, α_0 and γ_0 denote the true values of β, α and γ, respectively, and define $X_i(t) = (Z_i(t)', h(\mathcal{F}_{it})')'$, $\theta = (\beta', \alpha')'$, $\theta_0 = (\beta_0', \alpha_0')'$. First we will show that $\widetilde{Y}_i(t)d\widetilde{N}_i(t) - I(C_i \geq t)g\{\mu_0(t)e^{\theta_0' X_i(t)}\}e^{\gamma_0' Z_i(t)}d\Lambda_0(t)$ is a mean-zero stochastic process. This is true because under models (2.1) and (2.2) and the conditional independent assumptions for $Y_i^*(t)$, $N_i^*(t)$ and C_i, we have

$$\begin{aligned}
E\{\widetilde{Y}_i(t)d\widetilde{N}_i(t)\} &= E\Big[E\{I(C_i \geq t)Y_i^*(t)dN_i^*(t)|\mathcal{Z}_i(t), \mathcal{F}_{it}\}\Big] \\
&= E\Big[E\{I(C_i \geq t)|\mathcal{Z}_i(t)\}E\{Y_i^*(t)dN_i^*(t)|\mathcal{Z}_i(t), \mathcal{F}_{it}\}\Big] \\
&= E\Big[E\{I(C_i \geq t)|\mathcal{Z}_i(t)\}E\{Y_i^*(t)|D_i \geq t, \mathcal{Z}_i(t), \mathcal{F}_{it}\}E\{dN_i^*(t)|\mathcal{Z}_i(t)\}\Big] \\
&= E\Big[E\{I(C_i \geq t)g\{\mu_0(t)e^{\theta_0' X_i(t)}\}e^{\gamma_0' Z_i(t)}d\Lambda_0(t)|\mathcal{Z}_i(t), \mathcal{F}_{it}\}\Big] \\
&= E\Big[I(C_i \geq t)g\{\mu_0(t)e^{\theta_0' X_i(t)}\}e^{\gamma_0' Z_i(t)}d\Lambda_0(t)\Big],
\end{aligned}$$

where the third equality holds because

$$\begin{aligned}
& E\big\{Y_i^*(t)dN_i^*(t)|\mathcal{Z}_i(t), \mathcal{F}_{it}\big\} \\
= \; & E\Big\{E\big\{Y_i^*(t)dN_i^*(t)|D_i, \mathcal{Z}_i(t), \mathcal{F}_{it}\big\}\Big\} \\
= \; & E\big\{Y_i^*(t)dN_i^*(t)|D_i \geq t, \mathcal{Z}_i(t), \mathcal{F}_{it}\big\}P(D_i \geq t|\mathcal{Z}_i(t)) + 0 \times P(D_i < t|\mathcal{Z}_i(t)) \\
= \; & E\big\{Y_i^*(t)|D_i \geq t, \mathcal{Z}_i(t), \mathcal{F}_{it}\big\}E\big\{dN_i^*(t)|D_i \geq t, \mathcal{Z}_i(t)\big\}P(D_i \geq t|\mathcal{Z}_i(t)) \\
= \; & E\big\{Y_i^*(t)|D_i \geq t, \mathcal{Z}_i(t), \mathcal{F}_{it}\big\}E\big\{dN_i^*(t)|\mathcal{Z}_i(t)\big\}.
\end{aligned}$$

Note that in practice, C_i is unobservable when $D_i \leq C_i$. Thus the mean-zero stochastic process given above can not be directly used to construct estimating equations. To overcome this, we employ the inverse probability weighting procedure to replace

$I(C_i \geq t)$ $(i = 1, \ldots, n)$ in the process. Specifically, define $\omega_i(t) = I(T_i^* \geq t)/S(t|Z_i)$. Note that $E\{I(T_i^* \geq t)|\mathcal{Z}_i(t)\} = E\{I(C_i \geq t)|\mathcal{Z}_i(t)\}S(t|Z_i)$ based on the independence of C_i and D_i given $Z_i(\cdot)$. This gives that $E\{\omega_i(t)|\mathcal{Z}_i(t)\} = E\{I(C_i \geq t)|\mathcal{Z}_i(t)\}$. Define

$$dM_i(t; \theta, \gamma) = \widetilde{Y}_i(t)d\widetilde{N}_i(t) - \omega_i(t)g\{\mu_0(t)e^{\theta' X_i(t)}\}e^{\gamma' Z_i(t)}d\Lambda_0(t)$$

and $dM_i(t) = dM_i(t; \theta_0, \gamma_0)$. Then it follows from models (2.1) and (2.2) that $E[dM_i(t)] = 0$ for $i = 1, ..., n$. Note that here $\omega_i(t)$ is still unobservable, but it can be easily estimated by $\hat{\omega}_i(t) = I(T_i^* \geq t)/\hat{S}(t|Z_i)$, where $\hat{S}(t|Z_i) = \exp\big\{-\int_0^t \exp\{\hat{\delta}' Z_i(s)\}d\hat{\Delta}_0(s)\big\}$ with $\hat{\delta}$ and $\hat{\Delta}_0(t)$ being the maximum partial likelihood Breslow estimators of δ and $\Delta_0(t)$, respectively, given by model (2.3). By following the similar arguments as those in Lin, Wei, & Ying (2001), it can be shown that for large n, the estimator $\hat{\omega}_i(t)$ always exists and is unique and consistent.

For estimation of θ and $\mu_0(t)$, first assume that γ and Λ_0 are known. Then it is natural to employ the following estimating functions

$$\sum_{i=1}^{n} \Big[\widetilde{Y}_i(t)d\widetilde{N}_i(t) - \hat{\omega}_i(t)g\{\mu_0(t)e^{\theta' X_i(t)}\}e^{\gamma' Z_i(t)}d\Lambda_0(t)\Big] = 0, \;\; 0 \leq t \leq \tau, \tag{2.4}$$

and

$$U_\theta(\theta; \gamma) = \sum_{i=1}^{n} \int_0^\tau W(t)X_i(t)\Big[\widetilde{Y}_i(t)d\widetilde{N}_i(t) - \hat{\omega}_i(t)g\{\mu_0(t)e^{\theta' X_i(t)}\}e^{\gamma' Z_i(t)}d\Lambda_0(t)\Big] = 0\,, \tag{2.5}$$

where τ is the longest follow-up time and $W(t)$ is a possibly data-dependent weight function. Of course, γ and Λ_0 are unknown in general, but they can be easily estimated based on the recurrent event data observed on model (2.1) (Cook & Lawless, 2007).

Specifically, define

$$dM_i^*(t;\gamma) = d\widetilde{N}_i(t) - \omega_i(t)e^{\gamma' Z_i(t)}d\Lambda_0(t)$$

and $dM_i^*(t) = dM_i^*(t;\gamma_0)$. It is easy to see that $M_i^*(t)$ is a mean-zero stochastic process. It follows that the consistent estimators of γ and $\Lambda_0(t)$, denoted by $\hat{\gamma}$ and $\hat{\Lambda}_0(t)$, respectively, can be obtained by solving the following two estimating equations

$$U_\gamma(\gamma) = \sum_{i=1}^{n}\int_0^\tau \big\{Z_i(t) - \bar{Z}(t;\gamma)\big\}d\widetilde{N}_i(t) = 0, \tag{2.6}$$

and

$$\sum_{i=1}^{n}\Big[d\widetilde{N}_i(t) - \hat{\omega}_i(t)e^{\gamma' Z_i(t)}d\Lambda_0(t)\Big] = 0\,.$$

In the above, $\bar{Z}(t;\gamma) = S^{(1)}(t;\gamma)/S^{(0)}(t;\gamma)$ and $S^{(k)}(t;\gamma) = n^{-1}\sum_{i=1}^{n}\hat{\omega}_i(t)Z_i(t)^k e^{\gamma' Z_i(t)}$, $k = 0, 1$. In particular, we have

$$\hat{\Lambda}_0(t;\gamma) = \int_0^t \frac{d\bar{N}(u)}{S^{(0)}(u;\gamma)}, \tag{2.7}$$

where $\bar{N}(t) = n^{-1}\sum_{i=1}^{n}\widetilde{N}_i(t)$. Given $\hat{\gamma}$ and $\hat{\Lambda}_0(t)$, one can estimate θ and $\mu_0(t)$ by plugging them into Equations (2.4) and (2.5).

Let $\hat{\theta}$ and $\hat{\mu}_0(t;\hat{\theta},\hat{\gamma})$ denote the estimators of θ and $\mu_0(t)$ defined above. In general, there are no closed forms for these estimators except some special cases. One such case is $g(t) = t^m$ and in this situation, $\hat{\mu}_0(t;\theta,\gamma)$ has an explicit expression, where m is a positive number. Another special case is when $g(t) = \log t$ and in this case, we have

$$\begin{aligned}\hat{\theta} \;&= \Big\{\textstyle\sum_{i=1}^{n}\int_0^\tau W(t)\big\{X_i(t) - \bar{X}(t;\hat{\gamma})\big\}X_i'(t)\hat{\omega}_i(t)e^{\hat{\gamma}' Z_i(t)}d\hat{\Lambda}_0(t,\hat{\gamma})\Big\}^{-1}\\ &\quad\times\textstyle\sum_{i=1}^{n}\int_0^\tau W(t)\big\{X_i(t) - \bar{X}(t;\hat{\gamma})\big\}\tilde{Y}_i(t)d\tilde{N}_i(t),\end{aligned}$$

and

$$\hat{\mu}_0(t;\theta,\gamma) = \exp\Big\{\frac{\sum_{i=1}^n \tilde{Y}_i(t)d\tilde{N}_i(t)}{\sum_{i=1}^n \hat{\omega}_i(t)e^{\gamma' Z_i(t)}d\Lambda_0(t)} - \theta'\bar{X}(t;\gamma)\Big\},$$

where

$$\bar{X}(t;\gamma) = \frac{\sum_{i=1}^n X_i(t)\hat{\omega}_i(t)e^{\hat{\gamma}' Z_i(t)}}{\sum_{i=1}^n \hat{\omega}_i(t)e^{\hat{\gamma}' Z_i(t)}}.$$

To implement the estimation procedure proposed above, one needs to choose the link function g and the weight function W. As commented by Li, Sun, & Sun (2010) and others, this is usually difficult and a common strategy is to try several choices and compare the obtained results.

2.4 Asymptotic Properties of $\hat{\theta}$ and Model Assessment

In this section, we will establish the asymptotic properties of $\hat{\theta}$ and present a goodness-of-fit test procedure for assessing the appropriateness of model (2.2). To establish the asymptotic properties, define $N_i^d(t) = I(D_i \le t, D_i \le C_i)$ and $M_i^d(t) = N_i^d(t) - \int_0^t I(T_i^* \ge s)e^{\delta_0' Z_i(s)}d\Delta_0(s)$, $i = 1, ..., n$. Then it is easy to see that the $M_i^d(t)$'s are zero-mean martingale processes. Also define

$$\hat{M}_i^d(t) = N_i^d(t) - \int_0^t I(T_i^* \ge s)e^{\hat{\delta}' Z_i(s)}d\hat{\Delta}_0(s)\ , \hat{M}_i^*(t) = \widetilde{N}_i(t) - \int_0^t \hat{\omega}_i(s)e^{\hat{\gamma}' Z_i(s)}d\hat{\Lambda}_0(s;\hat{\gamma}),$$

$$\hat{M}_i(t) = \int_0^t \widetilde{Y}_i(s)d\widetilde{N}_i(s) - \int_0^t \hat{\omega}_i(s)g\{\hat{\mu}_0(s;\hat{\theta},\hat{\gamma})e^{\hat{\theta}' X_i(s)}\}e^{\hat{\gamma}' Z_i(s)}d\hat{\Lambda}_0(s;\hat{\gamma}),$$

$$\hat{E}_X(t;\theta,\gamma) = \frac{\sum_{i=1}^n X_i(t)\hat{\omega}_i(t)\dot{g}\{\hat{\mu}_0(t;\theta,\gamma)e^{\theta' X_i(t)}\}e^{\theta' X_i(t)+\gamma' Z_i(t)}}{\sum_{i=1}^n \hat{\omega}_i(t)\dot{g}\{\hat{\mu}_0(t;\theta,\gamma)e^{\theta' X_i(t)}\}e^{\theta' X_i(t)+\gamma' Z_i(t)}},$$

$$\hat{\Upsilon}(t;\theta,\gamma) = n^{-1}\sum_{i=1}^{n}\big\{X_i(t) - \hat{E}_X(t;\theta,\gamma)\big\}\hat{\omega}_i(t)g\big\{\hat{\mu}_0(t;\theta,\gamma)e^{\theta' X_i(t)}\big\}e^{\gamma' Z_i(t)}\,,$$

$$R^{(k)}(t;\delta) = n^{-1}\sum_{i=1}^{n} I(T_i^* \geq t)e^{\delta' Z_i(t)}Z_i(t)^{\otimes k},\ k = 0,1,2,$$

$$\hat{H}(t;Z_i) = \int_0^t e^{\hat{\delta}' Z_i(u)}\Big\{Z_i(u) - \frac{R^{(1)}(t;\hat{\delta})}{R^{(0)}(t;\hat{\delta})}\Big\}d\hat{\Delta}_0(u;\hat{\delta})\,,$$

and

$$\hat{\Omega}_\delta = n^{-1}\sum_{i=1}^{n}\int_0^\tau \left[\frac{R^{(2)}(t;\hat{\delta})}{R^{(0)}(t;\hat{\delta})} - \Big\{\frac{R^{(1)}(t;\hat{\delta})}{R^{(0)}(t;\hat{\delta})}\Big\}^{\otimes 2}\right]dN_i^d(t)\,.$$

In the above, $\dot{g} = dg(t)/dt$, $r^{(k)}(t) = \lim_{n\to\infty} R^{(k)}(t;\delta_0)$ with $k = 0,1,2$, and $v^{\otimes 2} = vv'$ for a vector v. Let $s^{(0)}(t)$, $s^{(1)}(t)$, $e_x(t)$, $\Upsilon(t)$ and Ω_δ denote the limits of $S^{(0)}(t;\gamma_0)$, $S^{(1)}(t;\gamma_0)$, $\hat{E}_X(t;\theta_0,\gamma_0)$, $\hat{\Upsilon}(t;\theta_0,\gamma_0)$ and $\hat{\Omega}_\delta$, respectively, and $\bar{z}(t) = s^{(1)}(t)/s^{(0)}(t)$. The following theorem gives the consistency and asymptotically normality of $\hat{\theta}$.

Theorem 2.1. Assume that the conditions (C1)-(C5) given in Appendix A.1 hold. Then $\hat{\theta}$ is a consistent estimator of θ_0 and the distribution of $n^{1/2}(\hat{\theta}-\theta_0)$ can be asymptotically approximated by the normal distribution with mean zero and the covariance matrix $\hat{A}_\theta^{-1}\hat{\Sigma}\hat{A}_\theta^{-1}$, where $\hat{\Sigma} = n^{-1}\sum_{i=1}^n(\hat{\xi}_{1i} - \hat{\xi}_{2i} - \hat{\xi}_{3i})^{\otimes 2}$,

$$\hat{\xi}_{1i} = \int_0^\tau W(t)\Big(X_i(t) - \hat{E}_X(t;\hat{\theta},\hat{\gamma})\Big)d\hat{M}_i(t)\,,$$

$$\hat{\xi}_{2i} = \int_0^\tau \Big\{\frac{W(t)\hat{\Upsilon}(t;\hat{\theta},\hat{\gamma})}{S^{(0)}(t;\hat{\gamma})} + \hat{A}_\gamma\hat{\Omega}_\gamma^{-1}\Big(Z_i(t) - \bar{Z}(t;\hat{\gamma})\Big)\Big\}d\hat{M}_i^*(t)\,,$$

$$\hat{\xi}_{3i} = \int_0^\tau \Big\{\hat{A}_\gamma\hat{\Omega}_\gamma^{-1}\hat{Q}_1\Big(Z_i(t) - \frac{R^{(1)}(t;\hat{\delta})}{R^{(0)}(t;\hat{\delta})}\Big) + \hat{A}_\gamma\hat{\Omega}_\gamma^{-1}\frac{\hat{Q}_2(t)}{R^{(0)}(t;\hat{\delta})} + \frac{\hat{B}_1(t)}{R^{(0)}(t;\hat{\delta})}$$

$$+\hat{B}_2\Big(Z_i(t) - \frac{R^{(1)}(t;\hat{\delta})}{R^{(0)}(t;\hat{\delta})}\Big)\Big\}d\hat{M}_i^d(t)\,,$$

$$\hat{A}_\gamma = n^{-1}\sum_{i=1}^{n}\int_0^\tau W(t)\hat{\omega}_i(t)g\{\hat{\mu}_0(t;\hat{\theta},\hat{\gamma})e^{\hat{\theta}'X_i(t)}\}e^{\hat{\gamma}'Z_i(t)}\big[X_i(t)-\hat{E}_X(t;\hat{\theta},\hat{\gamma})\big]$$

$$\times\big[Z_i(t)-\bar{Z}(t;\hat{\gamma})\big]'d\hat{\Lambda}_0(t;\hat{\gamma})\,,$$

$$\hat{\Omega}_\gamma = n^{-1}\sum_{i=1}^{n}\int_0^\tau \{Z_i(t)-\bar{Z}(t;\hat{\gamma})\}^{\otimes 2}\hat{\omega}_i(t)e^{\hat{\gamma}'Z_i(t)}d\hat{\Lambda}_0(t;\hat{\gamma})\,,$$

$$\hat{B}_1(t) = n^{-1}\sum_{i=1}^{n}e^{\hat{\delta}'Z_i(t)}\int_0^\tau I(t<s)\hat{B}_i^*(s)d\hat{\Lambda}_0(s;\hat{\gamma})\,,$$

$$\hat{B}_2 = n^{-1}\sum_{i=1}^{n}\int_0^\tau \hat{B}_i^*(t)\hat{H}(t;Z_i)'\hat{\Omega}_\delta^{-1}d\hat{\Lambda}_0(t;\hat{\gamma})\,,$$

$$\hat{B}_i^*(t) = W(t)\hat{\omega}_i(t)e^{\hat{\gamma}'Z_i(t)}\Big[\big\{X_i(t)-\hat{E}_X(t;\hat{\theta},\hat{\gamma})\big\}g\big\{\hat{\mu}_0(t;\hat{\theta},\hat{\gamma})e^{\hat{\theta}'X_i(t)}\big\}-\frac{\hat{\Upsilon}(t;\hat{\theta},\hat{\gamma})}{S^{(0)}(t;\hat{\gamma})}\Big],$$

$$\hat{Q}_1 = n^{-1}\sum_{i=1}^{n}\int_0^\tau \big\{Z_i(t)-\bar{Z}(t;\hat{\gamma})\big\}\hat{Q}_3(t;Z_i)'\hat{\Omega}_\delta^{-1}d\hat{M}_i^*(t)\,,$$

$$\hat{Q}_2(t) = n^{-1}\sum_{i=1}^{n}e^{\hat{\delta}'Z_i(t)}\int_0^\tau \big\{Z_i(u)-\bar{Z}(u;\hat{\gamma})\big\}I(u\geq t)d\hat{M}_i^*(u)\,,$$

and

$$\hat{Q}_3(t;Z_i) = \int_0^t \big\{Z_i(u)-\bar{Z}(u;\hat{\gamma})\big\}e^{\hat{\delta}'Z_i(u)}d\hat{\Delta}_0(u;\hat{\delta})\,.$$

For a given data set, one question of practical interest is to assess the adequacy of the models described in Section 2.2. For both models (2.1) and (2.3), note that one observes complete data and several procedures have been developed in the literature for checking their goodness-of-fits (Lin et al., 1993; Lin et al., 2010; Schoenfeld, 1982). So in the following, we will focus on model (2.2) and develop an omnibus goodness-of-fit procedure.

Let the $\widehat{M}_i(t)$'s be defined as above. Note that they represent the differences between the observed and model-predicted numbers of events by time t. Thus it is natural to

construct a test statistic based on them. Following Sun et al. (2007a), we consider the following cumulative sums of residual process

$$\mathcal{F}(t,x) = n^{-1/2}\sum_{i=1}^{n}\int_0^t I(X_i(u)\le x)d\hat{M}_i(u),$$

where the event $\{X_i(u)\le x\}$ means that each components of $X_i(u)$ is not greater than the respective component of x. We will show in Appendix A.2 that the null distribution of $\mathcal{F}(t,x)$ can be approximated by a zero-mean Gaussian process

$$\widehat{\mathcal{F}}(t,x) = n^{-1/2}\sum_{i=1}^{n}\Big\{\hat{\eta}_{1i}(t,x) - \hat{\Phi}_\gamma(t,x)\Omega_\gamma^{-1}\hat{\eta}_{2i} - \hat{\Phi}_\theta(t,x)\hat{A}_\theta^{-1}\hat{\eta}_{3i}\Big\}G_i\,. \tag{2.8}$$

In the above, $G_1,\ldots,G_n$ are independent standard normal variables independent of the observed data,

$$\hat{\eta}_{1i}(t,x) = \int_0^t\{I(X_i(u)\le x) - \hat{E}_I(u,x)\}d\hat{M}_i(u) - \int_0^t\frac{\tilde{\Upsilon}(u,x)}{S^{(0)}(u;\hat{\gamma})}d\hat{M}_i^*(u)$$

$$-\int_0^t\left\{\frac{\tilde{B}_1(u,t,x)}{R^{(0)}(u;\hat{\delta})} + \tilde{B}_2(t,x)\Big(Z_i(u) - \frac{R^{(1)}(u;\hat{\delta})}{R^{(0)}(u;\hat{\delta})}\Big)\right\}d\hat{M}_i^d(u)\,,$$

$$\hat{\eta}_{2i} = \int_0^\tau\Big[Q_1\Big\{Z_i(t) - \frac{R^{(1)}(t;\hat{\delta})}{R^{(0)}(t;\hat{\delta})}\Big\} + \frac{Q_2(t)}{R^{(0)}(t;\hat{\delta})}\Big]dM_i^d(t) + \int_0^\tau\big\{Z_i(t) - \bar{Z}(t;\hat{\gamma})\big\}dM_i^*(t)\,,$$

$$\hat{\eta}_{3i} = \hat{\xi}_{1i} - \hat{\xi}_{2i} - \hat{\xi}_{3i},$$

$$\tilde{\Upsilon}(s,x) = n^{-1}\sum_{i=1}^{n}\big\{I(X_i(s)\le x) - \hat{E}_I(s,x)\big\}\hat{\omega}_i(s)g\big\{\hat{\mu}_0(s)e^{\hat{\theta}'X_i(s)}\big\}e^{\hat{\gamma}'Z_i(s)},$$

$$\tilde{B}_1(u,t,x) = n^{-1}\sum_{i=1}^{n}e^{\hat{\delta}'Z_i(u)}\int_0^t I(u<s)\tilde{B}_i^*(s,x)d\hat{\Lambda}_0(s),$$

$$\tilde{B}_2(t,x) = n^{-1}\sum_{i=1}^{n}\int_0^t \tilde{B}_i^*(s,x)\hat{H}(s;Z_i)'\hat{\Omega}_\delta^{-1}d\hat{\Lambda}_0(s),$$

$$\tilde{B}_i^*(s,x) = \hat{\omega}_i(s)e^{\hat{\gamma}'Z_i(s)}\Big[\big\{I(X_i(s)\le x) - \hat{E}_I(s,x)\big\}g\big\{\hat{\mu}_0(s)e^{\hat{\theta}'X_i(s)}\big\} - \frac{\tilde{\Upsilon}(s;x)}{S^{(0)}(s;\hat{\gamma})}\Big],$$

$$\hat{\Phi}_\gamma(t,x) = n^{-1}\sum_{i=1}^{n}\int_0^t \big[I(X_i(u)\le x) - \hat{E}_I(u,x)\big]\big[Z_i(u) - \bar{Z}(u;\hat{\gamma})\big]'\hat{\omega}_i(u)$$

$$\times g\{\hat{\mu}_0(u)e^{\hat{\theta}'X_i(u)}\}e^{\hat{\gamma}'Z_i(u)}d\hat{\Lambda}_0(u;\hat{\gamma}),$$

$$\hat{\Phi}_\theta(t,x) = n^{-1}\sum_{i=1}^{n}\int_0^t I(X_i(u)\le x)\big\{X_i(t) - \hat{E}_X(t;\hat{\theta},\hat{\gamma})\big\}'\hat{\mu}_0(t)\hat{\omega}_i(t)$$

$$\times\dot{g}\{\hat{\mu}_0(t)e^{\hat{\theta}'X_i(t)}\}e^{\hat{\theta}'X_i(t)+\hat{\gamma}'Z_i(t)}d\hat{\Lambda}_0(t,\hat{\gamma}),$$

and

$$\hat{E}_I(u,x) = \frac{\sum_{i=1}^n I(X_i(u)\le x)\hat{\omega}_i(u)\dot{g}\{\hat{\mu}_0(u)e^{\hat{\theta}'X_i(u)}\}e^{\hat{\theta}'X_i(u)+\hat{\gamma}'Z_i(u)}}{\sum_{i=1}^n \hat{\omega}_i(u)\dot{g}\{\hat{\mu}_0(u)e^{\hat{\theta}'X_i(u)}\}e^{\hat{\theta}'X_i(u)+\hat{\gamma}'Z_i(u)}}.$$

Based on (2.8), it is easy to see that one could obtain a large number of realizations from $\widehat{\mathcal{F}}(t,x)$ by repeatedly generating the standard normal random sample $\{G_1,\ldots,G_n\}$ while fixing the observation data. Thus to check the validity of model (2.2), one can plot these realizations of $\widehat{\mathcal{F}}(t,x)$ along with the observed $\mathcal{F}(t,x)$ and examine any unusual pattern of $\mathcal{F}(t,x)$ compared to the realizations. Furthermore, a formal lack-of-fit test can be constructed based on the statistic $\sup_{0\le t\le\tau,x}|\mathcal{F}(t,x)|$ and the corresponding p-value can be obtained by comparing the observed value of $\sup_{0\le t\le\tau,x}|\mathcal{F}(t,x)|$ to a large number of realizations from $\sup_{0\le t\le\tau,x}|\widehat{\mathcal{F}}(t,x)|$.

2.5 A Simulation Study

In this section, we report some results obtained from a simulation study conducted

to assess the finite sample behavior of the estimation procedure proposed in the previous section. In the study, the covariate Z was assumed to be a Bernoulli random variable with the probability of success being 0.5 and the censoring time C was generated from the uniform distribution $U(\tau/4, \tau)$ with $\tau = 1$. To generate the correlation between the recurrent event process of interest and the terminal event, it was assumed that there is a latent variable v following the positive stable distribution with Laplace transformation $L(s) = \exp(-s^\rho)$ (Luonga & Doray, 2009). Here we took $\rho = 0.7$ or 1.0. Given v and Z, the death time D was assumed to have the hazard function $\lambda(t|Z, v) = 0.2v \exp\{0.5Z\}$. It can be shown that D satisfies model (2.3) with $S(t|Z_i) = \exp\{-(0.2te^{0.5Z_i})^\rho\}$. For the observation process, we assumed that given Z_i and T_i^*, $\tilde{N}_i(t)$ was a nonhomogeneous Poisson process on $[0, \tau]$ with $E\{d\tilde{N}_i(t)|Z_i, T_i^*\} = 10S^{-1}(t|Z_i)e^{0.5Z_i} I(T_i^* \geq t)$. This gives $E\{dN_i^*(t)|Z_i\} = e^{\gamma_0 Z_i} d\Lambda_0(t)$ with $\gamma_0 = 0.5$ and $\Lambda_0(t) = 10t$. Also given Z_i and T_i^*, the number of observations K_i was generated from the Poisson distribution with mean $10 \int_0^{T_i^*} S^{-1}(t|Z_i)e^{0.5Z_i} dt$ and the observation times $(t_{i,1}, \ldots, t_{i,K_i})$ were taken to be the order statistics of a random sample of size K_i from the uniform distribution over $(0, T_i^*)$.

For the generation of panel counts $Y_i^*(t_{i,j})$, given K_i and $(t_{i,1}, \ldots, t_{i,K_i})$, we assumed that

$$Y_i^*(t_{i,j}) = Y_i^{**}(t_{i,1}) + Y_i^{**}(t_{i,2} - t_{i,1}) + \cdots + Y_i^{**}(t_{i,j} - t_{i,j-1})$$

with $t_{i,0} = 0$, $j = 1, \ldots, K_i$. In the above, given v_i, Z_i and $\mathcal{F}_{i,j}$, it was assumed that $Y_i^{**}(s)$ and $Y_i^{**}(t - s)$ were the mixed Poisson distributions with the conditional mean functions

$$\phi(v_i, s) g\{\mu_0(s) e^{\beta Z_i + \alpha h(\mathcal{F}_{i,s})}\}$$

and

$$\phi(v_i,t)g\{\mu_0(t)e^{\beta Z_i+\alpha h(\mathcal{F}_{i,t})}\}-\phi(v_i,s)g\{\mu_0(s)e^{\beta Z_i+\alpha h(\mathcal{F}_{i,s})}\},$$

respectively, where

$$\phi(v_i,t)=e^{-v_i}\exp\{(1+\lambda_0^d t\,e^{\delta_0 Z_i})^\rho-(\lambda_0^d t\,e^{\delta_0 Z_i})^\rho\}.$$

One can show that $E\{\phi(v_i,t)|Z_i,\mathcal{F}_{it},D_i\geq t\}=1$. The results reported below are based on 500 replications and with the sample size $n=200$ or 300.

Table 2.1 presents the results obtained for estimation of β and α based on the simulated data with the true values of (β,α) being equal to $(0,0)$ $(0.5,0)$, $(0,0.1)$ or $(0.5,0.1)$, $g(t)=t$, $\mu_0(t)=t$, $h(\mathcal{F}_{i,t})=N_i(t-)$ and $W(t)=1$. The results include the estimated biases (BIAS) given by the averages of the estimators minus their true values, the sampling standard errors (SSE), the averages of the estimated standard errors (SEE), and the 95% empirical coverage probabilities (CP). The results suggest that the proposed approach seems to perform well. Specifically, they indicate that the proposed estimators seem to be unbiased and there is a good agreement between the estimated and empirical standard errors. Also the coverage probabilities are reasonable and consistent with the nominal levels and as expected, the estimated standard errors became smaller as the sample size increased.

The results presented in Table 2.2 were also about the estimation of β and α and obtained under the same set-up as those in Table 2.1 except that we used different link functions $g(t)=\log(t)$ and $\mu_0(t)=e^t$. It can be seen that they gave similar conclusions as those from Table 2.1. Both tables also suggest that it seems that the parameter α can be estimated more accurately than the parameter β. We also considered other

set-ups such as those with different link functions and obtained similar results.

2.6 An Application

Now we apply the statistical approach proposed in the previous sections to the panel count data arising from the bladder cancer study described above. For the analysis, following Li, Sun, & Sun (2010) and others, we will focus on the data from the 85 bladder cancer patients in thiotepa (38) and placebo (47) groups. As mentioned before, the original study includes another treatment but many authors have shown that it did not have any effect on the recurrence rate of the bladder tumors. Also as mentioned before, all patients had superficial bladder tumors when they entered the study and all these tumors were removed at the beginning. During the follow-up, the bladder tumors that were detected at each clinical visit were also removed. Of the 85 study subjects, there are 22 patients died before the end of the follow-up. For each patient, two covariates were measured and they are the number of initial tumors that the patients had before entering the study and the size of the largest initial tumor. The second covariate has been shown to have no effect on the recurrence rate of bladder tumors (Sun & Wei, 2000; Ghosh & Lin, 2002). Thus in the following, we will only consider the number of initial tumors.

For the analysis, define Z_1 to be the treatment indicator with $Z_1 = 1$ for the subjects in the thiotepa group and $Z_1 = 0$ otherwise and Z_2 the number of initial tumors. Then β_1 and β_2 will represent the effects of the thiotepa treatment and the number of initial tumors on the recurrence process of bladder tumors, respectively, while α gives the effect of the observation or visit process on the recurrent event process. Table 2.3

gives the obtained results, including the estimated effects, the 95% confidence intervals and the p-values for testing the estimated parameter to be zero. Here we considered three link functions for $g(t)$ and took $\tau = 53$ months, the longest observation time, and $h(\mathcal{F}_{i,t}) = N_i(t-)$, assuming that the bladder tumor recurrent process depends on the total number of observations or visits. It can be seen from the table that the results suggest that both the thiotepa treatment and the initial number of tumors had significant effects on the recurrence rate of the bladder tumor. In particular, the thiotepa treatment seems to significantly reduce the recurrence of bladder tumors. These results are similar to those given by other authors.

With respect to the relationship between the recurrence process of bladder tumors and the visit process, it seems that the total number of visits had no significant effect on the recurrence rate of bladder tumors. This differs from the result given in Li, Sun, & Sun (2010), which did not consider the terminal event death. One possible explanation for this is that the relationship detected in Li, Sun, & Sun (2010) may be due to the correlation between the bladder tumor occurrence process and the death. Instead of taking $h(\mathcal{F}_{i,t}) = N_i(t-)$, we also performed the analysis by letting $h(\mathcal{F}_{it}) = N_i(t-) - N_i(t-6)$, where $t_{i,j-1} < t \leq t_{i,j}$, meaning that the tumor recurrence process depends on the number of visits during the last six months. The obtained results are given in Table 2.4 and they gave similar conclusions as those in Table 2.3.

To assess the adequacy of model (2.2), we apply the goodness-of-fit procedure presented in Section 2.4 to the bladder tumor panel count data. Specifically, we calculated the statistic $\mathcal{F}(t, x)$ and obtained the p-value by comparing it to 1000 realizations of the statistic $\sup_{1<t\leq\tau,x} |\hat{\mathcal{F}}(t, x)|$. For the analysis with $h(\mathcal{F}_{it}) = N_i(t-)$, the p-values are 0.866 , 0.857 and 0.594 under the link functions $g(t) = t$, $g(t) = t^2$, and $g(t) = \log t$,

respectively. When taking $h(\mathcal{F}_{it}) = N_i(t-) - N_i(t-6)$, we obtained the p-values of 0.584, 0.6 and 0.454 for the same three link functions, respectively. These results indicate that model (2.2) seems to fit the data well.

2.7 Discussion and Concluding Remarks

This chapter considered regression analysis of panel count data in the presence of a terminal event. For the problem, a semiparametric transformation model was proposed, which can be seen as a generalization of the model studied by Li, Sun, & Sun (2010). For estimation of unknown parameters, the estimating equation approach and the inverse survival probability weighting technique were used and we established both finite and asymptotic properties of the resulting estimators. The simulation results indicate that the proposed approach works well for practical situations. In addition, we presented a goodness-of-fit test procedure for assessing the adequacy of the transformation model for the underlying recurrent event process of interest.

One of the focus of this chapter has been to take into account the dependent terminal event in the analysis of panel count data. It is worth noting that the models proposed may be of more clinical interest to some extent because it directly accounts for the covariate effects on the frequency of recurrent events among survivors without modeling the recurrent event process after the terminal events or the correlation between the rates of recurrent and terminal events. By models (2.2) and (2.3), the proposed procedure examined the effects of covariates on both the survival probability of the terminal event and the recurrent event rate among surviving subjects. In practice, if a treatment reduces the disease recurrence and death simultaneously or reduces

the disease recurrence but has no significant impact on survival, then the treatment is clearly preferred. However, if the treatment reduces the disease recurrence but increases mortality, then it is more subtle to make a judgment on the treatment and need to do further analysis.

To implement the proposed estimation procedure, one needs to choose the link function g, which determines the pattern of the underlying recurrent event process or the relationship between the recurrent event process and the covariate process. For this, although one can develop some procedures for selecting or estimating g, it is generally quite difficult as commented above. Of course, an alternative is to apply the goodness-of-fit test procedure given in Section 2.4. The same is true about the link function h, which represents the relationship between the underlying recurrent event process and the observation process. For modeling the terminal event, we used the proportional hazards model and in some situations, one may prefer some other models such as the additive hazards model, the accelerated failure time model and the linear transformation model, depending on the situation.

Chapter 3

SEMIPARAMETRIC ANALYSIS OF MULTIVARIATE PANEL COUNT DATA WITH A TERMINAL EVENT

3.1 Introduction

This chapter discusses regression analysis of panel count data for practical situations similar as in Chapter 2. We will now focus on multivariate panel count data in the presence of some terminal events (He et al., 2008; Zhao et al., 2011). Furthermore, both the observation process and the terminal event may be correlated with the recurrent event process of interest. As introduced by Chapter 1, panel count data arise in recurrent event studies when study subjects can be observed only at discrete time points instead of continuously. This is often the case in, for example, cohort studies, epidemiological studies, reliability studies and tumorigenicity experiments. In this

situation, the data take the form of counts of the cumulative numbers of the events of interest at observation time points along with explanatory covariates.

As discussed in Chapter 2, for regression analysis of panel count data, two complicated issues often arise. One is that the observation process that characterizes observation times on study subjects could be related to the underlying recurrent event process of interest even given covariates. The other is that there sometimes exists a terminal event such as death that stops the follow-up of the recurrent eventof interest or study subjects. More importantly, it is often the case that the terminal event is correlated with the recurrent event of interest. An example of this is that in a medical study, a patient may have an increasing rate of death when the rate of some disease-related recurrent event is unusually high. Note that terminal events are quite different from the ordinary censoring. This is because when a terminal event occurs, the recurrent event will be stopped permanently, while with censoring, the recurrent event may still occur continuously, just cannot be observed.

Many authors have considered regression analysis of univariate panel count data and these include Cheng and Wei (2000), Sun and Wei (2000), Wellner and Zhang (2007) and Zhang (2002) as mentioned in Section 1.2.2. However, the methods given by them assume that the underlying recurrent event process of interest and the observation process are independent completely or given covariates. Among others, He et al. (2009), Huang et al. (2006), Sun et al. (2007), and Li et al. (2010) studied the situation where the two processes may be correlated and proposed some approaches that directly model the relationship between the two processes. There also exist some procedures for regression analysis of multivariate panel count data (He et al., 2008; Li et al., 2011). But it does not seem to exist an established procedure for regression

analysis of multivariate panel count data with both dependent observation processes and a terminal event.

In the presence of terminal events, there exists considerable work on regression analysis of recurrent event data and in this case, two approaches are generally adopted. One is the marginal model approach, which focuses on the marginal rates of the recurrent and terminal events and does not specify the correlation between them (Cook and Lawless, 1997; Ghosh and Lin, 2002; Ye et al., 2007; Zhao et al., 2011). The other is the frailty model approach, which employs some latent variables to account for the correlation between the rates of recurrent and terminal events and assumes that these two event processes are independent given the frailty (Huang and Wang, 2004; Liu et al., 2004; Ye et al., 2007; Zeng and Cai, 2010). Similar approaches have been used for univariate panel count data, but not for multivariate panel count data. For the latter, an additional difficult issue is how to deal with the relationship among different types of recurrent events.

In this chapter, we propose a semiparametric marginal modeling approach for regression analysis of multivariate panel count data with dependent observation processes and a terminal event. In the approach, the additive model is employed for the mean functions of the underlying recurrent event processes and one advantage of such models is that they allow one to directly estimate the absolute difference between the rates of recurrent events. The proposed models are given in Section 3.2 along with some assumptions and leave both the correlation between the recurrent events and the terminal events and the correlation between different types of recurrent events unspecified. Section 3.3 presents an estimating equation-based procedure for estimation of regression parameters and the asymptotic properties of the resulting estimates are established. In

the procedure, the inverse probability weighting technique is used to take into account the facts that subjects who die cannot experience further observations or occurrence of the recurrent events of interests and that the whole observation process is informative. In addition, a model checking procedure is also given. An extensive simulation study is conducted in Section 3.4 and suggests that the proposed method works well for practical situations.

3.2 Models and Assumptions

Consider a recurrent event study that involves K different types of recurrent events. For each k ($k = 1, \ldots, K$), let $Y_k(t)$ denote the recurrent event process indicating the total number of occurrences of the kth type recurrent events of interest over the time interval $[0, t]$. Also let $Z(t)$ denote a vector of external covariate process (Kalbfleisch and Prentice, 2002) and define $\mathcal{Z}(t) = \{Z(s), 0 \leq s \leq t\}$, the history of covariates up to time t. In the following, we assume that each $Y_k(t)$ is observed only at discrete time points and will use the counting process $N_k(t)$ to denote the total number of observations up to time t. That is, $Y_k(t)$ can be observed only at the time points where $N_k(t)$ jumps.

Define $\mathcal{F}_{kt} = \{N_k(s), 0 \leq s < t\}$, the history or filtration of the observation process N_k up to time t^-. Assume that there exists a terminal event whose occurrence time is denoted by D and which may be correlated with the recurrent events of interest. In the following, to take into account the fact that the recurrent events will not occur further after the terminal event, we will focus on $Y_k^*(t) = Y_k(t \wedge D)$ and $N_k^*(t) = N_k(t \wedge D)$, the actual and adjusted recurrent event and observation processes. Note that both $N_k^*(t)$

and $Y_k^*(t)$ will remain constants after D.

To model the effects of covariates, we will assume that given $\mathcal{Z}(t)$, $\mathcal{F}_{kt}$ and D, the conditional mean function of $Y_k^*(t)$ has the form

$$E\{Y_k^*(t)|\mathcal{Z}(t), \mathcal{F}_{kt}, D \geq t\} = \mu_{0k}(t) + \beta' Z(t) + \alpha' h_k(\mathcal{F}_{kt}), \tag{3.1}$$

$k = 1, ..., K$. In the above, α and β are vectors of unknown regression parameters, $\mu_{0k}(t)$ is an unspecified baseline cumulative mean function, and $h_k(\cdot)$ is a vector of known functions of $\mathcal{F}_{kt}$. Furthermore, it is assumed that $\mu_{0k}(0) = h_k(\mathcal{F}_{k0}) = 0$ and $\mu_{0k}(t)$ is an increasing function of t for $t \leq D$. For the adjusted observation process $N_k^*(t)$, it will be assumed that it is a non-homogeneous Poisson process satisfying the following marginal rate model

$$E\{dN_k^*(t)|\mathcal{Z}(t)\} = e^{\gamma' Z(t)} d\Lambda_{0k}(t), \tag{3.2}$$

where γ is a vector of unknown regression parameters and $d\Lambda_{0k}(\cdot)$ is an unspecified baseline rate function.

Note that in both models (3.1) and (3.2), for the simplicity of presentation, we assume that regression parameters α, β and γ are the same for different k. It is straightforward to generalize the methodology proposed below to situations that they may differ for different k. Model (3.1) can be seen as a generalization of model (2) of Li et al. (2011) for $E\{Y_k^*(t)|\mathcal{Z}(t), \mathcal{F}_{kt}\}$ since if $D = +\infty$ or there is no terminal event, $E\{Y_k^*(t)|\mathcal{Z}(t), \mathcal{F}_{kt}, D \geq t\}$ reduces to $E\{Y_k^*(t)|\mathcal{Z}(t), \mathcal{F}_{kt}\}$. Define $S(t|Z) = P(D \geq$

$t|\mathcal{Z}(t))$. One can easily show that

$$E\{Y_k^*(t)|\mathcal{Z}(t), \mathcal{F}_{kt}\} = \int_0^t S(u|Z) E\{dY_k^*(u)|\mathcal{Z}(u), \mathcal{F}_{ku}, D \geq u\}\,.$$

This yields that

$$E\{Y_k^*(t)|\mathcal{Z}(t), \mathcal{F}_{kt}, D \geq t\} \;>\; E\{Y_k^*(t)|\mathcal{Z}(t), \mathcal{F}_{kt}\}$$

for t greater than the first observed terminal event time. In the following, we will assume that given $\mathcal{Z}(t)$, $\mathcal{F}_{kt}$ and D, the adjusted recurrent event process $Y_k^*(t)$ and the adjusted observation process $N_k^*(t)$ are independent.

In practice, covariates may have effects on terminal events too. For this, we will assume that the occurrence time D of the terminal event follows the proportional hazards model given by

$$\lambda^d(t|Z(t)) = e^{\delta' Z(t)} \lambda_0^d(t), \tag{3.3}$$

where $\lambda_0^d(t)$ is an unspecified baseline hazard function and δ is a vector of unknown regression parameters. Define $\Delta_0(t) = \int_0^t \lambda_0^d(s)ds$. Then we have

$$S(t|Z) = \exp\{-\int_0^t \exp\{\delta_0' Z(s)\} d\Delta_0(s)\}.$$

We remark that instead of modeling the adjusted rate or mean functions as in models (3.1) and (3.2), an alternative is to directly model the original recurrent event processes of interest and observation processes. An advantage of models (3.1) and (3.2) is that no assumption is needed for the recurrent event process after the terminal event (Luo et al.,2010). Among others, Cook and Lawless (1997) and Ghosh and Lin (2002)

discussed similar models for regression analysis of recurrent event data. An drawback of model (3.1) is that one cannot directly estimate the overall covariates effects.

3.3 Inference Procedures

Let $Y_k(t)$, $N_k(t)$, $Y_k^*(t)$, $N_k^*(t)$ and D be defined as in the previous section. In practice, in addition to the terminal event D, there may also exist a censoring time denoted by C. The actual follow-up time is then $T^* = C \wedge D$. In the following, for simplicity and as with D, it will be assumed that C is the same for all K types of recurrent events. Also we assume that C is independent of $\{N_k^*(t), Y_k^*(t), D\}$ conditional on $\mathcal{Z}(t)$. Define $\widetilde{Y}_k(t) = Y_k^*(t \wedge C)$ and $\widetilde{N}_k(t) = N_k^*(t \wedge C)$. Then for a study consisting of n independent subjects, the observed data have the form

$$\{\widetilde{N}_{ik}(t), \tilde{Y}_{ik}(t)\, d\tilde{N}_{ik}(t), Z_i(t), T_i^*, I(D_i \le C_i)\,;\ 0 \le t \le T_i^*,\ i = 1, \ldots, n;\ k = 1, \ldots, K\}\,.$$

To present the estimation procedure, define $X_{ik}(t) = (Z_i(t)', h_k(\mathcal{F}_{ikt})')'$ and $\theta = (\beta', \alpha')'$. Note that under models (3.1) and (3.2) and based on the conditional independent assumptions for $Y_{ik}^*(t)$, $N_{ik}^*(t)$ and C_i, one can show that

$$E\{\widetilde{Y}_{ik}(t) d\widetilde{N}_{ik}(t)\} = E\{I(C_i \ge t)\{\mu_{0k}(t) + \theta' X_{ik}(t)\} e^{\gamma' Z_i(t)} d\Lambda_{0k}(t)\}. \tag{3.4}$$

This naturally suggests the estimating equation

$$\sum_{i=1}^{n} \left[\widetilde{Y}_{ik}(t) d\widetilde{N}_{ik}(t) - I(C_i \ge t)\{\mu_{0k}(t) + \theta' X_{ik}(t)\} e^{\gamma' Z_i(t)} d\Lambda_{0k}(t)\right] = 0,\ \ 0 \le t \le \tau,$$

where τ is the longest follow-up time. The derivation of (3.4) will be given in Appendix B.1. On the other hand, in practice, C_i and thus $I(C_i \geq t)$ are unobservable when $D_i \leq C_i$ and thus the equation given above is not applicable. To deal with this, define $\omega_i(t) = I(T_i^* \geq t)/S(t|Z_i)$ and one can show that $E\{\omega_i(t)|\mathcal{Z}_i(t)\} = E\{I(C_i \geq t)|\mathcal{Z}_i(t)\}$ since $E\{I(T_i^* \geq t)|\mathcal{Z}_i(t)\} = E\{I(C_i \geq t)|\mathcal{Z}_i(t)\}\, S(t|Z_i)$. By employing the inverse probability weighting technique, this suggests that we can replace $I(C_i \geq t)$ $(i = 1, \ldots, n)$ by $\omega_i(t)$ and consider the zero-mean stochastic process

$$dM_{ik}(t;\theta,\gamma) = \widetilde{Y}_{ik}(t)d\widetilde{N}_{ik}(t) - \omega_i(t)\{\mu_{0k}(t) + \theta' X_{ik}(t)\}e^{\gamma' Z_i(t)}d\Lambda_{0k}(t)$$

for the construction of estimating equations.

To use $M_{ik}(t;\theta,\gamma)$, we need to estimate $\omega_i(t)$. It is apparent that a natural estimate is given by $\hat{\omega}_i(t) = I(T_i^* \geq t)/\hat{S}(t|Z_i)$, where $\hat{S}(t|Z_i) = \exp\big\{-\int_0^t \exp\{\hat{\delta}' Z_i(s)\}d\hat{\Delta}_0(s)\big\}$ with $\hat{\delta}$ and $\hat{\Delta}_0(t)$ denoting the maximum partial likelihood estimate of δ and the Breslow estimate of $\Delta_0(t)$, respectively, under model (3.3). By using the same arguments as those used in Lin et al. (2001) and Sun et al. (2005), one can show that for large n, $\hat{\omega}_i(t)$ always exists and is unique and consistent. Let $M_{ik}^*(t;\theta,\gamma)$ denote $M_{ik}(t;\theta,\gamma)$ with $\omega_i(t)$ replaced by $\hat{\omega}_i(t)$. Then if γ and Λ_{0k} are known, it is natural to estimate $\mu_{0k}(t)$ and θ by the following estimating equations

$$\sum_{i=1}^{n} dM_{ik}^*(t;\theta,\gamma) \;=\; 0, \;\; 0 \leq t \leq \tau, \tag{3.5}$$

and

$$U_\theta(\theta;\gamma) \;=\; \sum_{i=1}^{n}\sum_{k=1}^{K} \int_0^\tau W(t)X_{ik}(t)dM_{ik}^*(t;\theta,\gamma) \;=\; 0\,, \tag{3.6}$$

where $W(t)$ is a possibly data-dependent weight function.

Of course, in reality, γ and Λ_{0k} are unknown. On the other hand, they can be readily estimated based on the recurrent event data on the $N_{ik}^*(t)$'s. Specifically, define

$$S^{(j)}(t;\gamma) = n^{-1}\sum_{i=1}^{n}\hat{\omega}_i(t)Z_i(t)^j e^{\gamma' Z_i(t)}\,,\ j=0,1\,,$$

and $\bar{Z}(t;\gamma) = S^{(1)}(t;\gamma)/S^{(0)}(t;\gamma)$. Then the consistent estimates, denoted by $\hat{\gamma}$ and $\hat{\Lambda}_{0k}(t)$, of γ and $\Lambda_{0k}(t)$ can be obtained by solving the following two estimating equations

$$U_\gamma(\gamma) = \sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^\tau \big\{Z_i(t) - \bar{Z}(t;\gamma)\big\}d\widetilde{N}_{ik}(t) = 0 \tag{3.7}$$

and

$$\sum_{i=1}^{n}\Big[d\widetilde{N}_{ik}(t) - \hat{\omega}_i(t)e^{\gamma' Z_i(t)}d\Lambda_{0k}(t)\Big] = 0\,.$$

In particular, given γ, $\hat{\Lambda}_{0k}(t)$ has the closed form

$$\hat{\Lambda}_{0k}(t;\gamma) = \int_0^t \frac{d\bar{N}_k(u)}{S^{(0)}(u;\gamma)}, \tag{3.8}$$

where $\bar{N}_k(t) = n^{-1}\sum_{i=1}^{n}\widetilde{N}_{ik}(t)$.

Let $\hat{\theta}$ and $\hat{\mu}_{0k}$ denote the estimates of θ and $\mu_{0k}(t)$ given by equations (3.5) and (3.6) with all unknowns replaced by their estimates. Then one can easily show that

$$\hat{\theta} = \frac{1}{n}\hat{A}_\theta^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^\tau W(t)\big\{X_{ik}(t) - \bar{X}_k(t;\hat{\gamma})\big\}\tilde{Y}_{ik}(t)d\tilde{N}_{ik}(t)$$

and

$$\hat{\mu}_{0k}(t;\hat{\theta},\hat{\gamma}) = \frac{\sum_{i=1}^{n}\tilde{Y}_{ik}(t)d\tilde{N}_{ik}(t)}{\sum_{i=1}^{n}\hat{\omega}_i(t)e^{\hat{\gamma}' Z_i(t)}d\Lambda_{0k}(t)} - \hat{\theta}'\bar{X}_k(t;\hat{\gamma}),$$

where

$$\hat{A}_\theta = n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^\tau W(t)\big\{X_{ik}(t) - \bar{X}_k(t;\hat{\gamma})\big\}^{\otimes 2}\hat{\omega}_i(t)e^{\hat{\gamma}' Z_i(t)}d\hat{\Lambda}_{0k}(t,\hat{\gamma})$$

and

$$\bar{X}_k(t;\hat{\gamma}) = \frac{\sum_{i=1}^n X_{ik}(t)\hat{\omega}_i(t)e^{\hat{\gamma}' Z_i(t)}}{\sum_{i=1}^n \hat{\omega}_i(t)e^{\hat{\gamma}' Z_i(t)}}\,.$$

That is, they have closed forms, which makes their determination quite easy.

To present the asymptotic properties of $\hat{\theta}$, define $N_i^d(t) = I(D_i \le t, D_i \le C_i)$,

$$M_i^d(t) = N_i^d(t) - \int_0^t I(T_i^* \ge s)e^{\delta_0' Z_i(s)}d\Delta_0(s)\,,$$

$$\hat{M}_i^d(t) = N_i^d(t) - \int_0^t I(T_i^* \ge s)e^{\hat{\delta}' Z_i(s)}d\hat{\Delta}_0(s),$$

$$\hat{M}_{ik}^*(t) = \widetilde{N}_{ik}(t) - \int_0^t \hat{\omega}_i(s)e^{\hat{\gamma}' Z_i(s)}d\hat{\Lambda}_{0k}(s;\hat{\gamma}),$$

and

$$\hat{M}_{ik}(t) = \int_0^t \widetilde{Y}_{ik}(s)d\widetilde{N}_{ik}(s) - \int_0^t \hat{\omega}_i(s)\{\hat{\mu}_{0k}(s;\hat{\theta},\hat{\gamma}) + \hat{\theta}' X_{ik}(s)\}e^{\hat{\gamma}' Z_i(s)}d\hat{\Lambda}_{0k}(s;\hat{\gamma}).$$

Also define

$$\hat{\Upsilon}_k(t;\theta,\gamma) = n^{-1}\sum_{i=1}^{n}\big\{X_{ik}(t) - \bar{X}_k(t;\gamma)\big\}\hat{\omega}_i(t)\big\{\hat{\mu}_{0k}(t;\theta,\gamma) + \theta' X_{ik}(t)\big\}e^{\gamma' Z_i(t)},$$

$$R^{(j)}(t;\delta) = n^{-1}\sum_{i=1}^{n} I(T_i^* \ge t)Z_i(t)^{\otimes j}e^{\delta' Z_i(t)},\ j = 0,1,2,$$

$$\hat{H}(t;Z_i)=\int_0^t e^{\hat{\delta}'Z_i(u)}\Big\{Z_i(u)-\frac{R^{(1)}(t;\hat{\delta})}{R^{(0)}(t;\hat{\delta})}\Big\}d\hat{\Delta}_0(u;\hat{\delta}),$$

and

$$\hat{\Omega}_\delta=n^{-1}\sum_{i=1}^n\int_0^\tau\left[\frac{R^{(2)}(t;\hat{\delta})}{R^{(0)}(t;\hat{\delta})}-\Big\{\frac{R^{(1)}(t;\hat{\delta})}{R^{(0)}(t;\hat{\delta})}\Big\}^{\otimes 2}\right]dN_i^d(t)\,.$$

In the above, $r^{(j)}(t)=\lim_{n\to\infty}R^{(j)}(t;\delta_0)$ with $j=0,1,2$ and $v^{\otimes 2}=vv'$ for a vector v. Let θ_0 denote the true value of θ and $s^{(0)}(t)$, $s^{(1)}(t)$, $\bar{x}_k(t)$, $\Upsilon_k(t)$ and Ω_δ the limits of $S^{(0)}(t;\gamma_0)$, $S^{(1)}(t;\gamma_0)$, $\bar{X}_k(t;\gamma_0)$, $\hat{\Upsilon}_k(t;\theta_0,\gamma_0)$ and $\hat{\Omega}_\delta$, respectively. Define $\bar{z}(t)=s^{(1)}(t)/s^{(0)}(t)$. The following theorem gives the consistency and asymptotical normality of $\hat{\theta}$.

Theorem 3.1. *Assume that the conditions (C1)-(C5) described in Appendix B.2 hold. Then $\hat{\theta}$ is a consistent estimator of θ_0 and $n^{1/2}(\hat{\theta}-\theta_0)$ converges in distribution to a zero-mean normal random vector whose covariance matrix can be consistently estimated by $\hat{A}_\theta^{-1}\hat{\Sigma}\hat{A}_\theta^{-1}$, where $\hat{\Sigma}=n^{-1}\sum_{i=1}^n(\hat{\xi}_{1i}-\hat{\xi}_{2i}-\hat{\xi}_{3i})^{\otimes 2}$,*

$$\hat{\xi}_{1i}=\sum_{k=1}^K\int_0^\tau W(t)\{X_{ik}(t)-\bar{X}_k(t;\hat{\gamma})\}d\hat{M}_{ik}(t),$$

$$\hat{\xi}_{2i}=\sum_{k=1}^K\int_0^\tau\left[\frac{W(t)\hat{\Upsilon}_k(t;\hat{\theta},\hat{\gamma})}{S^{(0)}(t;\hat{\gamma})}+\hat{A}_\gamma\hat{\Omega}_\gamma^{-1}\{Z_i(t)-\bar{Z}(t;\hat{\gamma})\}\right]d\hat{M}_{ik}^*(t),$$

$$\hat{\xi}_{3i}=\sum_{k=1}^K\int_0^\tau\left[\hat{A}_\gamma\hat{\Omega}_\gamma^{-1}\hat{Q}_{1k}\Big\{Z_i(t)-\frac{R^{(1)}(t;\hat{\delta})}{R^{(0)}(t;\hat{\delta})}\Big\}+\hat{A}_\gamma\hat{\Omega}_\gamma^{-1}\frac{\hat{Q}_{2k}(t)}{R^{(0)}(t;\hat{\delta})}+\frac{\hat{B}_{1k}(t)}{R^{(0)}(t;\hat{\delta})}\right.$$
$$\left.+\hat{B}_{2k}\Big\{Z_i(t)-\frac{R^{(1)}(t;\hat{\delta})}{R^{(0)}(t;\hat{\delta})}\Big\}\right]d\hat{M}_i^d(t),$$

$$\hat{A}_\gamma=n^{-1}\sum_{i=1}^n\sum_{k=1}^K\int_0^\tau W(t)\hat{\omega}_i(t)e^{\hat{\gamma}'Z_i(t)}\{\hat{\mu}_{0k}(t;\hat{\theta},\hat{\gamma})+\hat{\theta}'X_{ik}(t)\}\{X_{ik}(t)-\bar{X}_k(t;\hat{\gamma})\}$$

$$\times\{Z_i(t) - \bar{Z}(t;\hat{\gamma})\}' d\hat{\Lambda}_{0k}(t;\hat{\gamma}),$$

$$\hat{\Omega}_\gamma = n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^\tau \{Z_i(t) - \bar{Z}(t;\hat{\gamma})\}^{\otimes 2} d\tilde{N}_{ik}(t),$$

$$\hat{B}_{1k}(t) = n^{-1}\sum_{i=1}^{n} e^{\hat{\delta}' Z_i(t)} \int_0^\tau I(t<s)\hat{B}^*_{ik}(s) d\hat{\Lambda}_{0k}(s;\hat{\gamma}),$$

$$\hat{B}_{2k} = n^{-1}\sum_{i=1}^{n}\int_0^\tau \hat{B}^*_{ik}(t)\hat{H}(t;Z_i)'\hat{\Omega}_\delta^{-1} d\hat{\Lambda}_{0k}(t;\hat{\gamma}),$$

$$\hat{B}^*_{ik}(t) = W(t)\hat{\omega}_i(t) e^{\hat{\gamma}' Z_i(t)}\Big[\big\{X_{ik}(t) - \bar{X}_k(t;\hat{\gamma})\big\}\big\{\hat{\mu}_{0k}(t;\hat{\theta},\hat{\gamma}) + \hat{\theta}' X_{ik}(t)\big\} - \frac{\hat{\Upsilon}_k(t;\hat{\theta},\hat{\gamma})}{S^{(0)}(t;\hat{\gamma})}\Big],$$

$$\hat{Q}_{1k} = n^{-1}\sum_{i=1}^{n}\int_0^\tau \big\{Z_i(t) - \bar{Z}(t;\hat{\gamma})\big\}\hat{Q}_3(t;Z_i)'\hat{\Omega}_\delta^{-1} d\hat{M}^*_{ik}(t),$$

$$\hat{Q}_{2k}(t) = n^{-1}\sum_{i=1}^{n} e^{\hat{\delta}' Z_i(t)}\int_0^\tau \big\{Z_i(u) - \bar{Z}(u;\hat{\gamma})\big\} I(u \geq t) d\hat{M}^*_{ik}(u),$$

and

$$\hat{Q}_3(t;Z_i) = \int_0^t \big\{Z_i(u) - \bar{Z}(u;\hat{\gamma})\big\} e^{\hat{\delta}' Z_i(u)} d\hat{\Delta}_0(u;\hat{\delta})\,.$$

For a given data set, one question of practical interest is to assess the adequacy of the models described in Section 3.2. Note that for both models (3.2) and (3.3), one observes complete data and several procedures have been developed in the literature for checking their goodness-of-fits (Lin, Wei, & Ying, 1993; Lin, et al., 2000; Schoenfeld, 1982). So in the following, we will focus on model (3.1) and develop an omnibus goodness-of-fit procedure.

Let the $\widehat{M}_{ik}(t)$'s be defined as above. Note that they represent the differences between the observed and model-predicted numbers of the kth type events by time t.

Thus it is natural to construct a test statistic based on them. Now we consider the following cumulative sums of residual process

$$\mathcal{F}(t,x) = n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^t I(X_{ik}(u)\le x)d\hat{M}_{ik}(u),$$

where the event $\{X_{ik}(u)\le x\}$ means that each components of $X_{ik}(u)$ is not greater than the respective component of x. We will show in Appendix B.3 that the null distribution of $\mathcal{F}(t,x)$ can be approximated by a zero-mean Gaussian process

$$\widehat{\mathcal{F}}(t,x) = n^{-1/2}\sum_{i=1}^{n}\Big\{\hat{\eta}_{1i}(t,x) - \hat{\Phi}_\gamma(t,x)\Omega_\gamma^{-1}\hat{\eta}_{2i} - \hat{\Phi}_\theta(t,x)\hat{A}_\theta^{-1}\hat{\eta}_{3i}\Big\}G_i\,. \tag{3.9}$$

In the above, $G_1,\ldots,G_n$ are independent standard normal variables independent of the observed data,

$$\hat{\eta}_{1i}(t,x) = \sum_{k=1}^{K}\int_0^t\{I(X_{ik}(u)\le x) - \hat{E}_k(u,x)\}d\hat{M}_{ik}(u) - \int_0^t\frac{\tilde{\Upsilon}_k(u,x)}{S^{(0)}(u;\hat{\gamma})}d\hat{M}^*_{ik}(u)$$

$$-\int_0^t\left\{\frac{\tilde{B}_{1k}(u,t,x)}{R^{(0)}(u;\hat{\delta})} + \tilde{B}_{2k}(t,x)\Big(Z_i(u) - \frac{R^{(1)}(u;\hat{\delta})}{R^{(0)}(u;\hat{\delta})}\Big)\right\}d\hat{M}^d_i(u)\,,$$

$$\hat{\eta}_{2i} = \sum_{k=1}^{K}\int_0^\tau\Big[\hat{Q}_{1k}\Big\{Z_i(t) - \frac{R^{(1)}(t;\hat{\delta})}{R^{(0)}(t;\hat{\delta})}\Big\} + \frac{\hat{Q}_{2k}(t)}{R^{(0)}(t;\hat{\delta})}\Big]d\hat{M}^d_i(t) + \int_0^\tau\big\{Z_i(t) - \bar{Z}(t;\hat{\gamma})\big\}d\hat{M}^*_{ik}(t)\,,$$

$$\hat{\eta}_{3i} = \sum_{k=1}^{K}\hat{\xi}_{1i} - \hat{\xi}_{2i} - \hat{\xi}_{3i},$$

$$\tilde{\Upsilon}_k(s,x) = n^{-1}\sum_{i=1}^{n}\big\{I(X_{ik}(s)\le x) - \hat{E}_k(s,x)\big\}\hat{\omega}_i(s)\{\hat{\mu}_{0k}(s) + \hat{\theta}'X_{ik}(s)\}e^{\hat{\gamma}'Z_i(s)},$$

$$\tilde{B}_{1k}(u,t,x) = n^{-1}\sum_{i=1}^{n}e^{\hat{\delta}'Z_i(u)}\int_0^t I(u<s)\tilde{B}^*_{ik}(s,x)d\hat{\Lambda}_{0k}(s),$$

$$\tilde{B}_{2k}(t,x) = n^{-1}\sum_{i=1}^{n}\int_0^t \tilde{B}^*_{ik}(s,x)\hat{H}(s;Z_i)'\hat{\Omega}_\delta^{-1}d\hat{\Lambda}_{0k}(s),$$

$$\tilde{B}^*_{ik}(s,x) = \hat{\omega}_i(s)e^{\hat{\gamma}'Z_i(s)}\Big[\big\{I(X_{ik}(s)\le x) - \hat{E}_k(s,x)\big\}\{\hat{\mu}_{0k}(s) + \hat{\theta}'X_{ik}(s)\} - \frac{\tilde{\Upsilon}_k(s;x)}{S^{(0)}(s;\hat{\gamma})}\Big],$$

$$\hat{\Phi}_\gamma(t,x) = n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^t \big[I(X_{ik}(u)\le x) - \hat{E}_k(u,x)\big]\big[Z_i(u) - \bar{Z}(u;\hat{\gamma})\big]'\hat{\omega}_i(u)$$

$$\times\{\hat{\mu}_{0k}(u) + \hat{\theta}'X_{ik}(u)\}e^{\hat{\gamma}'Z_i(u)}d\hat{\Lambda}_{0k}(u;\hat{\gamma}),$$

$$\hat{\Phi}_\theta(t,x) = n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^t I(X_{ik}(u)\le x)\big\{X_{ik}(u) - \bar{X}_k(t;\hat{\gamma})\big\}'\hat{\omega}_i(u)e^{\hat{\gamma}'Z_i(u)}d\hat{\Lambda}_{0k}(u,\hat{\gamma}),$$

and

$$\hat{E}_k(u,x) = \frac{\sum_{i=1}^n I(X_{ik}(u)\le x)\hat{\omega}_i(u)e^{\hat{\gamma}'Z_i(u)}}{\sum_{i=1}^n \hat{\omega}_i(u)e^{\hat{\gamma}'Z_i(u)}}.$$

Based on Equation (3.9), it is easy to see that one could obtain a large number of realizations from $\widehat{\mathcal{F}}(t,x)$ by repeatedly generating the standard normal random sample $\{G_1,\ldots,G_n\}$ while fixing the observation data. Thus to check the validity of model (3.2), one can plot these realizations of $\widehat{\mathcal{F}}(t,x)$ along with the observed $\mathcal{F}(t,x)$ and examine any unusual pattern of $\mathcal{F}(t,x)$ compared to the realizations. Furthermore, a formal lack-of-fit test can be constructed based on the statistic $\sup_{0\le t\le\tau,x}|\mathcal{F}(t,x)|$ and the corresponding p-value can be obtained by comparing the observed value of $\sup_{0\le t\le\tau,x}|\mathcal{F}(t,x)|$ to a large number of realizations from $\sup_{0\le t\le\tau,x}|\widehat{\mathcal{F}}(t,x)|$.

3.4 A Numerical Study

To examine the finite-sample behavior of the estimation procedure proposed in the previous section, an extensive simulation study was conducted. In the study,

we considered the situation of $K = 2$ and the covariate Z_i was assumed to follow the Bernoulli distribution with probability equal to 0.5. The censoring time C_i was generated from the uniform distribution $U(\tau/4, \tau)$ with τ being 1 as the follow-up time. A latent variable v_i was used to simulate the dependence between the recurrent event process and the terminal event, and v_i was assumed to follow the positive stable distribution with Laplace transformation $L(s) = E(e^{-sv_i}) = \exp(-s^\rho)$ (Luonga and Doray, 2009). We took $\rho = 0.7$ or $\rho = 1.0$. Given Z_i and v_i, we assumed the terminal event D_i has a hazard function $\lambda(t|Z_i, v_i) = 0.2v_i \exp\{0.5Z_i\}$. It could be easily shown that $S(t|Z_i) = \exp\{-\big(0.2te^{0.5Z_i}\big)^\rho\}$. For the observation process, it was assumed that given Z_i and T_i^*, $\tilde{N}_{ik}(t)$ was a nonhomogeneous Poisson process on $[0, \tau]$ with $E\{d\tilde{N}_{ik}(t)|Z_i, T_i^*\} = S^{-1}(t|Z_i)e^{0.5Z_i}d\Lambda_{0k}(t)I(T_i^* \geq t)$ for k=1,2. It also implies $E\{dN_{ik}^*(t)|Z\} = e^{0.5Z_i}d\Lambda_{0k}(t)$. Accordingly, the number of observations m_{ik} followed the Poisson distribution with mean $\int_0^{T_i^*} S^{-1}(t|Z_i)e^{0.5Z_i}d\Lambda_{0k}(t)$, and the observation times $(t_{ik,1}, \ldots, t_{ik,m_{ik}})$ were taken as the order statistics of a random sample of size m_{ik} from the uniform distribution over $(0, T_i^*)$. For the generation of panel counts $Y_{ik}^*(t_{ik,j})$, given m_{ik} and $(t_{ik,1}, \ldots, t_{ik,m_{ik}})$, we assumed that

$$Y_{ik}^*(t_{ik,j}) = Y_{ik}^{**}(t_{ik,1}) + Y_{ik}^{**}(t_{ik,2} - t_{ik,1}) + \cdots + Y_{ik}^{**}(t_{ik,j} - t_{ik,j-1})$$

with $t_{ik,0} = 0$, $j = 1, \ldots, m_{ik}$. In the above, given v_i, Z_i and $\mathcal{F}_{ik,j}$, it was assumed that $Y_{ik}^{**}(s)$ and $Y_{ik}^{**}(t-s)$ were the mixed Poisson distribution with the conditional mean functions

$$Q_i\phi(v_i, s)\big(\mu_{0k}(s) + \beta' Z_i + \alpha' h_k(\mathcal{F}_{ik,s})\big)$$

and

$$Q_i\phi(v_i,t)\big(\mu_{0k}(t)+\beta' Z_i+\alpha' h_k(\mathcal{F}_{ik,t})\big)-Q_i\phi(v_i,s)\big(\mu_{0k}(s)+\beta' Z_i+\alpha' h_k(\mathcal{F}_{ik,s})\big)\,,$$

respectively, where Q_i was sampled independently from a gamma distribution with mean 1 and variance 0.1, and

$$\phi(v_i,t)=e^{-v_i}\exp\{(1+0.2t\,e^{0.5Z_i})^{\rho}-(0.2t\,e^{0.5Z_i})^{\rho}\}\,.$$

One can show that $E\{Q_i\phi(v_i,t)|Z,\mathcal{F}_{ikt},D\geq t\}=1$. The results reported below are based on 500 replications and with the sample size $n=200$ or 300.

Table 3.1 presents the results obtained for estimation of β and α based on the simulated data with the true values of (β,α) being equal to $(0,0)$ $(0.5,0)$, $(0,0.1)$ or $(0.5,0.1)$, $\mu_{01}(t)=\mu_{02}(t)=t$, $\Lambda_{01}(t)=\Lambda_{02}(t)=10t$, $h_1(\mathcal{F}_{i1,t})=N_{i1}(t-)$, $h_2(\mathcal{F}_{i2,t})=N_{i2}(t-)$ and $W(t)=1$. The results include the estimated biases (BIAS) given by the averages of the estimators minus their true values, the sampling standard errors (SSE), the averages of the estimated standard errors (SEE), and the 95% empirical coverage probabilities (CP). They suggest that the proposed approach seems to perform well. Specifically, they indicate that the proposed estimators seem to be unbiased and there is a good agreement between the estimated and empirical standard errors. Also the coverage probabilities are reasonable and consistent with the nominal levels and as expected, the estimated standard errors became smaller as the sample size increased.

In addition to that discussed in Table 3.1, we investigated many other set-ups. For example, the results given in Table 3.2 were obtained under the same set-up as in Table 3.1 except that $\mu_{01}(t)=\sqrt{t}$, $\mu_{02}(t)=t$, $\Lambda_{01}(t)=8t$, $\Lambda_{02}(t)=12t$. Table 3.3 considered

the same setup as in Table 3.2 expect that $h_1(\mathcal{F}_{i1,t}) = N_{i1}(t-)$, $h_2(\mathcal{F}_{i2,t}) = N_{i2}(t-) - N_{i2}(t-0.5)$, while in Table 3.4, we employed $h_1(\mathcal{F}_{i1,t}) = N_{i1}(t-) - N_{i1}(t-0.75)$, $h_2(\mathcal{F}_{i2,t}) = N_{i2}(t-) - N_{i2}(t-0.75)$, $\mu_{01}(t) = \mu_{02}(t) = exp(t^2)$, $\Lambda_{01}(t) = \Lambda_{02}(t) = 8t$. It can be seen that all tables gave similar conclusions as those from Table 3.1.

Note that in the proposed methodology, we assume that the observation process $\tilde{N}_{ik}(t)$ is a non-homogeneous Poisson process and it is known that sometimes this may not hold. To investigate the robustness of the proposed estimation procedure to this assumption, we considered some situations where this Poisson assumption does not hold. For example, Table 3.5 presents the results obtained under the same set-up as in Table 3.2 except that $E\{dN^*_{ik}(t)|Z\} = Q'_i e^{0.5Z_i} d\Lambda_{0k}(t)$ with Q'_i generated from a gamma distribution with mean 1 and variance 0.01. One can see that they are similar to those given in Table 3.2. That is, the proposed procedure seems to still perform well.

3.5 Discussion and Concluding Remarks

Regression analysis of panel count data has been studied by many authors in the literature when there is no terminal event. However, in many medical longitudinal follow-up studies, there exist terminal events that stop permanently the further occurrence of the recurrent events of interest and make the analysis of panel count data more challenging. For the problem, we proposed an additive mean model and for estimation of regression parameters, the estimating equation approach and the inverse survival probability weighting technique were used. Both finite and asymptotic properties of the resulting estimators were established, and in addition, a lack-of-fit

test was also provided for assessing the adequacy of the model. Numerical results showed that the proposed procedures work well for practical situations. In the presence of a terminal event, the models proposed in this chapter are of more clinical interest to some extent because they directly account for the covariate effects on the frequency of recurrent events among survivals without modeling the recurrent event process after the terminal events or the correlation between the rates of recurrent and terminal events. In fact, the proposed estimation procedure is a joint analysis of the survival probability of the terminal event and the recurrent event rate among surviving subjects. This can be seen more clearly when the mean is expressed as $E\{Y^*(t)|\mathcal{Z}(t), \mathcal{F}_t\} = \int_0^t S(u|Z)E\{dY^*(u)|\mathcal{Z}(u), \mathcal{F}_u, D \geq u\}$. If a treatment reduces disease recurrences and death simultaneously or reduces disease recurrences but has no significant impact on survival, then the treatment is clearly preferred. However, if the treatment reduces disease recurrences but increases mortality, then it is more subtle to make a choice, and a marginal rate model is preferable.

One complication in the analysis of panel count data with terminal events is that the censoring time is not always observable. To deal with this, we applied the inverse probability of survival weighting technique that models the survival distribution. Instead one may use other weighting techniques too such as the inverse probability of censoring weighting (Ghosh and Lin, 2002). One advantage of using the survival weighting is that the survival distribution itself is usually of interest in clinical studies, but not censoring distribution. For the problem considered here, we have focused on the additive mean model, which has the advantage of giving direct estimation of absolute differences, the quantities often interested by clinicians. As alternatives, many other models could be used such as the multiplicative model or the semiparametric

transformation model.

Chapter 4

NONPARAMETRIC COMPARISON FOR PANEL COUNT DATA WITH UNEQUAL OBSERVATION PROCESSES

4.1 Introduction

In many medical studies producing panel count data, including the skin cancer study described in Section 1.1.2.3, treatment comparison is one of the most asked questions. The majority of existing test procedures assume identical observation processes across different treatment groups or involve the mean function estimators in their test statistics as discussed in Section 1.3. For example, Thall and Lachin (1988) suggested transforming the problem to a multivariate comparison one by grouping panel count data to multivariate data. Sun and Kalbfleisch (1993), Sun and Fang (2003) and Park, Sun and Zhao (2007) developed model-free approaches employing the isotonic

regression estimator (IRE) for the mean function. Zhang (2006) and Balakrishnan and Zhao (2011) used nonparametric maximum pseudo-likelihood estimator (NPMPLE) for multi-sample comparisons. Also Balakrishnan and Zhao (2009, 2010) employed the nonparametric maximum likelihood estimator (NPMLE) and proposed new classes of test statistcs. All the approaches above require an identical observation process across all study subjects, which may not be feasible in practice. For this, Zhao and Sun (2011) proposed a test procedure which allows for unequal observation processes. However, their test statistics also involved the estimation of the mean function and employed IRE.

Although the mean function estimators IRE, NPMPLE or NPMLE perform well in general, we noticed that they may be biased when the data are sparsely distributed. For example, in the skin cancer data described above, we noticed that the observed data are very sparsely distributed over all 1159 observation times made by 291 study subjects over the study. In this chapter, we propose a new class of nonparametric test procedures that allow different observation processes without employing the estimation of the mean function. The new test procedure is motivated by those used for recurrent event data. Unlike most test procedures listed above, the test statistics are constructed as contrasts of the sample means of the integrated weighted responses from the underlying recurrent event processes. It will be seen that the proposed test procedure performs well and especially for sparsely distributed data. The remainder of the chapter is organized as follows. Section 4.2 first considers the comparison problem for univariate panel count data and presents a class of test procedures. Section 4.3 generalizes the test procedure to multivariate panel count data. For both cases, the asymptotic normality of the test statistics is established. Section 4.4 investigates the finite sample properties of

the proposed test procedures through simulation studies and Section 4.5 applies the methodology to the skin cancer study described above. Some concluding remarks are provided in Section 4.6.

4.2 Nonparametric Comparison for Univariate Panel Count Data

We now consider m groups of independent subjects in a recurrent event study with total sample size n. For each subject, only panel count data are available, and the observation processes are different for subjects from different groups. Specifically, assume that there are n_l subjects in the lth group, $l = 1, \ldots, m$, and let S_l denote the set of indices for subjects in group l, where $\sum_{l=1}^{m} n_l = n$. Suppose Z_i is a group indicator of subject i $(i = 1, \ldots, n)$ which can always be labeled as a scalar variable. Without loss of generality, let $Z_i = 0$ for $i \in S_m$ (the control group). Also let $Y_i(t)$ be the counting process representing the total number of recurrent event occurrences up to time t from subject i. In addition, let C_i denote the censoring or follow-up time of subject i. It censors the observation times $T_{i,1} < T_{i,2} < \ldots$ in the sense that the event process $Y_i(\cdot)$ is observed only at jumps of $N_i(t) = N_i^*(C_i \wedge t)$, where $N_i^*(t) = \sum_{j=1}^{\infty}(T_{i,j} \leq t)$ and $a \wedge b$ denotes the minimum of a and b. Let m_i represent the total number of observation times for subject i and τ be the longest follow-up time. To account for the fact that subjects from different groups may have different observation processes, we assume that $N_i^*(t)$ depends on the treatment indicator Z_i through the rate model

$$E\{dN_i^*(t)|Z_i\} = \exp(\gamma Z_i)\lambda_0(t)dt, \tag{4.1}$$

where $\lambda_0(\cdot)$ is an unspecified continuous function and γ is an unknown regression parameter. Model (4.1) implies that Z_i has a multiplicative effect on the number of observations, and $\gamma = 0$ means that the observation processes are the same. Similar proportional models have been considered by many authors including Lin et al. (2000), Sun and Wei (2000), Lin et al. (2001), Sun et al. (2005) and Li et al. (2010) among others. The adequacy of model (4.1) is relatively easy to check since the observation process provides complete data.

Unlike the observation process, the recurrent event process associated with panel count data is not continuously observed and thus its model adequacy is generally difficult to check. In this chapter, we focus on a treatment comparison procedure which is model-free of the recurrent event process with panel count data while model (4.1) holds. Suppose that C_i is independent of Z_i, and given Z_i, C_i is independent of $\{Y_i(t), N_i^*(t)\}$. Also the observation process is assumed to be noninformative, that is, $Y_i(t)$ and $N_i^*(t)$ are independent given Z_i. The observed data consist of $\{N_i(t), Z_i, C_i, Y_i(T_{i,1}), \ldots, Y_i(T_{i,m_i}); \quad 0 \le t, T_{i,m_i} \le C_i,\ i = 1, \ldots, n\}$.

Our aim is to test the hypothesis

$$H_0 : E\{Y_i(t)|Z_i\} \text{ is independent of } Z_i,$$

that is, the occurrence rate of the recurrent event of interest is the same for different treatment groups. Let $\mu(t)$ denote the common mean function of $Y_i(t)$ under hypothesis H_0. Then under model (4.1) and the null hypothesis H_0, we have

$$E\Big\{ \sum_{j=1}^{m_i} Y_i(T_{i,j})|Z_i \Big\} = E\Big\{ \int_0^\tau Y_i(t) dN_i(t)|Z_i \Big\} = \int_0^\tau \mu(t)G(t)\exp(\gamma Z_i)\lambda_0(t)dt,$$

where $G(t) = P(C_i \geq t)$. Then

$$E\Big\{ \int_0^\tau \frac{Y_i(t)dN_i(t)}{\exp(\gamma Z_i)} | Z_i \Big\} = \int_0^\tau \mu(t)G(t)\lambda_0(t)dt.$$

Define

$$\tilde{Y}_i(t;\gamma) = \int_0^t \frac{Y_i(u)dN_i(u)}{\exp(\gamma Z_i)}. \tag{4.2}$$

Then H_0 can be formulated as

$$\tilde{H}_0 : E\{\tilde{Y}_i(t;\gamma)|Z_i\} \text{ is independent of } Z_i.$$

Note that $\tilde{Y}_i(t;\gamma)$ represents an integral of weighted responses from the underlying recurrent event process on subject i and is continuous on time t. Motivated by the idea commonly used for recurrent event data (Cook, Lawless and Nadeau (1996); Ghosh and Lin (2000); Wang and Chiang (2002)), we propose the following test statistic

$$\phi(\hat{\gamma}) = n^{\frac{1}{2}} \sum_{l=1}^{m} \int_0^\tau W(t) K_l d\hat{\mu}_l(t;\hat{\gamma}). \tag{4.3}$$

In the above, $W(t)$ is a predictable weighting process, $\hat{\gamma}$ represents an estimator of γ, $K_1, \ldots, K_m$ are a set of coefficients such that $\sum_{l=1}^m K_l = 0$, and

$$\hat{\mu}_l(t;\hat{\gamma}) = \frac{1}{n_l} \sum_{i \in S_l} \int_0^t d\tilde{Y}_i(u;\hat{\gamma}),\ l = 1, \ldots, m \tag{4.4}$$

is the sample mean of $\tilde{Y}_i(t;\hat{\gamma})$.

The test statistic $\phi(\hat{\gamma})$ represents a contrast of the sample means of the integrated weighted responses from the underlying recurrent event processes. The choice of

$K_1, \ldots, K_m$ depends on the comparison problem of interest and determines the interpretation of the contrast. To estimate γ, we can use the recurrent event data on the counting process $N_i^*(t)$, and in this case, according to Lin et al. (2000), γ can be consistently estimated by the unique solution to the estimating equation

$$U(\gamma) = n^{-1/2} \sum_{i=1}^{n} \int_0^{\tau} \Big\{ Z_i - \frac{\sum_{j=1}^{n} I(t \le C_j) \exp(\gamma Z_j) Z_j}{\sum_{j=1}^{n} I(t \le C_j) \exp(\gamma Z_j)} \Big\} dN_i(t) = 0. \tag{4.5}$$

We show in Appendix C that under regularity conditions (C1) to (C5), $\phi(\hat{\gamma})$ follows an asymptotic normal distribution with mean 0 and variance that can be consistently estimated by $\hat{\sigma}_\phi^2$. Therefore, a test of the hypothesis H_0 can be performed using $\phi(\hat{\gamma})/\hat{\sigma}_\phi$ based on the standard normal distribution.

4.3 Nonparametric Comparison for Multivariate Panel Count Data

Now suppose that there exist p ($p > 1$) types of recurrent events while model (4.1) still holds. Following the notation above, at each observation time, one observes $Y_i(t) = (Y_{i,1}(t), \ldots, Y_{i,p}(t))'$ with $Y_{i,k}(t)$ representing the total number of the kth type of recurrent event occurrences up to time t from subject i, $k = 1, \ldots, p$. Then the null hypothesis is

$$H_0^* : E\{Y_i(t)|Z_i\} \text{ is independent of } Z_i,$$

which can be formulated as

$$\tilde{H}_0^* : E\{\tilde{Y}_{i,k}(t;\gamma)|Z_i\} \text{ is independent of } Z_i,$$

where

$$\tilde{Y}_{i,k}(t;\gamma) = \int_0^t \frac{Y_i(u)dN_i(u)}{\exp(\gamma Z_i)}, \quad k = 1, \ldots, p. \tag{4.6}$$

This motivates the following test statistic for the hypothesis $\tilde{H}_0^*$

$$\phi^*(\hat{\gamma}) = n^{\frac{1}{2}} \sum_{l=1}^{m} \int_0^{\tau} W(t) K_l d\hat{\mu}_l^*(t;\hat{\gamma}), \tag{4.7}$$

where $W(t)$, $\hat{\gamma}$, $K_1, \ldots, K_m$ are defined in the same way as for univariate cases, and

$$\hat{\mu}_l^*(t;\hat{\gamma}) = \frac{1}{n_l} \sum_{i \in S_l} \int_0^t d\tilde{Y}_i^*(u;\hat{\gamma}), \text{ and } \tilde{Y}_i^*(t;\hat{\gamma}) = \sum_{k=1}^{p} \tilde{Y}_{i,k}(t;\hat{\gamma}), \; l = 1, \ldots, m. \tag{4.8}$$

By using the similar arguments given in Appendix C for univariate cases, one can show that $\phi^*(\hat{\gamma})$ is asymptotically normal with mean 0 and the variance that can be consistently estimated by

$$\hat{\sigma}_{\phi}^{*2} = \sum_{l=1}^{m} H_l^*(\hat{\gamma}) \hat{\Gamma}_l^* H_l^*(\hat{\gamma})',$$

In the above,

$$H_l^*(\hat{\gamma}) = (K_l \sqrt{\frac{n}{n_l}} \quad \sqrt{\frac{n_l}{n}} A^*(\hat{\gamma}) B^{-1}(\hat{\gamma})),$$

$$A^*(\gamma) = -\sum_{l=1}^{m-1} \frac{\sqrt{n}}{n_l} \sum_{i \in S_l} \sum_{k=1}^{p} \int_0^{\tau} W(t) K_l \frac{Z_i Y_{i,k}(t) dN_i(t)}{exp(\gamma Z_i)},$$

$$\hat{\Gamma}_l^* = n_l^{-1} \sum_{i \in S_l} \begin{pmatrix} \hat{a}_i^* \\ \hat{b}_i \end{pmatrix} (\hat{a}_i^* \;\; \hat{b}_i), \hat{a}_i^* = \int_0^{\tau} W(t) \{ d\tilde{Y}_i^*(t;\hat{\gamma}) - d\hat{\mu}_l^*(t;\hat{\gamma}) \}.$$

and B and $\hat{b}_i$ are given in Appendix C same as for univariate cases.

Therefore, a test of the hypothesis H_0 can be carried out by using the statistic

$\phi^*(\hat{\gamma})/\hat{\sigma}^*_\phi$ based on the standard normal distribution.

4.4 A Simulation Study

An extensive simulation study was conducted to assess the finite-sample properties of the test procedures described in Sections 4.2 and 4.3. In the study, we focused on the two-sample comparison problem with $m = 2$. Let $Z_i = 1$ for $i \in S_1$ (the treatment group) and $Z_i = 0$ for $i \in S_2$ (the control group). The follow-up time C_i was uniform on $(0.8\tau, \tau)$ with $\tau = 20$, $i = 1, \ldots, n$. We then generated the total number of observation times m_i from a Poisson distribution under model (4.1) with the mean $\Lambda_i(C_i)$,

$$\Lambda_i(t) = \exp(\gamma Z_i) \int_0^t \lambda_0(u)du, \tag{4.9}$$

and various choices of $\lambda_0(\cdot)$. The observation times T_{ij}'s were taken to be the order statistics of m_i random variables from a discrete uniform distribution over $(0, 0.1, 0.2, \ldots, C_i)$.

The panel count data $Y_i(t) = (Y_{i,1}(t), \ldots, Y_{i,p}(t))$ $(i = 1, \ldots, n)$ were assumed to follow non-homogeneous mixed Poisson processes. Specifically, for given T_{ij}'s and a latent variable Q_i, we generated $Y_i(T_{i,j})$ based on

$$Y_{i,k}(T_{i,j}) = Y^{**}_{i,k}(T_{i,1}) + Y^{**}_{i,k}(T_{i,2} - T_{i,1}) + ... + Y^{**}_{i,k}(T_{i,j} - T_{i,j-1})$$

for $j = 1, \ldots, m_i$, $k = 1, \ldots, p$. In the above, all $Y^{**}_{i,k}$ were assumed to follow Poisson distributions with the mean functions defined as, given Q_i and some baseline cumulative

mean function $\mu_k(t)$,

$$\begin{aligned} &\mathrm{E}\{\, Y_{i,k}^{**}(T_{i,1}) |\, Q_i \,\} = Q_i \, \mu_k(T_{i,1}) \exp(\beta Z_i), \\ &\mathrm{E}\{\, Y_{i,k}^{**}(T_{i,j} - T_{i,j-1}) |\, Q_i \,\} = Q_i \, \{\, \mu_k(T_{i,j}) - \mu_k(T_{i,j-1}) \,\} \exp(\beta Z_i) \end{aligned} \tag{4.10}$$

for $j = 2, ..., m_i$. Here β is a parameter representing the treatment difference and the Q_i's were generated from a Gamma distribution with mean 1 and variance 0.1. In the following, all results reported below are based on $W(t) = 1$ and 1000 replications with the significance level $\alpha = 0.05$.

Table 4.1 shows the estimated test sizes and powers with $\Lambda_i(t) = 0.75t \exp(\gamma Z_i)$ for (4.9) and univariate panel count data. In this case, the test statistic has the form

$$\phi(\hat{\gamma}) = n^{\frac{1}{2}} \int_0^\tau W(t) \{ d\hat{\mu}_1(t; \hat{\gamma}) - d\hat{\mu}_2(t; \hat{\gamma}) \}.$$

with $K_1 = 1$ and $K_2 = -1$, and the variance estimate

$$\hat{\sigma}_\phi^2 = H_1(\hat{\gamma}) \hat{\Gamma}_1 H_1(\hat{\gamma})' + H_2(\hat{\gamma}) \hat{\Gamma}_2 H_2(\hat{\gamma})',$$

where $H_1(\hat{\gamma}) = (\sqrt{\frac{n}{n_1}} \quad \sqrt{\frac{n_1}{n}} A_1(\hat{\gamma}) B^{-1}(\hat{\gamma}))$ and $H_2(\hat{\gamma}) = (-\sqrt{\frac{n}{n_2}} \quad \sqrt{\frac{n_2}{n}} A_1(\hat{\gamma}) B^{-1}(\hat{\gamma}))$. When $\gamma = 0$, both groups have the same observation process. Otherwise, the two observation processes are different. We took $\mu_1(t) = 0.25t$ and $\mu_1(t) = log(1 + t)$ for (4.10). It shows that the test sizes are all close to the nominal level 0.05. Also as expected, the powers increase when the sample size increases. We also considered other set-ups for univariate panel count data such as different values of γ or other forms of $\mu_1(\cdot)$ and obtained similar results.

Table 4.2 presents the test results on bivariate panel count data ($p = 2$). In this case, the test statistic has the form

$$\phi^*(\hat{\gamma}) = n^{\frac{1}{2}} \int_0^\tau W(t)\{d\hat{\mu}_1^*(t;\hat{\gamma}) - d\hat{\mu}_2^*(t;\hat{\gamma})\}.$$

when $K_1 = 1$ and $K_2 = -1$, and the variance estimate

$$\hat{\sigma}_\phi^{*2} = H_1^*(\hat{\gamma})\hat{\Gamma}_1^* H_1^*(\hat{\gamma})' + H_2^*(\hat{\gamma})\hat{\Gamma}_2^* H_2^*(\hat{\gamma})',$$

with

$$H_1^*(\hat{\gamma}) = (\sqrt{\frac{n}{n_l}} \quad \sqrt{\frac{n_1}{n}}A^*(\hat{\gamma})B^{-1}(\hat{\gamma})), H_2^*(\hat{\gamma}) = (-\sqrt{\frac{n}{n_2}} \quad \sqrt{\frac{n_2}{n}}A^*(\hat{\gamma})B^{-1}(\hat{\gamma})).$$

As with univariate cases, the proposed procedure also seems to perform well. Besides, comparing the same set-up when $\mu_1(t) = 0.25t$ in Tables 4.1 and 4.2, one may find that the test powers given by a bivariate analysis could be higher than those by a univariate analysis when $\beta \neq 0$. Similar comparisons can also be found for other set-ups between the univariate and multivariate cases.

For the analysis of univariate panel count data, we also investigated the performances of the proposed test procedure in comparison with the procedure given in Zhao and Sun (2011). The results are shown by Table 4.3. For both test procedures we took $\Lambda_i(t) = 0.75t\exp(\gamma Z_i)$ and $\mu_1(t) = \log(1+t)$, with C_i or $T_{i,j}$ generated in special schemes. We focused on the test sizes given by both procedures for varied follow-up times and different settings of data. Scheme 1 represents a regular setting of data, where $T_{i,j}$ followed a discrete uniform distribution on $(0, 0.1, 0.2, \ldots, C_i)$ like for

all the above tables, and C_i followed a uniform distribution on $(0.5\tau,\ \tau)$. Schemes 2 and 3 represent settings of sparsely distributed data with differently varied follow-up times. For both of those schemes, $T_{i,j}$'s followed a discrete uniform distribution on $(0, 0.01, 0.02, \ldots, C_i)$, with C_i being uniform on $(0.5\tau,\ \tau)$ or $(0.8\tau,\ \tau)$. We found that the test sizes given by Zhao and Sun (2011) seem to be inflated a little under Scheme 1, especially when sample sizes are small. However, results from Schemes 2 and 3 indicate that the test procedure in Zhao and Sun (2011) could overestimate the test sizes a lot when the data are sparsely distributed. One explanation might be related to the IRE $\hat{\Lambda}_n^{(l)}(t)$ employed for the mean function. When there are too few observations at some observation times, $\hat{\Lambda}_n^{(l)}(t)$ at those observation times may not perform well. Extremely, if in one group the responses are time non-decreasing with only one observation made at each observation time, then the procedure in Zhao and Sun (2011) will not be applicable since their variance estimator $\hat{\sigma}_l^2$ will be 0 on denominator.

4.5 An Application

In this section, we applied the proposed test procedure described in the previous sections to the data from the skin cancer chemoprevention trial introduced in Section 1.1.2.3. In the study, 291 patients were randomly assigned to the placebo or the DFMO group. Subjects were scheduled to be assessed every six months, but the actual observation times varied a lot. Figure 4.1 shows the distribution of the numbers of observations made on all the 1159 observation times. Of the observation times, 70.6% of them had 1 or 2 observations and 92.3% of them had observations 5 or less. In other words, the observed data are very sparsely distributed. As implied by Table 4.3

of the simulation study, one may not wish to employ the mean function estimators for analyzing such data since their performances may be highly affected.

For the analysis of DFMO treatment in reducing the occurrences of the two types of related non-melanoma skin cancers: basal cell carcinoma (BCC) and squamous cell carcinoma (SCC), we will focus on the 290 patients with at least one observation by applying the proposed test procedure. Among the patients, 143 were from the DFMO group ($Z_i = 1$, $i = 1, \ldots, 143$), and others were from the placebo group ($Z_i = 0$, $i = 144, \ldots, 290$). We took the last observation times as each follow-up time C_i and the largest follow-up time $\tau = 1879$. To check whether model (4.1) is appropriate for the study, Figure 4.2 shows the Aalen-Breslow-type estimates of the mean numbers of observation times $N(t)$. It appears that the two group means are quite proportional to each other. The results are given by Table 4.4 with $W(t) = 1$.

We first did two univariate analyses on the occurrences of BCC and SCC separately. Let $Y_i(t)$ denote the total number of the occurrences of BCC or SCC up to time t on subject i, then treatment comparisons can be conducted by applying the proposed test procedure described in Section 4.2. With the significance level 0.05, we concluded that DFMO treatment was significantly effective on reducing the occurrences of BCC, but not on SCC.

Next, we did a bivariate analysis on both BCC and SCC jointly. Let $Y_{i,1}(t)$ and $Y_{i,2}(t)$ denote the total numbers of the occurrences of BCC and SCC, respectively, up to time t on subject i, and apply the proposed test procedure described in Section 4.3. With respect to the overall treatment effectiveness on reducing the non-melanoma skin cancers, the results indicate there was no enough evidence to conclude a significant effect of the 0.5g/m2/day PO DFMO. In comparison, the semiparametric regression

model proposed by Li et al. (2011) also lead to a similar conclusion.

4.6 Discussions

This chapter proposed a class of test procedures for comparing panel count data with different observation processes. The proposed test statistics were constructed as contrasts of the sample means of the integrated weighted responses from the underlying recurrent event processes. In comparison, the test statistics in Zhao and Sun (2011) were formulated as the sums of the differences between the integrated weighted mean function estimators and their averages with IRE employed. As mentioned above, although the procedure in Zhao and Sun (2011) is more general, its performance may be affected due to much varied follow-up times or too few observations at some observation times. The proposed test procedure has the advantage that it works well in such cases.

In this chapter, for the observation process $N_i^*(t)$, the proportional rates model assumption was used to account for the fact that it may depend on the treatment indicator. In practice, however, one could model $N_i^*(t)$ differently. Here we considered a constant proportional factor in model (4.1). In reality, a treatment may play differently on the observation process over time, so one may wish to consider time-dependent γ or $Z_i(\cdot)$, both of which are currently under investigation.

Chapter 5

FUTURE RESEARCH

In this chapter, we discuss some potential directions of future research for semiparametric and nonparametric analysis of panel count data following Chapters 2 to 4.

5.1 Analyzing Panel Count Data with Dependent Observation Processes and a Terminal Event

There exist several directions for future research on the problem discussed here. One of the model assumptions is that conditioning on $\mathcal{Z}(t)$, the time of censoring C is independent of $\{N^*(t), Y^*(t), D\}$. However, this may not be always true in reality since C_i may also be informative about the event process and death. It would be useful to extend the model to cases with informative censoring.

When building up estimating equations for the analysis, we also assumed that given Z_i and T_i^*, $\tilde{N}_i(t)$ was a nonhomogeneous Poisson process on $[0, \tau]$. Instead, we could

generalize the estimation method considered here to observation processes without the Poisson assumption.

In models (2.1) to (2.3), we considered multiplicative covariate effects. Instead, it is possible that the covariates effects are in other forms. Meanwhile, with respect to the correlated relationships considered here, we employed a function of filtration on the observation process in (2.2). One may consider similar terms regarding the recurrent event process as well and incorporate them into (2.1) and (2.3). For example, for the terminal event model (2.3), it is possible that the total number of recurrent events can increase the intensity of the terminal event.

A practical problem with model (2.2) involves the choice of $h(\cdot)$. Regarding the transformation variable of $g(\cdot)$, $h(\cdot)$ can take many forms representing various dependent structures between $Y^*(t)$ and $N^*(t)$. However, for a given $g(\cdot)$, $h(\cdot)$ cannot be arbitrary since the time non-decreasing property of $E\{Y^*(t)|\mathcal{Z}(t), \mathcal{F}_t, D \geq t\}$ must be theoretically satisfied. One may use a more general form instead for the transformation variable or extend $g(\cdot)$ to be functionals.

5.2 Semiparametric Analysis of Multivariate Panel Count Data with a Terminal Event

Similar as the univariate panel count data analysis discussed above, for multivariate panel count data analysis, one can make the semiparametric regression models more practical, for example, by allowing informative censoring for the response variable or considering other counting processes for possible observation processes. Although the simulation results show the proposed procedure is robust to the nonhomogeneous

Poisson assumption on the observation processes, one may wish to use other models that can be applied more generally.

For the recurrent event processes, we considered an additive model in (3.1). By doing this, the effects from covariate or a dependent observation process are straightforward to interpret. Apart from this, however, one may easily extend (3.1) to other forms. For example, transformation models considered for the univariate panel count data analysis in Chapter 2 can also be employed here.

One hidden model assumption with model (3.1) is that $Z(t)$ plays the same for all recurrent events of interest because a common β is employed. A similar assumption also exists for the observation process with model (3.2). Although theoretically such models make sense, practically they may not. For example, if a covariate affects oppositely on two types of recurrent events, results lead by model (3.1) may be hard to interpret.

5.3 Nonparametric Comparison for Panel Count Data with Unequal Observation Processes

There exist several directions for future work. For the observation process $N_i^*(t)$, the proportional rates model assumption was used to account for the fact that it may depend on the treatment indicator. In practice, however, one could model $N_i^*(t)$ differently. Here we considered a constant proportional factor in model (4.1). In reality, a treatment may play differently on the observation process over time, so one may wish to consider time-dependent $\gamma(\cdot)$ or $Z_i(\cdot)$, both of which are currently under investigation.

In the analysis, we assumed noninformative $N_i^*(\cdot)$ with respect to $Y_i(\cdot)$ and inde-

pendent identically distributed C_i for simplicity. In realistic cases, any of the above assumptions could be violated. Especially, a way to account for possible dependence between $N_i^*(\cdot)$ and $Y_i(\cdot)$ into nonparametric comparison is a challenging direction for future work.

Appendix A

A.1 Proof of Theorem 2.1

To derive the asymptotic properties of the proposed estimator $\hat{\theta}$, we need the following regularity conditions.

(C1). $\{\tilde{N}_i(\cdot), \tilde{Y}_i(\cdot), T_i^*, I(D_i \leq C_i), Z_i(\cdot)\}_{i=1}^n$ are independent and identically distributed.

(C2). There exists a $\tau > 0$ such that $P(T_i^* \geq \tau) > 0$.

(C3). Both $\tilde{N}_i(\tau)$ and $\tilde{Y}_i(\tau)$ $(i = 1, \ldots, n)$ are bounded.

(C4). $W(t)$ and $Z_i(\cdot)$, $i = 1, \ldots, n$, have bounded variations and $W(t)$ converges almost surely to a deterministic function $w(t)$ uniformly in $t \in [0, \tau]$.

(C5). $A_\theta = E\Big[\int_0^\tau w(t)\big\{X_1(t) - e_x(t)\big\}^{\otimes 2}\omega_1(t)\dot{g}\{\mu_0(t)e^{\theta_0' X_1(t)}\}e^{\theta_0' X_1(t)+\gamma_0' Z_1(t)}\mu_0(t)d\Lambda_0(t)\Big]$, Ω_δ and $\Omega_\gamma = E\Big[\int_0^\tau \big\{Z_1(t) - \bar{z}(t)\big\}^{\otimes 2} I(C \geq t|Z_1)e^{\gamma_0' Z_1(t)}d\Lambda_0(t)\Big]$ are all positive definite.

Define

$$U_1(\theta;\gamma)=\sum_{i=1}^{n}\int_0^{\tau}W(t)X_i(t)\Big[\widetilde{Y}_i(t)d\widetilde{N}_i(t)-\hat{\omega}_i(t)g\big\{\hat{\mu}_0(t;\theta,\gamma)e^{\theta' X_i(t)}\big\}e^{\gamma' Z_i(t)}d\hat{\Lambda}_0(t,\gamma)\Big],$$

and note that $\hat{\mu}_0(t;\theta,\gamma)$ satisfies

$$\sum_{i=1}^{n}\Big[\widetilde{Y}_i(t)d\widetilde{N}_i(t)-\hat{\omega}_i(t)g\big\{\hat{\mu}_0(t;\theta,\gamma)e^{\theta' X_i(t)}\big\}e^{\gamma' Z_i(t)}d\hat{\Lambda}_0(t;\gamma)\Big]=0. \qquad (A.1)$$

Let $\hat{A}_\theta(\theta)=-n^{-1}\partial U_1(\theta,\hat{\gamma})/\partial\theta'$, $\hat{A}_\gamma(\gamma)=-n^{-1}\partial U_1(\theta_0,\gamma)/\partial\gamma'$, $A_\theta=\lim_{n\to\infty}\hat{A}_\theta(\theta_0)$ and $A_\gamma=\lim_{n\to\infty}\hat{A}_\gamma(\gamma_0)$. Then the Taylor series expansions of $U_1(\hat{\theta};\hat{\gamma})$ at $(\theta_0;\hat{\gamma})$ and (θ_0,γ_0) yield $n^{1/2}(\hat{\theta}-\theta_0)=A_\theta^{-1}n^{-1/2}U_1(\theta_0;\hat{\gamma})+o_p(1)=A_\theta^{-1}\Big\{n^{-1/2}U_1(\theta_0;\gamma_0)-A_\gamma n^{1/2}(\hat{\gamma}-\gamma_0)\Big\}+o_p(1)$. To prove Theorem 2.1, we will need the following four steps (i)-(iv).

(i) First, using some derivation operation to $U_1(\theta;\gamma)$ and (A.1), we can get

$$\hat{A}_\theta(\theta)=n^{-1}\sum_{i=1}^{n}\int_0^{\tau}W(t)\big\{X_i(t)-\hat{E}_X(t;\theta,\hat{\gamma})\big\}^{\otimes 2}p_i^x(t)\hat{\mu}_0(t;\theta,\hat{\gamma})d\hat{\Lambda}_0(t,\hat{\gamma}),$$

where $p_i^x(t)=\hat{\omega}_i(t)\dot{g}\{\hat{\mu}_0(t;\theta,\hat{\gamma})e^{\theta' X_i(t)}\}e^{\theta' X_i(t)+\hat{\gamma}' Z_i(t)}$.

(ii) The use of the Taylor expansion to $U_1(\theta_0;\gamma_0)$ yields

$$U_1(\theta_0;\gamma_0)=\sum_{i=1}^{n}\int_0^{\tau}W(t)X_i(t)\Big[\widetilde{Y}_i(t)d\widetilde{N}_i(t)-\hat{\omega}_i(t)\Big[g\big\{\mu_0(t)e^{\theta_0' X_i(t)}\big\}$$

$$+\dot{g}\big\{\mu^*(t)e^{\theta_0' X_i(t)}\big\}e^{\theta_0' X_i(t)}\big\{\hat{\mu}_0(t;\theta_0,\gamma_0)-\mu_0(t)\big\}\Big]e^{\gamma_0' Z_i(t)}d\hat{\Lambda}_0(t,\gamma_0)\Big],$$

where μ^* lies on the line segment between $\mu_0(t)$ and $\hat{\mu}_0(t;\theta_0,\gamma_0)$. This and the linear expansion of (A.1) at $\theta=\theta_0$ and $\gamma=\gamma_0$ give us

$$\big\{\hat{\mu}_0(t;\theta_0,\gamma_0)-\mu_0(t)\big\}d\hat{\Lambda}_0(t,\gamma_0)=\frac{\sum_{i=1}^n\Big[\widetilde{Y}_i(t)d\widetilde{N}_i(t)-\hat{\omega}_i(t)g\{\mu_0(t)e^{\theta_0'X_i(t)}\}e^{\gamma_0'Z_i(t)}d\hat{\Lambda}_0(t;\gamma_0)\Big]}{\sum_{i=1}^n\hat{\omega}_i(t)\dot{g}\{\mu^{**}(t)e^{\theta_0'X_i(t)}\}e^{\theta_0'X_i(t)+\gamma_0'Z_i(t)}},$$

where μ^{**} lies between $\mu_0(t)$ and $\hat{\mu}_0(t;\theta_0,\gamma_0)$. Hence we have

$$n^{-1/2}U_1(\theta_0;\gamma_0)=n^{-1/2}\sum_{i=1}^n\int_0^\tau W(t)\big\{X_i(t)-e_x(t)\big\}d_i+o_p(1),$$

where

$$d_i=dM_i(t)+\{\omega_i(t)-\hat{\omega}_i(t)\}g\big\{\mu_0(t)e^{\theta_0'X_i(t)}\big\}e^{\gamma_0'Z_i(t)}d\Lambda_0(t)$$

$$-\hat{\omega}_i(t)g\big\{\mu_0(t)e^{\theta_0'X_i(t)}\big\}e^{\gamma_0'Z_i(t)}\big\{d\hat{\Lambda}_0(t,\gamma_0)-d\Lambda_0(t)\big\}.$$

Then it follows from Equation (2.7) that

$$d\hat{\Lambda}_0(t,\gamma_0)-d\Lambda_0(t)=n^{-1}\sum_{i=1}^n\Big[\frac{dM_i^*(t)}{s^{(0)}(t)}+\frac{\omega_i(t)-\hat{\omega}_i(t)}{s^{(0)}(t)}e^{\gamma_0'Z_i(t)}d\Lambda_0(t)\Big]+o_p(n^{-1/2}).$$

According to the functional delta method (van der Vaart & Wellner, 1996, Theorem 3.9.4, page 374) and the martingale central limit theorem, we have

$$\hat{\omega}_i(t)-\omega_i(t)=n^{-1}\omega_i(t)\Big[\sum_{j=1}^n\int_0^t\frac{e^{\delta_0'Z_i(u)}dM_j^d(u)}{r^{(0)}(u)}$$

$$+H(t;Z_i)'\Omega_\delta^{-1}\sum_{j=1}^n\int_0^\tau\Big\{Z_j(u)-\frac{r^{(1)}(u)}{r^{(0)}(u)}\Big\}dM_j^d(u)\Big]+o_p(n^{-1/2}).$$

This gives

$$n^{-1/2}U_1(\theta_0;\gamma_0) = n^{-1/2}\sum_{i=1}^{n}\Big[\int_0^\tau W(t)\big\{X_i(t)-e_x(t)\big\}dM_i(t)-\int_0^\tau \frac{W(t)\Upsilon(t)}{s^{(0)}(t)}dM_i^*(t)\Big]$$

$$-n^{-1/2}\sum_{i=1}^{n}\int_0^\tau\left[\frac{B_1(u)}{r^{(0)}(u)}+B_2\Big\{Z_i(u)-\frac{r^{(1)}(u)}{r^{(0)}(u)}\Big\}\right]dM_i^d(u)+o_p(1),$$

where

$$B_i^*(t) = W(t)\omega_i(t)e^{\gamma_0' Z_i(t)}\Big[\big\{X_i(t)-e_x(t)\big\}g\big\{\mu_0(t)e^{\theta_0' X_i(t)}\big\}-\frac{\Upsilon(t)}{s^{(0)}(t)}\Big],$$

$B_1(u) = n^{-1}\sum_{i=1}^n e^{\delta_0' Z_i(u)}\int_0^\tau I(u<t)B_i^*(t)d\Lambda_0(t)$, and
$B_2 = n^{-1}\sum_{i=1}^n\int_0^\tau B_i^*(t)H(t;Z_i)'\Omega_\delta^{-1}d\Lambda_0(t)$.

(iii) By the Taylor series expansions and differentiation of (A.1) with respect to γ, we can obtain

$$\hat{A}_\gamma(\gamma) = n^{-1}\sum_{i=1}^{n}\int_0^\tau W(t)\big[X_i(t)-\hat{E}_X(t;\theta_0,\gamma)\big]\hat{\omega}_i(t)g\{\hat{\mu}_0(t;\theta_0,\gamma)e^{\theta_0' X_i(t)}\}$$

$$\times e^{\gamma_0' Z_i(t)}\big[Z_i(t)-\bar{Z}(t;\gamma)\big]'d\hat{\Lambda}_0(t;\gamma)\,.$$

(iv) According to Equation (2.6) and the arguments similar as Ghosh & Lin (2002), one can show that

$$n^{1/2}\{\hat{\gamma}-\gamma_0\} = \Omega_\gamma^{-1}n^{-1/2}\sum_{i=1}^{n}\Big[\int_0^\tau\Big\{Q_1\Big(Z_i(t)-\frac{r^{(1)}(t)}{r^{(0)}(t)}\Big)+\frac{Q_2(t)}{r^{(0)}(t)}\Big\}dM_i^d(t)$$

$$+\int_0^\tau\big\{Z_i(t)-\bar{z}(t)\big\}dM_i^*(t)\Big]+o_p(1). \qquad (A.2)$$

where $\Omega_\gamma = E\Big[\int_0^\tau \big\{Z_1(t) - \bar{z}(t)\big\}^{\otimes 2} I(C \geq t|Z_1) e^{\gamma_0' Z_1(t)} d\Lambda_0(t)\Big]$, $Q_1 = \lim_{n\to\infty} n^{-1} \sum_{i=1}^n$ $\int_0^\tau \big\{Z_i(t) - \bar{Z}(t;\gamma)\big\} Q_3(t;Z_i)' \Omega_\delta^{-1} dM_i^*(t)$, $Q_2(t) = \lim_{n\to\infty} n^{-1} \sum_{i=1}^n \int_0^\tau \big\{Z_i(u) - \bar{Z}(u;\gamma)\big\}$ $e^{\delta' Z_i(t)} I(u \geq t) dM_i^*(u)$, and $Q_3(t;Z_i) = \int_0^t \big\{Z_i(u) - \frac{r^{(1)}(u)}{r^{(0)}(u)}\big\} e^{\delta' Z_i(u)} d\Delta_0(u)$.

Combining the results in steps (i)-(iv), we have

$$U_1(\theta_0;\hat{\gamma}) = \sum_{i=1}^n \Big[\int_0^\tau W(t)\big\{X_i(t) - e_x(t)\big\} dM_i(t) - \int_0^\tau \frac{W(t)\Upsilon(t)}{s^{(0)}(t)} dM_i^*(t)\Big] - \sum_{i=1}^n \int_0^\tau \Big[\frac{B_1(t)}{r^{(0)}(t)}$$

$$+ B_2\Big\{Z_i(t) - \frac{r^{(1)}(t)}{r^{(0)}(t)}\Big\}\Big] dM_i^d(t) - A_\gamma \Omega_\gamma^{-1} \sum_{i=1}^n \Big[\int_0^\tau \Big\{Q_1\big(Z_i(t) - \frac{r^{(1)}(t)}{r^{(0)}(t)}\big)$$

$$+ \frac{Q_2(t)}{r^{(0)}(t)}\Big\} dM_i^d(t) + \int_0^\tau \big\{Z_i(t) - \bar{z}(t)\big\} dM_i^*(t)\Big] + o_p(n^{1/2}).$$

Then it follows from the multivariate central limit theorem that the conclusion holds.

A.2 Proof of the Null Distribution of $\mathcal{F}(t,x)$ in Chapter 2

Let $V(\hat{\theta},\hat{\gamma}) = \sum_{i=1}^n \int_0^t I(X_i(u) \leq x) d\hat{M}_i(u;\hat{\theta},\hat{\gamma})$. Then the Taylor series expansion gives

$$\mathcal{F}(t,x;\hat{\theta},\hat{\gamma}) = n^{-1/2} V(\theta_0,\gamma_0) + \frac{\partial V(\theta_0,\gamma_0)}{n\partial\gamma'} \sqrt{n}(\hat{\gamma} - \gamma_0) + \frac{\partial V(\theta_0,\hat{\gamma})}{n\partial\theta'} \sqrt{n}(\hat{\theta} - \theta_0) + o_p(1).$$

Using the arguments and algebra manipulations similar to those in Appendix A.1, one can show that $V(\theta_0,\gamma_0) = \sum_{i=1}^n \eta_{1i}(t,x)$. Note that one can estimate $n^{-1}\partial V(\theta_0,\gamma_0)/\partial\gamma'$ and $n^{-1}\partial V(\theta_0,\hat{\gamma})/\partial\theta'$ by $-\hat{\Phi}_\gamma(t,x)$ and $-\hat{\Phi}_\theta(t,x)$, respectively. It then follows from

(A.2) that

$$\sqrt{n}(\hat{\gamma} - \gamma_0) = \Omega_\gamma^{-1} n^{-1/2} \sum_{i=1}^{n} \eta_{2i} + o_p(1)\,.$$

Also it follows from Theorem 2.1 that

$$\sqrt{n}(\hat{\theta} - \theta_0) = A_\theta^{-1} n^{-1/2} \sum_{i=1}^{n} (\xi_{1i} - \xi_{2i} - \xi_{3i}) + o_p(1)\,.$$

This indicates that $\mathcal{F}(t, x; \hat{\theta}, \hat{\gamma})$ can be expressed as a sum of i.i.d. zero-mean terms for fixed t and thus by the multivariate central limit theorem, $\mathcal{F}(t, x)$ converges in finite-dimensional distributions to a zero-mean Gaussian process. Since $\mathcal{F}(t, x)$ is tight based on the empirical process theory, $\mathcal{F}(t, x)$ converges weakly to a zero-mean Gaussian process that can be approximated by the zero-mean Gaussian process $\hat{\mathcal{F}}(t, x)$ given by Equation (2.8).

Appendix B

B.1 Derivation of Equation (3.4)

$$
\begin{aligned}
& E\{\widetilde{Y}_{ik}(t)d\widetilde{N}_{ik}(t)\} \\
= \;& E\Big\{E\big\{I(C_i \geq t)Y_i^*(t)dN_i^*(t)|Z_i(t), \mathcal{F}_{it}\big\}\Big\} \\
= \;& E\Big\{E\big\{I(C_i \geq t)|Z_i(t)\big\}E\big\{Y_{ik}^*(t)dN_{ik}^*(t)|Z_i(t), \mathcal{F}_{ikt}\big\}\Big\} \\
= \;& E\Big\{E\big\{I(C_i \geq t)|Z_i(t)\big\}E\big\{Y_{ik}^*(t)|D_i \geq t, Z_i(t), \mathcal{F}_{ikt}\big\}E\big\{dN_{ik}^*(t)|Z_i(t)\big\}\Big\} \\
= \;& E\Big\{E\{I(C_i \geq t)g\{\mu_{0k}(t)e^{\theta' X_i(t)}\}e^{\gamma' Z_i(t)}d\Lambda_0(t)|Z_i(t), \mathcal{F}_{it}\}\Big\} \\
= \;& E\big\{I(C_i \geq t)\{\mu_{0k}(t) + \theta' X_{ik}(t)\}e^{\gamma' Z_i(t)}d\Lambda_{0k}(t)\big\},
\end{aligned}
$$

where the third equality holds because

$$
\begin{aligned}
& E\big\{Y_{ik}^*(t)dN_{ik}^*(t)|Z_i(t), \mathcal{F}_{it}\big\} \\
= \;& E\Big\{E\big\{Y_i^*(t)dN_i^*(t)|D_i, Z_i(t), \mathcal{F}_{it}\big\}\Big\} \\
= \;& E\big\{Y_{ik}^*(t)dN_{ik}^*(t)|D_i \geq t, Z_i(t), \mathcal{F}_{ikt}\big\}P(D_i \geq t|Z_i(t)) + 0 \times P(D_i < t|Z_i(t)) \\
= \;& E\big\{Y_{ik}^*(t)|D_i \geq t, Z_i(t), \mathcal{F}_{ikt}\big\}E\big\{dN_{ik}^*(t)|D_i \geq t, Z_i(t)\big\}P(D_i \geq t|Z_i(t)) \\
= \;& E\big\{Y_{ik}^*(t)|D_i \geq t, Z_i(t), \mathcal{F}_{ikt}\big\}E\big\{dN_{ik}^*(t)|Z_i(t)\big\}.
\end{aligned}
$$

B.2 Proof of Theorem 3.1

To derive the asymptotical properties of the proposed estimator $\hat{\theta}$, we need the following regularity conditions for $i = 1, \ldots, n$.

(C1). $\{\tilde{N}_{ik}(\cdot), \tilde{Y}_{ik}(\cdot), T_i^*, I(D_i \leq C_i), Z_i(\cdot)\}$ are independent and identically distributed.

(C2). There exists a $\tau > 0$ such that $P(T_i^* \geq \tau) > 0$.

(C3). Both $\tilde{N}_{ik}(\tau)$ and $\tilde{Y}_{ik}(\tau)$ are bounded.

(C4). $W(t)$ and $Z_i(\cdot)$ have bounded variations and $W(t)$ converges almost surely to a deterministic function $w(t)$ uniformly in $t \in [0, \tau]$.

(C5). $A_\theta = \sum_{k=1}^K E\Big[\int_0^\tau w(t)\big\{X_{1k}(t) - \bar{x}_k(t)\big\}^{\otimes 2}\omega_1(t)e^{\gamma_0' Z_1(t)}d\Lambda_{0k}(t)\Big]$, Ω_δ and $\Omega_\gamma = \sum_{k=1}^K E\Big[\int_0^\tau \big\{Z_1(t) - \bar{z}(t)\big\}^{\otimes 2} d\tilde{N}_{1k}(t)\Big]$ are all positive definite.

Define

$$U_1(\theta;\gamma) = \sum_{i=1}^n \sum_{k=1}^K \int_0^\tau W(t)X_{ik}(t)\Big[\widetilde{Y}_{ik}(t)d\widetilde{N}_{ik}(t) - \hat{\omega}_i(t)\big\{\hat{\mu}_{0k}(t;\theta,\gamma) + \theta' X_{ik}(t)\big\}e^{\gamma' Z_i(t)}d\hat{\Lambda}_{0k}(t;\gamma)\Big],$$

and note that $\hat{\mu}_{0k}(t;\theta,\gamma)$ satisfies

$$\sum_{i=1}^n \Big[\widetilde{Y}_{ik}(t)d\widetilde{N}_{ik}(t) - \hat{\omega}_i(t)\big\{\hat{\mu}_{0k}(t;\theta,\gamma) + \theta' X_{ik}(t)\big\}e^{\gamma' Z_i(t)}d\hat{\Lambda}_{0k}(t;\gamma)\Big] = 0. \qquad (B.1)$$

Let $\hat{A}_\theta = -n^{-1}\partial U_1(\theta,\hat{\gamma})/\partial\theta'$, $\hat{A}_\gamma(\gamma) = -n^{-1}\partial U_1(\theta_0,\gamma)/\partial\gamma'$, $A_\theta = \lim_{n\to\infty}\hat{A}_\theta$ and $A_\gamma = \lim_{n\to\infty}\hat{A}_\gamma(\gamma_0)$. Taylor expansions of $U_1(\hat{\theta};\hat{\gamma})$ at $(\theta_0;\hat{\gamma})$ and $U_1(\theta_0;\hat{\gamma})$ at (θ_0,γ_0) yields $n^{1/2}(\hat{\theta}-\theta_0) = A_\theta^{-1}n^{-1/2}U_1(\theta_0;\hat{\gamma}) = A_\theta^{-1}\Big\{n^{-1/2}U_1(\theta_0;\gamma_0) - A_\gamma n^{1/2}(\hat{\gamma}-\gamma_0)\Big\} + o_p(1)$.

(1) First, using some derivation operation to $U_1(\theta;\gamma)$ and (B.1), we can get

$$\hat{A}_\theta = n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^\tau W(t)\big\{X_{ik}(t)-\bar{X}_k(t;\hat{\gamma})\big\}^{\otimes 2}\hat{\omega}_i(t)e^{\hat{\gamma}'Z_i(t)}d\hat{\Lambda}_{0k}(t,\hat{\gamma}).$$

(2) Taylor expansion to $U_1(\theta_0;\gamma_0)$ yields

$$U_1(\theta_0;\gamma_0) = \sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^\tau W(t)X_{ik}(t)\Big[\widetilde{Y}_{ik}(t)d\widetilde{N}_{ik}(t)-\hat{\omega}_i(t)e^{\gamma_0'Z_i(t)}$$

$$\times\Big(\mu_{0k}(t)+\theta_0'X_{ik}(t)+\hat{\mu}_{0k}(t;\theta_0,\gamma_0)-\mu_{0k}(t)\Big)d\hat{\Lambda}_{0k}(t,\gamma_0)\Big].$$

From the linear expansion of (B.1) with $\theta=\theta_0$ and $\gamma=\gamma_0$, we have

$$\big\{\hat{\mu}_{0k}(t;\theta_0,\gamma_0)-\mu_{0k}(t)\big\}d\hat{\Lambda}_{0k}(t,\gamma_0)$$

$$=\frac{\sum_{i=1}^n\Big[\widetilde{Y}_{ik}(t)d\widetilde{N}_{ik}(t)-\hat{\omega}_i(t)\{\mu_{0k}(t)+\theta_0'X_{ik}(t)\}e^{\gamma_0'Z_i(t)}d\hat{\Lambda}_{0k}(t;\gamma_0)\Big]}{\sum_{i=1}^n\hat{\omega}_i(t)e^{\gamma_0'Z_i(t)}}.$$

Hence, $n^{-1/2}U_1(\theta_0;\gamma_0)=n^{-1/2}\sum_{i=1}^n\int_0^\tau W(t)\big\{X_{ik}(t)-\bar{x}_k(t)\big\}d_{ik}+o_p(1)$, where $d_{ik}=dM_{ik}(t)+\big\{\omega_i(t)-\hat{\omega}_i(t)\big\}\big\{\mu_{0k}(t)+\theta_0'X_{ik}(t)\big\}e^{\gamma_0'Z_i(t)}d\Lambda_{0k}(t)-\hat{\omega}_i(t)\big\{\mu_{0k}(t)+\theta_0'X_{ik}(t)\big\}e^{\gamma_0'Z_i(t)}\big\{d\hat{\Lambda}_{0k}(t,\gamma_0)-d\Lambda_{0k}(t)\big\}$. Then it follows from (3.8) that

$$d\hat{\Lambda}_{0k}(t,\gamma_0)-d\Lambda_{0k}(t)=n^{-1}\sum_{i=1}^{n}\Big[\frac{dM_{ik}^*(t)}{s^{(0)}(t)}+\frac{\omega_i(t)-\hat{\omega}_i(t)}{s^{(0)}(t)}e^{\gamma_0'Z_i(t)}d\Lambda_{0k}(t)\Big]+o_p(n^{-1/2}).$$

According to the functional delta method (van der Vaart and Wellner, 1996, Theorem 3.9.4, page 374) and the martingale central limit theorem, we have

$$\hat{\omega}_i(t)-\omega_i(t)=n^{-1}\omega_i(t)\Big[\sum_{j=1}^{n}\int_0^t\frac{e^{\delta_0'Z_i(u)}dM_j^d(u)}{r^{(0)}(u)}$$

$$+H(t;Z_i)'\Omega_\delta^{-1}\sum_{j=1}^{n}\int_0^\tau\Big\{Z_j(u)-\frac{r^{(1)}(u)}{r^{(0)}(u)}\Big\}dM_j^d(u)\Big]+o_p(n^{-1/2}).$$

Then, we obtain that

$$n^{-1/2}U_1(\theta_0;\gamma_0)=n^{-1/2}\sum_{k=1}^{K}\sum_{i=1}^{n}\Big[\int_0^\tau W(t)\big\{X_{ik}(t)-\bar{x}_k(t)\big\}dM_{ik}(t)-\int_0^\tau\frac{W(t)\Upsilon_k(t)}{s^{(0)}(t)}dM_{ik}^*(t)\Big]$$

$$-n^{-1/2}\sum_{i=1}^{n}\int_0^\tau\left[\frac{B_{1k}(u)}{r^{(0)}(u)}+B_{2k}\Big\{Z_i(u)-\frac{r^{(1)}(u)}{r^{(0)}(u)}\Big\}\right]dM_i^d(u)+o_p(1),$$

where $B_{ik}^*(t)=W(t)\omega_i(t)e^{\gamma_0'Z_i(t)}\Big[\big\{X_{ik}(t)-\bar{x}_k(t)\big\}\big\{\mu_{0k}(t)+\theta_0'X_{ik}(t)\big\}-\frac{\Upsilon_k(t)}{s^{(0)}(t)}\Big]$, $B_{1k}(u)=n^{-1}\sum_{i=1}^{n}e^{\delta_0'Z_i(u)}\int_0^\tau I(u<t)B_{ik}^*(t)d\Lambda_{0k}(t)$, and $B_{2k}=n^{-1}\sum_{i=1}^{n}\int_0^\tau B_{ik}^*(t)H(t;Z_i)'\Omega_\delta^{-1}d\Lambda_{0k}(t)$.

(3) Based on some Taylor expansions and differentiation of (B.1) with respect to γ, we have

$$\hat{A}_\gamma(\gamma)=n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^\tau W(t)\big\{X_{ik}(t)-\bar{x}_k(t)\big\}\big\{Z_i(t)-\bar{Z}(t;\gamma)\big\}'\hat{\omega}_i(t)e^{\gamma_0'Z_i(t)}$$

$$\times\big\{\hat{\mu}_{0k}(t;\theta_0,\gamma)+\theta_0'X_{ik}(t)\big\}d\hat{\Lambda}_{0k}(t;\gamma),$$

(4) According to equation (3.7) and arguments similar as Ghosh and Lin (2002), we have

$$n^{1/2}\{\hat{\gamma}-\gamma_0\}=\Omega_\gamma^{-1}n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{K}\Big[\int_0^\tau\Big\{Q_{1k}\Big(Z_i(t)-\frac{r^{(1)}(t)}{r^{(0)}(t)}\Big)+\frac{Q_{2k}(t)}{r^{(0)}(t)}\Big\}dM_{ik}^d(t)$$

$$+\int_0^\tau\big\{Z_i(t)-\bar{z}(t)\big\}dM_{ik}^*(t)\Big]+o_p(1).\qquad(B.2)$$

where $\Omega_\gamma = \sum_{k=1}^K E\Big[\int_0^\tau \big\{Z_1(t) - \bar{z}(t)\big\}^{\otimes 2} d\tilde{N}_{1k}(t)\Big]$, $Q_{1k} = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \int_0^\tau \big\{Z_i(t) - \bar{Z}(t;\gamma)\big\} Q_3(t;Z_i)' \Omega_\delta^{-1}\, dM_{ik}^*(t)$, $Q_{2k}(t) = \lim_{n\to\infty} n^{-1}\sum_{i=1}^n \int_0^\tau \big\{Z_i(u) - \bar{Z}(u;\gamma)\big\} e^{\delta' Z_i(t)} I(u \geq t) dM_{ik}^*(u)$, and $Q_3(t;Z_i) = \int_0^t \big\{Z_i(u) - \frac{r^{(1)}(u)}{r^{(0)}(u)}\big\} e^{\delta' Z_i(u)} d\Delta_0(u)$.

Combining the results in steps (1)-(4), we have

$$U_1(\theta_0;\hat{\gamma}) = \sum_{i=1}^n \sum_{k=1}^K \Big[\int_0^\tau W(t)\big\{X_{ik}(t) - \bar{x}_k(t)\big\} dM_{ik}(t) - \int_0^\tau \Big\{ \frac{W(t)\Upsilon_k(t)}{s^{(0)}(t)} + A_\gamma \Omega_\gamma^{-1}\big(Z_i(t) - \bar{z}(t)\big)\Big\}$$

$$dM_{ik}^*(t)\Big] - \sum_{i=1}^n \sum_{k=1}^K \int_0^\tau \Big[\frac{B_{1k}(t)}{r^{(0)}(t)} + B_{2k}\Big\{Z_i(t) - \frac{r^{(1)}(t)}{r^{(0)}(t)}\Big\} + A_\gamma \Omega_\gamma^{-1}\Big\{ Q_{1k}\big(Z_i(t) - \frac{r^{(1)}(t)}{r^{(0)}(t)}\big) + \frac{Q_{2k}(t)}{r^{(0)}(t)}\Big\} dM_i^d(t)$$

Then it follows from the multivariate central limit theorem that the conclusion holds.

B.3 Proof of the Null Distribution of $\mathcal{F}(t,x)$ in Chapter 3

Let $V(\hat{\theta},\hat{\gamma}) = \sum_{i=1}^n \sum_{k=1}^K \int_0^t I(X_{ik}(u) \leq x) d\hat{M}_{ik}(u;\hat{\theta},\hat{\gamma})$. Then the Taylor series expansion gives

$$\mathcal{F}(t,x;\hat{\theta},\hat{\gamma}) = n^{-1/2} V(\theta_0,\gamma_0) + \frac{\partial V(\theta_0,\gamma_0)}{n\partial\gamma'}\sqrt{n}(\hat{\gamma} - \gamma_0) + \frac{\partial V(\theta_0,\hat{\gamma})}{n\partial\theta'}\sqrt{n}(\hat{\theta} - \theta_0) + o_p(1).$$

Using the arguments and algebra manipulations similar to those in Appendix B.2, one can show that $V(\theta_0,\gamma_0) = \sum_{i=1}^n \eta_{1i}(t,x)$. Note that one can estimate $n^{-1}\partial V(\theta_0,\gamma_0)/\partial\gamma'$ and $n^{-1}\partial V(\theta_0,\hat{\gamma})/\partial\theta'$ by $-\hat{\Phi}_\gamma(t,x)$ and $-\hat{\Phi}_\theta(t,x)$, respectively. It then follows from

(B.2) that

$$\sqrt{n}(\hat{\gamma} - \gamma_0) = \Omega_\gamma^{-1} n^{-1/2} \sum_{i=1}^{n} \eta_{2i} + o_p(1)\,.$$

Also it follows from Theorem 3.1 that

$$\sqrt{n}(\hat{\theta} - \theta_0) = A_\theta^{-1} n^{-1/2} \sum_{i=1}^{n} (\xi_{1i} - \xi_{2i} - \xi_{3i}) + o_p(1)\,.$$

This indicates that $\mathcal{F}(t, x; \hat{\theta}, \hat{\gamma})$ can be expressed as a sum of i.i.d. zero-mean terms for fixed t and thus by the multivariate central limit theorem, $\mathcal{F}(t, x)$ converges in finite-dimensional distributions to a zero-mean Gaussian process. Since $\mathcal{F}(t, x)$ is tight based on the empirical process theory, $\mathcal{F}(t, x)$ converges weakly to a zero-mean Gaussian process that can be approximated by the zero-mean Gaussian process $\hat{\mathcal{F}}(t, x)$ given by Equation (3.9).

Appendix C

C.1 The Asymptotic Distribution of $\phi(\hat{\gamma})$ in Chapter 4

To derive the asymptotic distribution of $\phi(\hat{\gamma})$, we need the following regularity conditions:

(C1). $\{N_i(\cdot), Y_i(\cdot), C_i, Z_i\}_{i=1}^n$ are independent and identically distributed.

(C2). There exists a $\tau > 0$ such that $P(C_i \geq \tau) > 0$.

(C3). Both $N_i(\tau)$ and $Y_i(\tau)$ $(i = 1, \ldots, n)$ are bounded.

(C4). $W(t)$ and Z_i, $i = 1, \ldots, n$, have bounded variations and $W(t)$ converges almost surely to a deterministic function $w(t)$ uniformly in $t \in [0, \tau]$.

(C5). $B_\gamma = E\Big[\int_0^\tau \big\{ Z_1 - \frac{s^{(1)}(t,\gamma_0)}{s^{(0)}(t,\gamma_0)} \big\}^2 I(C \geq t | Z_1) e^{\gamma_0 Z_1} \lambda_0(t) dt \Big]$ is positive definite.

Now consider $\phi(\hat{\gamma})$, which can be written as

$$\phi(\hat{\gamma}) = \sum_{l=1}^{m} \phi_l(\hat{\gamma}),$$

where

$$\phi_l(\hat{\gamma}) = n^{\frac{1}{2}} \int_0^\tau W(t) K_l d\hat{\mu}_l(t;\hat{\gamma}), \quad l = 1, \ldots, m. \tag{C.1}$$

At the true value γ_0, $\phi_l(\gamma_0)$, $l = 1, \ldots, m$ are independent and all asymptotically normal, so that $\phi(\gamma_0)$ has mean 0 when $\tilde{H}_0$ is true. With respect to $\phi_l(\hat{\gamma})$, it follows from the definition of $\hat{\mu}_l(t;\hat{\gamma})$ that

$$\phi_l(\hat{\gamma}) = \frac{\sqrt{n}}{n_l} \sum_{i \in S_l} \int_0^\tau W(t) K_l \frac{Y_i(t) dN_i(t)}{\exp(\hat{\gamma} Z_i)} \quad l = 1, ...m.$$

Especially when $l = m$ for the control group,

$$\phi_m(\hat{\gamma}) = \frac{\sqrt{n}}{n_m} \sum_{i \in S_m} \int_0^\tau W(t) K_m Y_i(t) dN_i(t) \triangleq \phi_m,$$

which does not involve $\hat{\gamma}$ since $Z_i = 0$ for $i \in S_m$. Then we apply Taylor series expansion to $\phi_l(\hat{\gamma})$ $(l = 1, \ldots, m-1)$ in $\phi(\hat{\gamma})$,

$$\phi_l(\hat{\gamma}) = \phi_l(\gamma_0) + A_{l,\gamma} B_\gamma^{-1} U(\gamma_0).$$

In the above, γ_0 denotes the true value of γ, $A_{l,\gamma} = \lim_{n\to\infty} A_l(\gamma_0)$, $B_\gamma = \lim_{n\to\infty} B(\gamma_0)$,

$$A_l(\gamma) = \frac{1}{\sqrt{n}} \frac{\partial \phi_l(\gamma)}{\partial \gamma} = -\frac{1}{n_l} \sum_{i \in S_l} \int_0^\tau W(t) K_l \frac{Z_i Y_i(t) dN_i(t)}{\exp(\gamma Z_i)},$$

$$B(\gamma) = -\frac{1}{\sqrt{n}}\frac{\partial U(\gamma)}{\partial \gamma} = \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \frac{S^{(2)}(t,\gamma)S^{(0)}(t,\gamma) - S^{2(1)}(t,\gamma)}{S^{2(0)}(t,\gamma)} dN_i(t),$$

$S^{(r)}(t,\gamma) = \sum_{j=1}^{n} I(t \le C_j)\exp(\gamma Z_j) Z_j^r, \quad r = 0, 1, 2.$

Also by simple manipulation and the expression of $U(\gamma)$ in (4.5),

$$U(\gamma_0) = n^{-1/2}\sum_{i=1}^{n}\int_0^\tau \Big\{ Z_i - \frac{S^{(1)}(t,\gamma_0)}{S^{(0)}(t,\gamma_0)}\Big\} dM_i(t;\gamma_0),$$

where

$$dM_i(t;\gamma_0) = dN_i(t) - I(t \le C_i)\exp(\gamma_0 Z_i)\lambda_0(t)dt, \quad i = 1, \ldots, n,$$

are mean-zero stochastic processes under model (4.1). Then asymptotically,

$$\begin{aligned}
\phi(\hat{\gamma}) &= \sum_{l=1}^{m}\phi_l(\hat{\gamma}) \\
&= \sum_{l=1}^{m-1}\phi_l(\gamma_0) + A_\gamma B_\gamma^{-1} U(\gamma_0) + \phi_m \\
&= \sqrt{n}\Big(\sum_{l=1}^{m-1}\frac{K_l}{n_l}\sum_{i\in S_l} a_i + \frac{K_m}{n_m}\sum_{i\in S_m} a_i + \frac{1}{n}A_\gamma B_\gamma^{-1}\sum_{i=1}^{n} b_i\Big) \\
&= \sum_{l=1}^{m-1}\frac{1}{\sqrt{n_l}}\sum_{i\in S_l}\Big(K_l\sqrt{\frac{n}{n_l}}a_i + \sqrt{\frac{n_l}{n}}A_\gamma B_\gamma^{-1} b_i\Big) + \frac{1}{\sqrt{n_m}}\sum_{i\in S_m}\Big(K_m\sqrt{\frac{n}{n_m}}a_i + \sqrt{\frac{n_m}{n}}A_\gamma B_\gamma^{-1} b_i\Big),
\end{aligned} \tag{C.2}$$

where $a_i = \int_0^\tau W(t)\frac{Y_i(t)dN_i(t)}{\exp(\gamma_0 Z_i)}$, and $b_i = \int_0^\tau \Big\{ Z_i - \frac{s^{(1)}(t,\gamma_0)}{s^{(0)}(t,\gamma_0)}\Big\} dM_i(t;\gamma_0)$, for $i = 1, \ldots, n$,
$A_\gamma = \lim_{n\to\infty} A(\gamma)$, $A(\gamma) = \sum_{l=1}^{m-1} A_l(\gamma)$ and $s^{(r)}(t,\gamma_0) = \lim_{n\to\infty} S^{(r)}(t,\gamma_0)$, $r = 0, 1$.

For univariate cases, by the multivariate central limit theorem and some arguments similar as those in Lin et al. (2000) (Appendix A.2.), $\phi(\hat{\gamma})$ is asymptotically normal with mean 0 and the variance that can be consistently estimated by

$$\hat{\sigma}_\phi^2 = \sum_{l=1}^{m} H_l(\hat{\gamma})\hat{\Gamma}_l H_l(\hat{\gamma})',$$

where $H_l(\hat{\gamma}) = (K_l\sqrt{\frac{n}{n_l}} \quad \sqrt{\frac{n_l}{n}}A(\hat{\gamma})B^{-1}(\hat{\gamma}))$, $\hat{\Gamma}_l = n_l^{-1}\sum_{i\in S_l}\binom{\hat{a}_i}{\hat{b}_i}(\hat{a}_i \;\; \hat{b}_i)$,
$\hat{a}_i = \int_0^\tau W(t)\{d\tilde{Y}_i(t;\hat{\gamma}) - d\hat{\mu}_l(t;\hat{\gamma})\}$ and $\hat{b}_i = \int_0^\tau \left\{Z_i - \frac{S^{(1)}(t,\hat{\gamma})}{S^{(0)}(t,\hat{\gamma})}\right\} d\hat{M}_i(t;\hat{\gamma})$,
with

$$d\hat{M}_i(t;\hat{\gamma}) = dN_i(t) - I(t \le C_i)\exp(\hat{\gamma}Z_i)d\hat{\Lambda}_0(t) \tag{C.3}$$

and the Aalen-Breslow-type estimator for the true cumulative baseline function $\Lambda_0(t) = \int_0^t \lambda_0(u)du$,

$$\hat{\Lambda}_0(t) = \sum_{i=1}^{n}\int_0^t \frac{dN_i(u)}{S^{(0)}(u,\hat{\gamma})}.$$

Bibliography

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. (1972). *Statistical Inference under Order Restrictions.* New York: Wiley.

Balakrishnan, N. and Zhao, X. (2009). New multi-sample nonparametric tests for panel count data. *Annals of Statistics* **37**, 1112-1149.

Balakrishnan, N. and Zhao, X. (2010). A nonparametric test for the equality of counting processes with panel count data. *Computational Statistics and Data Analysis* **54**, 135-142.

Balakrishnan, N. and Zhao, X. (2011). A class of multi-sample nonparametric tests for panel count data. *Annals of the Institute of Statistical Mathematics* **63**, 135-156.

Cheng, S. C. and Wei, L. J. (2000). Inferences for a semiparametric model with panel data. *Biometrika* **87**, 89-97.

Cook, R. J., Lawless, J. F. and Nadeau, C. (1996). Robust tests for treatment comparisons based on recurrent event responses. *Biometrics* **52**, 557-571.

Cook, R. J. and Lawless, J. F. (1997). Marginal analysis of recurrent events and a terminating event. *Statistics in Medicine* **16**, 911-924.

Cook, R. J. and Lawless, J. F. (2007). *The Statistical Analysis of Recurrent Events*. Springer, New York.

Ghosh, D. and Lin, D. Y. (2002). Marginal regression models for recurrent and terminal events. *Statistica Sinica* **12**, 663-688.

Hu, X. J., Sun J. and Wei L. J. (2003). Regression parameter estimation from panel counts. *Scandinavian Journal of Statistics* **30**, 25-43.

He, X., Tong, X. and Sun, J. (2009). Semiparametric analysis of panel count data with correlated observation and follow-up times. *Lifetime Data Analysis* **15**, 177-196.

Hu X. J., Lagakos, S. W. and Lockhart R. A. (2009a). Marginal analysis of panel counts through estimating functions. *Biometrika* **96**, 445-456.

Hu, X. J., Lagakos, S. W. and Lockhart, R. A. (2009b). Generalized least squares estimation of the mean function of a counting process based on panel counts. *Statistica Sinica* **19**, 561-580.

Huang, C. Y. and Wang, M. C. (2004). Joint modeling and estimation of recurrent event processes and failure time. *Journal of the American Statistical Association* **99**, 1153-1165.

Huang, C. Y., Wang, M. C., and Zhang, Y. (2006). Analysing panel count data with informative observation times. *Biometrika* **93**, 763-775.

James L. F. (2003). Bayesian calculus for gamma processes with applications to semiparametric intensity models. *Sankhya, Series A* **65**, 196-223.

Kalbfleisch, J. D. and Lawless, J. F. (1985). The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association* **80**, 863-871.

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

Kim, Y. J. (2007). Analysis of panel count data with measurement errors in the covariates. *Journal of Statistical Computation and Simulation* **77**, 109-117.

Li, N., Sun, L. Q. and Sun, J. (2010). Semiparametric transformation models for panel count data with dependent observation process. *Statistics in Bioscience* **2**, 191-210.

Li, N., Park, D. H., Sun, J., and Kim, K. (2011). Semiparametric transformation models for multivariate panel count data with dependent observation process. *Canadian Journal of Statistics* **39**, 458-474.

Liang, Y., Lu, W. B. 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 functions of recurrent events. *Journal of the Royal Statistical Society, Series B* **62**, 711-730.

Lin, D. Y., Wei, L. J. and Ying, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika* **80**, 557-572.

Lin, D. Y., Wei, L. J. and Ying, Z. (2001). Semiparametric transformation models for point processes. *Journal of the American Statistical Association* **96**, 620-628.

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

Liu, L., Wolfe, R. A. and Huang, X. (2004). Shared frailty models for recurrent events and a terminal event. *Biometrics* **60**, 747-756.

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.

Luo, X. H. and Huang, C. Y. (2010). A comparison of various rate functions of a recurrent event process in the presence of a terminal event. *Statistical Methods in Medical Research* **19**, 167-182.

Luonga, A. and Doray, L. (2009). Inference for the positive stable laws based on a special quadratic distance. *Statistical Methodology* **6**, 147-156.

Park, D. H., Sun, J. and Zhao, X. (2007). A class of two-sample nonparametric tests for panel count data. *Communications in Statistics - Theory and Methods* **36,** 1611-1625.

Robertson, T., Wright, F. T. and Dykstra, R. (1988). *Order Restricted Statistical Inference.* New York: Wiley.

Schoenfield, L. J., Lachin, J. M., the Steering Committee and the NCGS Group (1981). Chenodiol (Chenodeoxycholic Acid) for Dissolution of Gallstones: The National Cooperative Gallstone StudyA Controlled Trial of Efficacy and Safety. *Annals of Internal Medicine* **95**(3), 257-282.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika* **69**, 239-241.

Sun, J. (1999). A nonparametric test for current status data with unequal censoring. *Journal of the Royal Statistical Society, Series B* **61**, 243-250.

Sun, J. and Fang, H. B. (2003). A nonparametric test for panel count data. *Biometrika* **90**, 199-208.

Sun, J. and Kalbfleisch, J. D. (1993). The analysis of current status data on point processes. *Journal of American Statistical Association* **88**, 1149-1154.

Sun, J. and Kalbfleisch, J. D. (1995). Estimation of the mean function of point processes based on panel count data. *Statistica Sinica* **5**, 279-289.

Sun, J., Park D., Sun L. Q. 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. Q. and Liu, D. (2007a). Regression Analysis of Longitudinal Data in the Presence of Informative Censoring and Observation Times. *Journal of the American Statistical Association* **102**, 1397-1406.

Sun, J., Tong, X. and He, X. (2007b). Regression analysis of panel count data with dependent observation times. *Biometrics* **63**, 1053-1059.

Sun, J. and Wei, L. J. (2000). Regression analysis of panel count data with covariate-dependent observation and censoring times. *Journal of the Royal Statistical Society, Series B* **62**, 293-302.

Thall, P. F. and Lachin, J. M. (1988). Analysis of recurrent events: nonparametric methods for random-interval count data. *Journal of American Statistical Association* **83**, 339-347.

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

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.

Wellner, J. A. and Zhang, Y. (2007). Two likelihood-based semiparametric estimation methods for panel count data with covariates. *Annals of Statistics* **35**, 2106-2142.

Ye, Y. N., Kalbfleisch, J. D. and Schaubel, D. E. (2007). Semiparametric analysis of correlated recurrent and terminal events. *Biometrics* **63**, 78-87.

Zhao, X., Balakrishnan, N. and Sun, J. (2011). Nonparametric inference based on panel count data. *Test* **20**, 1-42.

Zhao, X. and Tong, X. (2011). Semiparametric regression analysis of panel count data with informative observation times. *Computational Statistics and Data Analysis* **55**(1), 291-300.

Zhao, X. and Sun, J. (2011). Nonparametric comparison for panel count data with unequal observation processes. *Biometrics* **67**, 770-779.

Zhao, X., Zhou, J. and Sun, L. Q. (2011). Semiparametric transformation models with time-varying coefficients for recurrent and terminal events. *Biometrics* **67**, 404-414.

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

Zhang, Y. (2006). Nonparametric k-sample tests with panel count data. *Biometrika* **93**, 777-790.

Zhou H. and Pepe, M. S. (1995). Auxiliary covariate data in failure time regression. *Biometrika* **82**, 139-149.

Zhou H. and Wang C. Y. (2000). Failure time regression analysis with measurement error in covariates. *Journal of Royal Statistics Society, Series B* **62**, 657-665.

Zeng, D. and Cai, J. W. (2010). A semiparametric additive rate model for recurrent events with an informative terminal event. *Biometrika* **97**, 699-712.

Table 2.1. Results for estimation of β and α with $g(t) = t$ and $\mu_0(t) = t$

	$n = 200$				$n = 300$			
	$\rho = 1$		$\rho = 0.7$		$\rho = 1$		$\rho = 0.7$	
	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$
$\theta = (0, 0)$								
BIAS	0.0070	-0.0064	-0.0987	0.0013	0.0062	-0.0039	-0.0001	-0.0036
SEE	0.2853	0.0381	0.3178	0.0472	0.2328	0.0310	0.2727	0.0331
SSE	0.2812	0.0413	0.3191	0.0478	0.2210	0.0328	0.2746	0.0360
CP	0.942	0.924	0.950	0.944	0.964	0.924	0.946	0.926
$\theta = (0.5, 0)$								
BIAS	-0.0073	-0.0041	0.0304	0.0027	0.0061	-0.0036	-0.0021	-0.0035
SEE	0.2559	0.0323	0.2856	0.0403	0.2025	0.0295	0.2396	0.0337
SSE	0.2700	0.0332	0.2907	0.0408	0.1862	0.0299	0.2419	0.0352
CP	0.932	0.943	0.960	0.920	0.960	0.950	0.956	0.922
$\theta = (0, 0.1)$								
BIAS	-0.0068	-0.0062	-0.0183	-0.0063	-0.0083	-0.0028	-0.0054	-0.0063
SEE	0.2028	0.0220	0.2597	0.0324	0.1666	0.0185	0.2125	0.0269
SSE	0.2006	0.0242	0.2639	0.0369	0.1653	0.0202	0.2096	0.0295
CP	0.952	0.908	0.948	0.910	0.954	0.914	0.944	0.918
$\theta = (0.5, 0.1)$								
BIAS	-0.0216	-0.0040	-0.0072	-0.0052	-0.0203	-0.0046	-0.0083	-0.0037
SEE	0.1900	0.0231	0.2421	0.0303	0.1562	0.0189	0.1994	0.0256
SSE	0.1938	0.0265	0.2357	0.0350	0.1511	0.0207	0.1812	0.0286
CP	0.934	0.906	0.948	0.900	0.952	0.910	0.956	0.912

Table 2.2. Results for estimation of β and α with $g(t) = \log(t)$ and $\mu_0(t) = e^t$

	$n = 200$				$n = 300$			
	$\rho = 1$		$\rho = 0.7$		$\rho = 1$		$\rho = 0.7$	
	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$
$\theta = (0, 0)$								
BIAS	-0.0053	-0.0026	-0.0063	-0.0008	-0.0055	-0.0020	-0.0055	-0.0008
SEE	0.0845	0.0196	0.0927	0.0224	0.0700	0.0153	0.0776	0.0161
SSE	0.0846	0.0202	0.0916	0.0256	0.0703	0.0165	0.0749	0.0172
CP	0.962	0.930	0.964	0.914	0.958	0.910	0.960	0.928
$\theta = (0.5, 0)$								
BIAS	0.0026	-0.0034	-0.0126	-0.0019	-0.0063	-0.0025	0.0003	-0.0019
SEE	0.1182	0.0272	0.1397	0.0278	0.0964	0.0229	0.1159	0.0233
SSE	0.1176	0.0298	0.1348	0.0288	0.1028	0.0241	0.1082	0.0253
CP	0.948	0.912	0.954	0.918	0.938	0.934	0.950	0.936
$\theta = (0, 0.1)$								
BIAS	-0.0076	-0.0031	-0.0066	-0.0037	-0.0032	-0.0058	-0.0033	-0.0029
SEE	0.1306	0.0364	0.1606	0.0425	0.1088	0.0307	0.1298	0.0374
SSE	0.1319	0.0391	0.1569	0.0481	0.1144	0.0329	0.1270	0.0409
CP	0.964	0.912	0.954	0.904	0.934	0.914	0.958	0.912
$\theta = (0.5, 0.1)$								
BIAS	-0.0107	-0.0048	-0.0202	-0.0076	-0.0046	-0.0037	-0.0190	0.0010
SEE	0.1626	0.0403	0.1959	0.0516	0.1339	0.0336	0.1725	0.0420
SSE	0.1577	0.0414	0.1917	0.0543	0.1335	0.0350	0.1652	0.0445
CP	0.956	0.936	0.950	0.912	0.956	0.934	0.962	0.926

Table 2.3. Estimation results with $h(\mathcal{F}_{it}) = N_i(t-)$ for the bladder tumor study

$g(t)$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\alpha}$
	95% CI for $\hat{\beta}_1$	95% CI for $\hat{\beta}_2$	95% CI for $\hat{\alpha}$
	p-value for $\hat{\beta}_1$	p-value for $\hat{\beta}_2$	p-value for $\hat{\alpha}$
$g(t) = t$	-1.8955	0.2961	0.0398
	(-2.6442, -1.1467)	(0.1487, 0.4436)	(-0.0086, 0.0883)
	< 0.001	< 0.001	0.1074
$g(t) = t^2$	-0.9474	0.1481	0.0199
	(-1.3217, -0.5731)	(0.0743, 0.2218)	(-0.0043, 0.0441)
	< 0.001	< 0.001	0.1075
$g(t) = \log t$	-4.0501	0.8464	0.0352
	(-5.9544, -2.1459	(0.2636, 1.4292)	(-0.1260, 0.1964)
	< 0.001	0.0044	0.6683

Table 2.4. Estimation results with $h(\mathcal{F}_{it}) = N_i(t-) - N_i(t-6)$ for the bladder tumor study

$g(t)$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\alpha}$
	95% CI for $\hat{\beta}_1$	95% CI for $\hat{\beta}_2$	95% CI for $\hat{\alpha}$
	p-value for $\hat{\beta}_1$	p-value for $\hat{\beta}_2$	p-value for $\hat{\alpha}$
$g(t) = t$	-1.6750	0.2901	0.0764
	(-2.3786, -0.9713)	(0.1483, 0.4318)	(-0.0639, 0.2165)
	< 0.001	< 0.001	0.2858
$g(t) = t^2$	-0.8373	0.1450	0.0382
	(-1.1890, -0.4854)	(0.0742, 0.2159)	(-0.0319, 0.1083)
	< 0.001	< 0.001	0.2861
$g(t) = \log t$	-4.1338	0.8492	0.2189
	(-6.2092, -2.0584)	(0.2780, 1.4205)	(-0.0703, 0.5080)
	< 0.001	0.0036	0.1379

Table 3.1. Results for estimation of β and α with $\mu_{01}(t) = \mu_{02}(t) = t$, $\Lambda_{01}(t) = \Lambda_{02}(t) = 10t$, $h_1(\mathcal{F}_{i1,t}) = N_{i1}(t-)$, $h_2(\mathcal{F}_{i2,t}) = N_{i2}(t-)$

	n=200				n=300			
	$\rho = 1$		$\rho = 0.7$		$\rho = 1$		$\rho = 0.7$	
	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$
$\theta = (0, 0)$								
BIAS	-0.0084	-0.0024	-0.0057	0.0002	-0.0086	-0.0012	-0.0042	-0.0002
SEE	0.0629	0.0136	0.0745	0.0149	0.0523	0.0113	0.0608	0.0121
SSE	0.0596	0.0148	0.0703	0.0149	0.0500	0.0119	0.0594	0.0128
CP	0.952	0.912	0.964	0.938	0.964	0.934	0.952	0.922
$\theta = (0.5, 0)$								
BIAS	-0.0025	-0.0031	-0.0017	-0.0007	0.0089	-0.0012	-0.0062	-0.0012
SEE	0.0919	0.0196	0.1159	0.0219	0.0753	0.0160	0.0945	0.0176
SSE	0.0911	0.0215	0.1057	0.0226	0.0716	0.0178	0.0887	0.0174
CP	0.946	0.904	0.952	0.94	0.964	0.912	0.96	0.942
$\theta = (0, 0.2)$								
BIAS	-0.0159	-0.0046	-0.0237	-0.0011	-0.0067	-0.0040	-0.0156	-0.0002
SEE	0.1532	0.0405	0.2131	0.0527	0.1241	0.0343	0.1785	0.0452
SSE	0.1474	0.0439	0.1858	0.0539	0.1173	0.0366	0.1667	0.0455
CP	0.966	0.908	0.970	0.926	0.97	0.920	0.968	0.936
$\theta = (0.5, 0.2)$								
BIAS	-0.0100	-0.0100	-0.0094	0.0023	-0.0079	-0.0050	-0.0233	-0.0007
SEE	0.1731	0.0438	0.2480	0.0582	0.1411	0.0369	0.2070	0.0492
SSE	0.1691	0.0466	0.2327	0.0627	0.1345	0.0379	0.2000	0.0526
CP	0.96	0.916	0.962	0.914	0.968	0.924	0.968	0.934

Table 3.2. Results for estimation of β and α with $\mu_{01}(t) = \sqrt{t}$, $\mu_{02}(t) = t$, $\Lambda_{01}(t) = 8t$, $\Lambda_{02}(t) = 12t$, $h_1(\mathcal{F}_{i1,t}) = N_{i1}(t-)$, $h_2(\mathcal{F}_{i2,t}) = N_{i2}(t-)$

	n=200				n=300			
	$\rho = 1$		$\rho = 0.7$		$\rho = 1$		$\rho = 0.7$	
	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$
$\theta = (0, 0)$								
BIAS	-0.0038	-0.0017	-0.0030	-0.0015	-0.0029	-0.0014	0.0019	-0.0013
SEE	0.0742	0.0146	0.0881	0.0154	0.0611	0.0122	0.0725	0.0128
SSE	0.0740	0.0145	0.0857	0.0173	0.0610	0.0123	0.0708	0.0136
CP	0.958	0.952	0.948	0.916	0.950	0.934	0.964	0.914
$\theta = (0.5, 0)$								
BIAS	0.0040	-0.0044	-0.0120	-0.0010	0.0059	-0.0019	-0.0058	0.0002
SEE	0.1014	0.0200	0.1273	0.0220	0.0831	0.0167	0.1062	0.0184
SSE	0.1003	0.0199	0.1187	0.0239	0.0762	0.0167	0.1022	0.0201
CP	0.950	0.930	0.948	0.932	0.966	0.946	0.962	0.924
$\theta = (0, 0.2)$								
BIAS	-0.0181	-0.0046	0.0125	-0.0056	-0.0130	-0.0038	-0.0073	0.0005
SEE	0.1621	0.0424	0.2335	0.0567	0.1323	0.0350	0.1934	0.0488
SSE	0.1522	0.0460	0.2260	0.0627	0.1314	0.0360	0.1853	0.0490
CP	0.972	0.924	0.964	0.922	0.956	0.936	0.958	0.940
$\theta = (0.5, 0.2)$								
BIAS	-0.0072	-0.0058	-0.0158	-0.0013	0.0091	-0.0067	-0.0065	-0.0041
SEE	0.1552	0.0399	0.2266	0.0527	0.1493	0.0381	0.2195	0.0511
SSE	0.1506	0.0437	0.2179	0.0555	0.1434	0.0417	0.2007	0.0539
CP	0.962	0.93	0.974	0.942	0.964	0.916	0.970	0.938

Table 3.3. Results for estimation of β and α with $\mu_{01}(t) = \sqrt{t}$, $\mu_{02}(t) = t$, $\Lambda_{01}(t) = 8t$, $\Lambda_{02}(t) = 12t$, $h_1(\mathcal{F}_{i1,t}) = N_{i1}(t-)$, $h_2(\mathcal{F}_{i2,t}) = N_{i2}(t-) - N_{i2}(t-0.5)$

	n=200				n=300			
	$\rho = 1$		$\rho = 0.7$		$\rho = 1$		$\rho = 0.7$	
	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$
$\theta = (0, 0)$								
BIAS	-0.0026	-0.0021	-0.0022	-0.0005	0.0025	-0.0020	-0.0048	0.0003
SEE	0.0745	0.0145	0.0890	0.0157	0.0612	0.0122	0.0620	0.0120
SSE	0.0703	0.0158	0.0847	0.0167	0.0589	0.0124	0.0599	0.0124
CP	0.962	0.926	0.974	0.924	0.952	0.944	0.954	0.93
$\theta = (0.5, 0)$								
BIAS	-0.0038	-0.0018	-0.0044	0.0002	-0.0020	-0.0015	-0.0021	-0.0008
SEE	0.0923	0.0195	0.1277	0.0222	0.0758	0.0163	0.0954	0.0177
SSE	0.0908	0.0213	0.1211	0.0241	0.0755	0.0165	0.0865	0.0184
CP	0.942	0.910	0.960	0.930	0.956	0.932	0.954	0.940
$\theta = (0, 0.2)$								
BIAS	0.0208	0.0002	0.0165	0.0014	0.0015	-0.0018	0.0105	-0.0002
SEE	0.1585	0.0408	0.2285	0.0527	0.1302	0.0339	0.1875	0.0445
SSE	0.1517	0.0427	0.2126	0.0574	0.1251	0.0353	0.1743	0.0479
CP	0.950	0.934	0.974	0.900	0.954	0.940	0.968	0.902
$\theta = (0.5, 0.2)$								
BIAS	0.0172	-0.0050	0.0160	0.0023	0.0100	-0.0049	-0.0050	0.0024
SEE	0.1779	0.0443	0.2610	0.0588	0.1461	0.0370	0.2151	0.0497
SSE	0.1637	0.0475	0.2456	0.0628	0.1332	0.0378	0.1917	0.0510
CP	0.966	0.922	0.968	0.922	0.964	0.938	0.970	0.932

Table 3.4. Results for estimation of β and α with $\mu_{01}(t) = \mu_{02}(t) = exp(t^2)$, $\Lambda_{01}(t) = \Lambda_{02}(t) = 8t$, $h_1(\mathcal{F}_{i1,t}) = N_{i1}(t-) - N_{i1}(t-0.75)$, $h_2(\mathcal{F}_{i2,t}) = N_{i2}(t-) - N_{i2}(t-0.75)$

	n=200				n=300			
	$\rho = 1$		$\rho = 0.7$		$\rho = 1$		$\rho = 0.7$	
	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$
$\theta = (0, 0)$								
BIAS	0.0055	-0.0043	0.0060	0.0025	0.0003	-0.0069	0.0285	-0.0006
SEE	0.1501	0.0305	0.2062	0.0350	0.1223	0.0252	0.1691	0.0288
SSE	0.1398	0.0319	0.1947	0.0391	0.1142	0.0262	0.1497	0.0287
CP	0.970	0.934	0.956	0.916	0.966	0.930	0.970	0.936
$\theta = (-1, 0)$								
BIAS	-0.0062	-0.0059	0.0024	-0.0063	0.0097	-0.0054	0.0033	-0.0045
SEE	0.1122	0.0190	0.1499	0.0206	0.0917	0.0154	0.1232	0.0172
SSE	0.1078	0.0190	0.1380	0.0216	0.0884	0.0153	0.1226	0.0174
CP	0.964	0.928	0.970	0.930	0.944	0.936	0.954	0.940
$\theta = (0, 0.05)$								
BIAS	0.0089	-0.0082	0.0158	-0.0049	0.0022	-0.0056	0.0046	-0.0061
SEE	0.1635	0.0359	0.2305	0.0428	0.1342	0.0303	0.1876	0.0351
SSE	0.1551	0.0377	0.2105	0.0464	0.1279	0.0326	0.1754	0.0379
CP	0.964	0.914	0.958	0.904	0.956	0.922	0.964	0.916
$\theta = (-1, 0.05)$								
BIAS	-0.0041	-0.0041	0.0203	-0.0054	0.0030	-0.0050	0.0242	-0.0039
SEE	0.1276	0.0257	0.1748	0.0299	0.1047	0.0214	0.1425	0.0246
SSE	0.1204	0.0272	0.1750	0.0300	0.1015	0.0226	0.1297	0.0257
CP	0.966	0.932	0.944	0.926	0.958	0.930	0.958	0.924

Table 3.5. Results for estimation of β and α with $\mu_{01}(t) = \sqrt{t}$, $\mu_{02}(t) = t$, $E\{dN_{ik}(t)|Z\} = Q_i' e^{0.5Z_i} d\Lambda_{0k}(t)$, $\Lambda_{01}(t) = 8t$, $\Lambda_{02}(t) = 12t$, $h_1(\mathcal{F}_{i1,t}) = N_{i1}(t-)$, $h_2(\mathcal{F}_{i2,t}) = N_{i2}(t-)$

	n=200				n=300			
	$\rho = 1$		$\rho = 0.7$		$\rho = 1$		$\rho = 0.7$	
	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$
$\theta = (0, 0)$								
BIAS	-0.0098	0.0035	-0.0202	0.0040	-0.0191	0.0026	-0.0218	0.0041
SEE	0.0745	0.0147	0.0893	0.0157	0.0615	0.0122	0.0739	0.0132
SSE	0.0751	0.0159	0.0892	0.0166	0.0543	0.0127	0.0706	0.0142
CP	0.948	0.932	0.940	0.944	0.970	0.930	0.944	0.940
$\theta = (0.5, 0)$								
BIAS	-0.0223	0.0052	-0.0402	0.0074	-0.0191	0.0052	-0.0214	0.0045
SEE	0.1000	0.0204	0.1269	0.0223	-0.0191	0.0169	0.1053	0.0185
SSE	0.1026	0.0220	0.1195	0.0248	0.0808	0.0170	0.1037	0.0196
CP	0.924	0.916	0.942	0.922	0.950	0.940	0.938	0.934
$\theta = (0, 0.2)$								
BIAS	-0.0574	0.0165	-0.0589	0.0192	-0.0565	-0.0176	-0.0562	0.0198
SEE	0.1658	0.0444	0.2417	0.0593	0.1360	0.0370	0.1958	0.0499
SSE	0.1587	0.0501	0.2374	0.0663	0.1386	0.0391	0.1862	0.0545
CP	0.954	0.926	0.954	0.922	0.928	0.930	0.952	0.940
$\theta = (0.5, 0.2)$								
BIAS	-0.0698	0.0172	-0.0696	0.0222	-0.0779	0.0202	–0.0702	0.0210
SEE	0.1824	0.0473	0.2702	0.0656	0.1512	0.0399	0.2220	0.0551
SSE	0.1724	0.0527	0.2412	0.0707	0.1322	0.0415	0.2095	0.0597
CP	0.950	0.926	0.964	0.922	0.956	0.930	0.948	0.928

Table 4.1. Estimated sizes and powers when $\Lambda_i(t) = 0.75t \exp(\gamma Z_i)$, $\mathrm{E}\{Y_{i,1}(t)|Q_i\} = Q_i\, \mu_1(t) \exp(\beta Z_i)$.

	$n_1 = n_2 = 50$		$n_1 = n_2 = 100$		$n_1 = n_2 = 50$		$n_1 = n_2 = 100$	
β	$\gamma = 0$	$\gamma = 0.2$	$\gamma = 0$	$\gamma = 0.2$	$\gamma = 0$	$\gamma = 0.2$	$\gamma = 0$	$\gamma = 0.2$
	$\mu_1(t) = 0.25t$				$\mu_1(t) = \log(1+t)$			
-0.1	0.123	0.135	0.184	0.178	0.096	0.111	0.168	0.170
-0.2	0.264	0.277	0.516	0.508	0.268	0.288	0.470	0.449
-0.3	0.547	0.561	0.830	0.820	0.508	0.502	0.775	0.782
0	0.047	0.053	0.042	0.056	0.051	0.054	0.053	0.045
0.1	0.128	0.119	0.182	0.168	0.105	0.132	0.162	0.177
0.2	0.340	0.337	0.556	0.558	0.299	0.312	0.501	0.508
0.3	0.607	0.667	0.878	0.887	0.535	0.576	0.863	0.873

Table 4.2. Estimated sizes and powers of the proposed multivariate test procedure when $\Lambda_i(t) = 0.75t\exp(\gamma Z_i)$, $\mathrm{E}\{Y_{i,1}(t)|Q_i\} = Q_i\,\mu_1(t)\exp(\beta Z_i)$, $\mathrm{E}\{Y_{i,2}(t)|Q_i\} = Q_i\,\mu_2(t)\exp(\beta Z_i)$.

	$n_1 = n_2 = 50$		$n_1 = n_2 = 100$		$n_1 = n_2 = 50$		$n_1 = n_2 = 100$	
β	$\gamma = 0$	$\gamma = 0.2$	$\gamma = 0$	$\gamma = 0.2$	$\gamma = 0$	$\gamma = 0.2$	$\gamma = 0$	$\gamma = 0.2$
	$\mu_1(t) = 0.25t$,$\mu_2(t) = 0.15t$				$\mu_1(t) = \mu_2(t) = \log(1+t)$			
-0.1	0.148	0.137	0.229	0.207	0.138	0.171	0.230	0.211
-0.2	0.400	0.399	0.642	0.630	0.396	0.429	0.631	0.672
-0.3	0.690	0.698	0.927	0.943	0.726	0.710	0.932	0.939
0	0.043	0.057	0.049	0.043	0.047	0.048	0.049	0.055
0.1	0.150	0.137	0.231	0.252	0.143	0.149	0.240	0.245
0.2	0.425	0.436	0.706	0.720	0.430	0.435	0.708	0.723
0.3	0.746	0.755	0.956	0.958	0.757	0.784	0.963	0.971

Table 4.3. Estimated sizes of the proposed test procedure and the procedure in Zhao and Sun (2011), when $\Lambda_i(t) = 0.75t\exp(\gamma Z_i)$, $\mathrm{E}\{Y_{i,1}(t)|Q_i\} = Q_i\,\log(1+t)\exp(\beta Z_i)$, C_i and $T_{i,j}$ generated in special schemes.

	Proposed		Zhao&Sun(2011)	
$\beta = 0$	$\gamma = 0$	$\gamma = 0.2$	$\gamma = 0$	$\gamma = 0.2$
		$n_1 = n_2 = 50$		
Scheme 1	0.049	0.056	0.077	0.067
Scheme 2	0.054	0.046	0.219	0.228
Scheme 3	0.051	0.048	0.213	0.228
		$n_1 = n_2 = 100$		
Scheme 1	0.041	0.044	0.060	0.058
Scheme 2	0.052	0.042	0.142	0.152
Scheme 3	0.045	0.048	0.149	0.141

Scheme 1: The censoring time C_i followed a uniform distribution from 0.5τ to τ; $T_{i,j}$ followed a discrete uniform distribution on $(0, 0.1, 0.2, \ldots, C_i)$.

Scheme 2: The censoring time C_i followed a uniform distribution from 0.5τ to τ; $T_{i,j}$ followed a discrete uniform distribution on $(0, 0.01, 0.02, \ldots, C_i)$.

Scheme 3: The censoring time C_i followed a uniform distribution from 0.8τ to τ; $T_{i,j}$ followed a discrete uniform distribution on $(0, 0.01, 0.02, \ldots, C_i)$.

Table 4.4. p-values for the effectiveness of DFMO treatment on non-melanoma skin cancers

Basal cell carcinoma	Squamous cell carcinoma	Overall
0.0231	0.583	0.0872

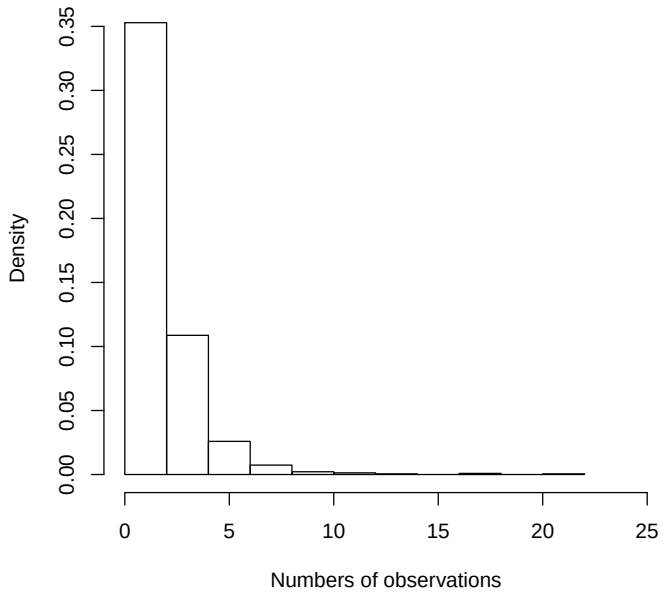

Figure 4.1. Distribution of the numbers of observation times

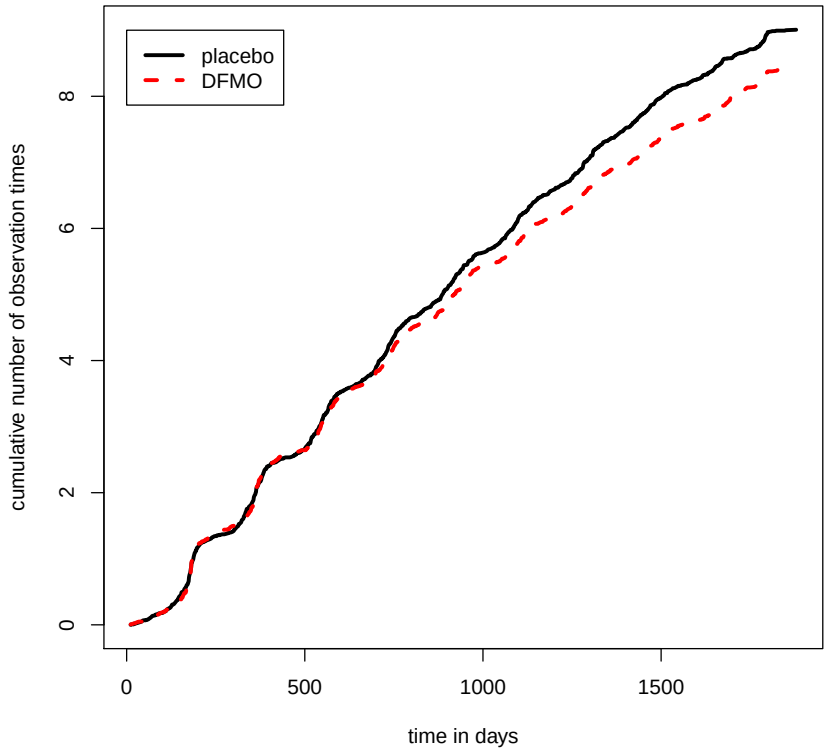

Figure 4.2. Estimated means of observation times for different groups

VITA

Yang Li was born on January 18, 1983, in Tonghua, Jilin Providence, People's Republic of China. After attending public schools in Changchun, she received her B.E. from Beijing Institute of Technology. In May 2005, she received her M.S. from University of Arkansas at Little Rock. She joined the graduate program in the Department of Statistics at the University of Missouri in August 2008. She will receive her Ph.D. in Statistics in May 2013. As of August 2013, she will be working as an Assistant Professor in the Department of Mathematics and Statistics at The University of North Carolina at Charlotte in Charlotte, North Carolina.