SEMIPARAMETRIC ANALYSIS OF PANEL COUNT DATA

A Dissertation

Presented to

the Faculty of the Graduate School

University of Missouri-Columbia

In Partial Fulfillment

Of the Requirements for the Degree

Doctor of Philosophy

---

by

XIN HE

Dr. (Tony) Jianguo Sun, Dissertation Supervisor

AUGUST 2007

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

SEMIPARAMETRIC ANALYSIS OF PANEL COUNT DATA

Presented by Xin He

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. Paul L. Speckman _____

Dr. Athanasios C. Micheas _____

Dr. Lori A. Thombs _____

Dr. (Shawn) Xiaoguang Ni _____

Dedicated to my parents,

He, Jianmin and He, Deping

ACKNOWLEDGEMENTS

I would like to acknowledge many people for helping me during my doctoral work. I am especially indebted to my esteemed advisor, Dr. (Tony) Jianguo Sun, for his critical inspiration, constant support and endless patience. His infectious enthusiasm and continual encouragement have been major driving forces through my study and research. This dissertation would not have been possible without his generous guidance.

I extend my gratitude to members of my exceptional doctoral committee: Drs. Paul Speckman, Athanasios Micheas, Lori Thombs and Xiaoguang Ni for their continuous support and insightful comments on my work. I owe a special note of gratitude to Dr. Xingwei Tong for his great assistance and helpful advice throughout my research process.

Many professors in the Department of Statistics assisted and encouraged me in various ways during my course of studies. I would like to express my deep gratitude to Dr. Nancy Flournoy for inspiring and helping me to pursue a career in Biostatistics. I would also like to express my appreciation to Drs. Chris Wikle, Dongchu Sun, Chong He, Min Yang and Scott Holan for all that they have taught me. I am very appreciative and thankful for the support of Dr. Larry Ries for helping me to be a better instructor.

I would like to take this opportunity to extend many thanks to my colleagues and friends in the Department of Statistics, especially Zhigang Zhang, Do-Hwan Park, Chao Zhu and Lianming Wang for their kind help and useful discussions.

Finally, I am particularly grateful to my parents whom I owe everything I am today. I thank them for always being at my side, listening to me and giving me support. I would also like to thank my fiancee and best friend, Qiongman Kong. Her love and support have given me a lot of courage and strength in all my work.

Table of Contents

ACKNOWLEDGEMENTS **ii**

LIST OF TABLES **vi**

LIST OF FIGURES **vii**

ABSTRACT **viii**

1 INTRODUCTION **1**

1.1 Longitudinal Data 1

1.1.1 Introduction 1

1.1.2 Some Examples 3

1.2 Analysis of Recurrent Event Data 6

1.3 Analysis of Panel Count Data 9

1.3.1 Inferences on the Mean Function 10

1.3.2 Regression Models for Panel Count Data 14

1.4 Counting Processes 18

1.4.1 Introduction 18

1.4.2 Counting Processes 19

1.4.3 Multivariate Counting Processes 20

1.4.4 Martingales and Stochastic Integrals 21

1.5 Outline of the Dissertation 25

2 REGRESSION ANALYSIS OF PANEL COUNT DATA WITH DEPENDENT OBSERVATION TIMES 27

2.1 Introduction 27

2.2 Models and Notation 30

2.3 Estimation Procedures 33

2.4 Numerical Results 39

2.5 An Illustrative Example 42

2.6 Discussion 44

3 REGRESSION ANALYSIS OF MULTIVARIATE PANEL COUNT DATA 46

3.1 Introduction 46

3.2 Models and Notation 49

3.3 Estimation Procedures 51

3.3.1 Estimation with Covariate-Independent Observation Processes . 51

3.3.2 Estimation with Covariate-Dependent Observation Processes . . 53

3.4 Simulation Studies 57

3.5 An Application 59

3.6 Concluding Remarks 62

4 SEMIPARAMETRIC ANALYSIS OF PANEL COUNT DATA WITH CORRELATED OBSERVATION AND FOLLOW-UP TIMES 65

4.1 Introduction 65

4.2 Models and Notation 67

4.3 Estimation of Regression Parameters 69

4.3.1 Estimation of Model (4.2) 69

4.3.2 Estimation of Model (4.3) 70

4.3.3 Estimation of Model (4.1) 73

4.4 Numerical Results 75
4.5 An Application 77
4.6 Discussion 78

5 FUTURE RESEARCH **80**
5.1 More Efficient Estimation for Regression Parameters 80
5.2 Regression Analysis of Multivariate Panel Count Data with Time-Dependent Covariates 81
5.3 Likelihood-Based Approach to the Analysis of Panel Count Data with Dependent Observation and Follow-up Times 81

APPENDIX **83**

BIBLIOGRAPHY **93**

VITA **116**

List of Tables

2.1 Estimation of β with a Homogeneous Observation Process 100

2.2 Estimation of β with a Nonhomogeneous Observation Process 101

2.3 Estimation of β with Covariate-Dependent Follow-up Times 102

3.1 Simulation Results for Covariate-Independent Observation Processes . . 103

3.2 Simulation Results for Covariate-Dependent Observation Processes . . 104

3.3 Results of Joint and Univariate Analyses of Radiological and Functional Joint Damage Data from the University of Toronto Psoriatic Arthritis Clinic 105

4.1 Estimation of β_1 with $\beta_2 = \beta_3 = 0$ 106

4.2 Estimation of β_1 with $\beta_2 = \beta_3 = 0.2$ 107

List of Figures

2.1 Quantile Plot with a Homogeneous Observation Process 108

2.2 Quantile Plot with a Nonhomogeneous Observation Process 109

2.3 Estimates of the Baseline Mean Functions 110

3.1 Timeline Diagram of Visits, Event Counts and Follow-up Durations for Sample of Patients in the University of Toronto Psoriatic Arthritis Clinic 111

3.2 Crude Event Rates for Radiological Damage Against Functional Damage 112

3.3 Estimated Baseline Cumulative Mean Functions from Univariate and Bivariate Regression Models Applied to Data from the University of Toronto Psoriatic Arthritis Clinic 113

4.1 Quantile Plot with $\beta_1 = 0$, $\beta_2 = 0$, and $\beta_3 = 0$ 114

4.2 Quantile Plot with $\beta_1 = 1$, $\beta_2 = 0$, and $\beta_3 = 0$ 115

SEMIPARAMETRIC ANALYSIS OF PANEL COUNT DATA

Xin He

Dr. (Tony) Jianguo Sun, Dissertation Supervisor

ABSTRACT

Statistical analysis of longitudinal data is required in a number of applied fields including econometrics, epidemiology, psychology, and public health. In these studies, one may be interested in some events, and subjects could experience the event of interest multiple times. One type of these studies monitors study subjects continuously and thus produces recurrent event data, which record the times of all occurrences of events. In the other type, study subjects are observed only at discrete time points, and only the numbers of occurrences of the events between observation times are known, resulting in so-called panel count data.

For the analysis of panel count data, most of the existing approaches assume that the observation and longitudinal response processes are independent and therefore rely on the conditional inference procedures given the observation times. In practice, however, this assumption may not be true. The first part of this dissertation will consider regression analysis of panel count data with dependent observation times. We introduce

a class of semiparametric models to characterize the possible correlation between the observation and response processes. The inference procedures are proposed based on estimating equations, and the asymptotic distributions of the resulting estimates are established. Some results from simulation studies are presented, and the method is applied to a bladder cancer study.

In the second part of this dissertation, we will focus on regression analysis of multivariate panel count data. In many applications, several recurrent event processes may be correlated, and appropriate dependence structures are difficult to accommodate. We present a class of marginal mean models that leave the dependence structures for the related types of recurrent events completely unspecified. Some estimating equations are developed for inference, and the resulting estimates of regression parameters are consistent and asymptotically normal. Simulation studies are conducted for practical situations, and the methodology is applied to a psoriatic arthritis study.

The last part of the dissertation considers the same problem studied in the first part and provides an approach that allows both observation and follow-up times to be correlated with the recurrent events of interest. Many researchers have developed the methods which assume that the follow-up time is noninformative, but this assumption may be violated in some clinical trials when the drop-out time is informative. We propose some shared frailty models that allow potential dependence among the response process, the observation process and the follow-up process. Estimating equations and an EM algorithm are developed for estimation of regression parameters. The proposed estimates are consistent and have asymptotically a normal distribution. Their finite sample properties are evaluated through simulation, and the method is used to reanalyze the data from the bladder cancer study.

Chapter 1

INTRODUCTION

1.1 Longitudinal Data

1.1.1 Introduction

Statistical analysis of longitudinal data is required in many applied fields including demographical studies, econometrics, epidemiology, public health, and reliability studies. One special type of longitudinal study is the event history study in which subjects experience some events of interest multiple times. The resulting data are usually referred to as event history data.

The event history studies can be generally classified into two types. One monitors study subjects continuously and thus produces recurrent event data (Byar, 1980; Prentice *et al.*, 1981; Pepe and Cai, 1993), which record the times of all occurrences of events. In the other type, study subjects are observed only at discrete time points, and only the numbers of occurrences of the events between observation times are known; they are often referred to as panel count data (Kalbfleisch and Lawless, 1985; Thall

and Lachin, 1988; Sun and Kalbfleisch, 1995).

Recurrent event data are frequently encountered in epidemiological and medical studies where the event could be adverse event of drugs, sickness leave from work, and hospitalization. For each subject, there could be multiple occurrences that are correlated and naturally ordered. The observation of recurrent events could be censored by loss to follow-up, end of the study, or a fatal event such as death.

Panel count data are often collected in periodic follow-up studies in which it may be impractical or not realistic to keep subjects under observation over the entire study period. Since the number of observation times and the observation times themselves usually vary across the subjects, there are two main components in panel count data. One is the set of observation times that can be regarded as realizations of an observation process, and the other is the set of observed counts for the underlying recurrent event process. The methods dealing with panel count data depend on the relationship between these two processes and the research interest. In some situations, there is only one observation time for each subject; for example, death or onset of disease; in this case panel count data are often referred to as current status data (Dinse and Lagakos, 1983; Diamond *et al.*, 1986; Diamond and McDonald, 1991; Sun and Kalbfleisch, 1993).

In the following, we present four examples to further illustrate the basic concepts and general structures of recurrent event data and panel count data.

1.1.2 Some Examples

1.1.2.1 CGD Study

The Chronic Granulomatous Disease (CGD) study is a placebo-controlled randomized trial of gamma interferon, which was conducted from 1988 to 1989 (Fleming and Harrington, 1991). A total of 128 patients participated in the study, and by the end of the study, 30 of 65 patients in the control group and 14 of 63 patients in the treatment group had experienced at least one infection. The objectives of the CGD study were to compare the treatments with respect to the occurrence rate of infection and assess the effect of covariates. Among others, Lin *et al.* (2000) analyzed the study and concluded that the gamma interferon was effective in reducing the rate of infection, and the patient's age at enrollment had significant effect.

1.1.2.2 National Cooperative Gallstone Study

The National Cooperative Gallstone Study (NCGS) is a 10-year, multicenter, double-blinded, placebo-controlled clinical trial on the use of the natural bile acid chenodeoxycholic acid (chenodiol) for the dissolution of cholesterol gallstones (Schoenfield *et al.*, 1981). The original study consists of 916 patients who were randomly assigned into each of the three treatments: high dose, low dose and placebo. A part of the resulting data was presented in Thall and Lachin (1988), including the successive visit times and the associated counts of episodes of nausea for 111 patients with floating gallstones in high-dose and placebo groups. During the study, patients were scheduled to return for

clinical observations at 1, 2, 3, 6, 9, and 12 months during the first year follow-up, but the actual visit or observation times differed from patient to patient. One of the primary objectives of the study was to test the difference of the two treatments with respect to the incidence rate of nausea. Thall and Lachin (1988) suggested that the recurrence rate of nausea for the placebo group was significantly greater than that of the high-dose group over the first six months of follow-up, with no substantial difference thereafter.

1.1.2.3 Bladder Cancer Study

The bladder cancer study conducted by the Veterans Administration Cooperative Urological Research Group is a follow-up study of patients with superficial bladder tumors (Byar, 1980; Andrews and Herzberg, 1985; Wei *et al.*, 1989; Wellner and Zhang, 1998). The total 116 patients were randomly divided into three treatment groups: placebo, thiotepa and pyridoxine. At the beginning of the study, all tumors were removed transurethrally, and many patients had multiple recurrences of tumors during the study. At each follow-up visit, the visit time and the number of recurrent tumors between visits were recorded, and then the recurrent tumors were removed. Note that the observation times and follow-up time periods varied among patients. Furthermore, for each patient, two baseline covariates were observed: the number of initial tumors and the size of the largest initial tumor. The purpose of the study was to determine the covariate effects on the tumor recurrence rate. Among others, Sun and Wei (2000) analyzed the study and showed that thiotepa effectively reduced the recurrences of

tumors and that the number of initial tumors was a useful prognostic factor.

1.1.2.4 Psoriatic Arthritis Study

The psoriatic arthritis study was conducted by the University of Toronto Psoriatic Arthritis Clinic which was established in 1978 in order to collect data on the functional and radiological courses of disease for patients with psoriatic arthritis (Gladman *et al.*, 1995). Functional assessments were scheduled annually, during which patients underwent a detailed physical examination including a careful assessment of each of 64 joints. A joint was classified as damaged by a *functional assessment* if there was evidence of deformity or ankylosis, if it flailed, or if it became damaged to the point that surgery was required. Radiological assessments were scheduled to be performed on patients at two year intervals. From the resulting films, a joint was classified as damaged according to *radiological assessment* if there was evidence of surface erosions of the bone in the joint, joint space narrowing, "disorganization" of the joint, or surgery was required. While these two types of assessments were scheduled at regular (but different) times, the actual times and frequency of functional and radiological assessments varied considerably from patient to patient. Covariates of interest included the presence of a family history of psoriasis (yes/no), arthritis duration (years), and the number of active (defined as tender or swollen) joints at clinic entry. The study goal was to investigate the covariate effects on the respective rate functions and estimate the cumulative mean numbers of damaged joints according to each criterion.

The first data set from the CGD study is a typical recurrent event data set, and the other three examples give panel count data. For the reminder of this chapter, first we will begin with a brief introduction on the analysis of recurrent event data in Section 1.2. Section 1.3 discusses statistical inferences about panel count data, and Section 1.4 presents a brief review of counting processes. The outline of the dissertation is given in Section 1.5.

1.2 Analysis of Recurrent Event Data

To analyze recurrent event data, a number of statistical approaches have been developed with respect to modeling the intensity process. For example, following the traditional development of survival analysis, Prentice *et al.* (1981) proposed conditional models that generalized the Cox proportional hazard function (Cox, 1972) for subjects with multiple failures. Andersen and Gill (1982) introduced intensity-based counting process modeling techniques. Specifically, let $\tilde{N}(t)$ denote the number of events that occur over the interval $(0, t]$ and $\mathbf{Z}(t) = (Z_1(t), \ldots, Z_p(t))'$ be a p-dimensional covariate process. Let $\mathcal{F}_{t-}$ represent the σ-field generated by $\{\tilde{N}(s), \mathbf{Z}(s) : 0 \leq s < t\}$. The intensity process $\lambda(t; \mathbf{Z})$ can be defined by

$$\lambda(t; \mathbf{Z})dt = E\{d\tilde{N}(t)|\mathcal{F}_{t-}\} .$$

The Andersen-Gill intensity model has the form as

$$\lambda(t;\mathbf{Z}) = \lambda_0(t)exp\{\beta'\mathbf{Z}(t)\}\ , \tag{1.1}$$

where $\lambda_0(t)$ is an unknown continuous baseline intensity function and β denotes a $p \times 1$ vector of regression parameters. Using martingale methods, Andersen and Gill (1982) gave a thorough account for the asymptotic properties of the estimators from model (1.1) based on the partial likelihood (Cox, 1975). Lawless (1987) presented a class of parametric and semiparametric procedures based on a nonhomogeneous Poisson process and proportional intensity assumptions with random effects. Wei *et al.* (1989) considered marginal models to analyze the multivariate failure time data. Huang and Wang (2004) studied the joint modeling of a recurrent event process and a failure time where latent variables were used to model the association between the intensity of recurrent event process and the hazard function of the failure time.

Although the intensity process may be of interest in some settings, a number of authors have focused on modeling the mean number of events since the mean function is sometimes more interpretable than the intensity process. Denoting $E\{d\tilde{N}(t)|\mathbf{Z}(t)\}$ by $d\mu(t;\mathbf{Z})$, we can specify the following Cox model

$$d\mu(t;\mathbf{Z}) = d\mu_0(t)exp\{\beta'\mathbf{Z}(t)\}\ , \tag{1.2}$$

where $\mu_0(t)$ is an unspecified continuous baseline mean function. If $\mathbf{Z}$ is time-independent,

then we can express the resulting model as

$$\mu(t;\mathbf{Z}) = \mu_0(t)exp\{\beta'\mathbf{Z}\} \ . \tag{1.3}$$

Models (1.2) and (1.3) characterize the rate and mean functions of the underlying counting process $\tilde{N}(t)$, respectively. Among others, Pepe and Cai (1993) gave an estimating procedure for the rate function of recurrence after the first event. Lawless and Nadeau (1995) discussed a class of marginal models when the failure times are discrete. Lin *et al.* (2000) investigated the robust inferences of semiparametric regression models with absolutely continuous failure times through modern empirical process theory. Lin *et al.* (1998) extended the conventional accelerated failure time model for survival data to the mean function of counting process for recurrent events and developed the corresponding asymptotic distribution theory. Cook and Lawless (1997) studied the marginal mean and rate models by including terminal events, which are correlated with the recurrent event process. Wang *et al.* (2001) adopted a marginal mean model with a multiplicative random effect for recurrent events with informative censoring. Cai and Schaubel (2004b) proposed a class of semiparametric marginal mean and rate models for correlated multiple-type recurrent event data.

For recurrent event studies, in addition to the occurrence rate of the event of interest, sometimes one may be interested in the recurrent gap time. For example, Lin *et al.* (1999) and Wang and Chang (1999) presented nonparametric estimators of the distribution function of the gap times between successive events. Huang and Chen

(2003) considered the proportional hazards model for regression analysis of recurrent gap times. Chen *et al.* (2004) proposed proportional reverse-time hazards models to estimate the longitudinal pattern parameter of the recurrent gap times. Sun *et al.* (2006) investigated the analysis of recurrent gap times with the additive hazards model.

1.3 Analysis of Panel Count Data

By panel count data, we mean that, for each subject, observations are taken at finite discrete time points, and only the number of recurrent events that have occurred before each observation time is known. In particular, no information is available on subjects about the specific timing of events between observation time points. Furthermore, the set of observation times may vary from subject to subject.

For the analysis of panel count data, several parametric approaches have been proposed. Kalbfleisch and Lawless (1985) discussed the fitting of a finite state Markov model to panel count data. Hinde (1982) and Breslow (1984) considered regression analysis of Poisson count data. Thall (1988) investigated some regression models for mixed Poisson processes. Liang and Zeger (1986) and Thall and Vail (1990) presented quasi-likelihood regression models with a generalized estimating equation (GEE) approach by treating panel count data as longitudinal count data. Also, Diggle *et al.* (1994) gave a review of parametric approaches at this point.

Besides the above parametric approaches, many nonparametric and semiparametric methods have been developed for the analysis of panel count data. Due to the incomplete nature of panel count data, it is more convenient to regard the observations

as underlying counting processes. In this section, we focus on inferences on the mean function and regression models for panel count data.

1.3.1 Inferences on the Mean Function

Consider a follow-up study involving n subjects who may experience recurrent events, and let $N_i(t)$ represent the cumulative number of events that have occurred prior to time t for the ith subject with $N_i(0) = 0$. Define $\mu(t) = E\{N_i(t)\}$ to be the mean function of the process N_i. Suppose that $N_i(.)$ is observed only at finite time points $t_{i,1} < \ldots < t_{i,m_i}$, where m_i denotes the potential number of observation times for subject i, $i = 1, \ldots, n$.

In a simple situation, $t_{i,j} = s_j$ for $j = 1, \ldots, m_i$ and $i = 1, \ldots, n$ with $m_i \leq m$, where m is the total number of distinct observation times. In this case, all the subjects have the same observation times except that the numbers of observations may be different. The Nelson-Aalen estimator of $\mu(s_j)$ can be written as

$$\hat{\mu}(s_j) = \sum_{l=1}^{j} \frac{\sum_{i=1}^{n} I(s_l \leq t_{i,m_i})\{N_i(s_l) - N_i(s_{l-1})\}}{\sum_{i=1}^{n} I(s_l \leq t_{i,m_i})} .$$

For the general situation, the observation times may vary from subject to subject, there is no extension of the Nelson-Aalen estimator. Thall and Lachin (1988) described an approach for estimation of the rate function $d\mu(t)$, which can be used to estimate the mean function $\mu(t)$ by taking the integral of the rate function. In their method, the rate function, which is assumed to be constant between observation times for each

subject, can be estimated by

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

Thus the estimator of the mean function is given by

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

Utilizing the monotonic property of the mean function of a counting process, Sun and Kalbfleisch (1995) applied isotonic regression techniques to estimate the mean function. Let $s_1, \ldots, s_m$ denote the ordered distinct observation times in the set $\{t_{i,l} : l = 1, \ldots, m_i, i = 1, \ldots, n\}$. Let l_j and $\bar{n}_j$ represent the number and mean value respectively of observations made at s_j, $j = 1, \ldots, m$. The isotonic regression estimators $\hat{\mu}(s_1), \ldots, \hat{\mu}(s_m)$ are defined as the $\mu(s_1), \ldots, \mu(s_m)$ that minimize the weighted sum of squares

$$\sum_{j=1}^{m} l_j \{\bar{n}_j - \mu(s_j)\}^2$$

subject to the order restriction $\mu(s_1) \leq \ldots \leq \mu(s_m)$. Using the max-min formula for isotonic regression (Barlow *et al.*, 1972; Robertson *et al.*, 1988), the isotonic estimator for $\mu(s_j)$ is given by

$$\hat{\mu}(s_j) = max_{r \leq j} min_{u \geq j} \frac{\sum_{v=r}^{u} l_v \bar{n}_v}{\sum_{v=r}^{u} l_v} = min_{u \geq j} max_{r \leq j} \frac{\sum_{v=r}^{u} l_v \bar{n}_v}{\sum_{v=r}^{u} l_v} \ . \tag{1.4}$$

Wellner and Zhang (2000) showed that the isotonic regression estimator (1.4) is equivalent to a pseudo-maximum likelihood estimator based on the following pseudo log likelihood function

$$l_p(\mu) = \sum_{i=1}^{n}\sum_{j=1}^{m_i}\{N_i(t_{i,j})log\mu(t_{i,j}) - \mu(t_{i,j})\} = \sum_{j=1}^{m} l_j\{\bar{n}_j log\mu_j - \mu_j\}$$

assuming that the counting process is a non-homogeneous Poisson process. Under the same assumption, they proposed a nonparametric maximum likelihood estimator for the mean function based on the log full likelihood function, which is proportional to

$$l(\mu) = \sum_{i=1}^{n}\sum_{j=1}^{m_i}\{N_i(t_{i,j}) - N_i(t_{i,j-1})\}log\{\mu(t_{i,j}) - \mu(t_{i,j-1})\} - \sum_{i=1}^{n}\mu(t_{i,m_i}) \ ,$$

where $t_{i,0} = 0$ and $N_i(0) = 0$. Wellner and Zhang (2000) also investigated the asymptotic properties of both estimators.

Since the above methods ignored the dependence among the counts within a subject, Zhang and Jamshidian (2003) introduced the gamma frailty variable γ_i to account for correlation among the panel counts. Their model can be written as

$$\mu_i(t|\gamma_i) = E\{N_i(t)|\gamma_i\} = \gamma_i\mu(t) \ ,$$

assuming that $\gamma_i \sim Gamma(\alpha, 1/\alpha)$ such that $E(\gamma_i) = 1$ and thus $\mu(t) = E\{N_i(t)\}$, $i = 1, \ldots, n$. Furthermore, Zhang and Jamshidian (2003) developed an EM algorithm for a maximum pseudo-likelihood estimate of the mean function.

Besides estimation of mean functions, some methods have also been proposed for the comparison of mean functions between two treatment groups. Thall and Lachin (1988) considered to represent each subject's empirical rate function as a vector corresponding to K fixed time intervals and constructed a two-sample comparison by the Wei-Lachin vector of Wilcoxon-like rank statistics (Wei and Lachin, 1984). However, one limitation of this method is that the number of selected intervals and the intervals themselves would affect the test result.

Sun and Kalbfleisch (1993) developed a statistic for testing the equality of the mean functions of point processes for current status data, which are special cases of panel count data when each subject is observed only once. Sun and Fang (2003) extended that test statistic to general panel count data. For the two-sample comparison problem, let z_i be the group indicator (0 or 1). Then to test the hypothesis $H_0 : \mu_1(t) = \mu_2(t)$, the statistic is given by

$$U_n = \sum_{i=1}^{n} z_i \sum_{j=1}^{m_i} \{N_i(t_{i,j}) - \hat{\mu}(t_{i,j})\} \ ,$$

where $\hat{\mu}(t_{i,j})$ is the isotonic regression estimator in (1.4). The asymptotic distribution of U_n was derived. Furthermore, Sun and Rai (2001) investigated a class of nonparametric test statistics for the comparison of several point processes with respect to the intensity functions for the situation when the observation processes are non-informative. Although this method is not our focus in this section, it can serve as an alternative approach for the comparison procedure of panel count data.

1.3.2 Regression Models for Panel Count Data

For regression analysis of panel count data, several methods have been proposed based on the rate and mean functions, which are similar to those for recurrent event data mentioned in Section 1.2. Let the $N_i(t)$ and $t_{i,j}$ for $j = 1, \ldots, m_i$ be defined as in the previous section. Suppose that for subject i, $\mathbf{Z}_i$ represents a p-dimensional vector of covariates that are assumed to be time-independent; C_i denotes the follow-up or censoring time. The proportional means model for panel count data can be expressed as

$$\mu_i(t) = \mu_0(t)exp(\beta'\mathbf{Z}_i) \tag{1.5}$$

for $i = 1, \ldots, n$, where $\mu_0(t)$ is an unspecified continuous baseline mean function and β represents a $p \times 1$ vector of regression parameters. Let $Y_i(t) = I(t \leq C_i)$, indicating if subject i is at risk of experiencing the recurrent events at time t. Then the proportional rates model for panel count data is given by

$$\lambda_i(t) = \lambda_0(t)Y_i(t)exp(\beta'\mathbf{Z}_i) \tag{1.6}$$

for $i = 1, \ldots, n$, where $\lambda_0(t)$ denotes an unknown continuous baseline rate function. The two models above are the most commonly used models for regression analysis of panel count data.

Several authors developed inference procedures for the proportional rates model (1.6). Sun and Matthews (1997) considered it for the case when the observation times

on each subject can be regarded as random variables with respect to prespecified observation times. They assumed that there exist k fixed time points denoted by t_l's and that each $t_{i,j}$ can be modeled by

$$t_{i,j} = t_l + \epsilon_{i,j}$$

for some l, $1 \leq l \leq k$, $j = 1, \ldots, m_i$, $i = 1, \ldots, n$, where the $\epsilon_{i,j}$'s are random variables with mean 0 and finite variance. They also proposed an estimating equation for estimation of regression parameters and derived the corresponding asymptotic properties of the estimator. Staniswalis *et al.* (1997) used smoothing splines to obtain a nonparametric baseline rate function estimate and profile likelihood to estimate the covariate parameters. Lawless and Zhan (1998) discussed a maximum likelihood-based approach and extended generalized estimating equations by using a piecewise constant rate function.

Besides the proportional rates model (1.6), many semiparametric methods have been proposed for the proportional means model (1.5). Among others, Sun and Wei (2000) considered estimating equation methods when both observation times and follow-up times may depend on the covariates. Let $H_i(t) = \sum_{j=1}^{m_i} I(t_{i,j} \leq t)$ be the underlying observation process that represents the potential number of observations up to time t on subject i and define $\tilde{N}_i(t) = H_i\{min(t, C_i)\}$ for $i = 1, \ldots, n$. Thus $\tilde{N}_i(t)$ is the real observation process and jumps only at the observation times for subject i. Analogous to the proportional means model (1.5) for $\{N_i(t)\}$, assume that the mean

function of the counting process $\{H_i(t)\}$ has the form

$$\tilde{\mu}_i(t) = \tilde{\mu}_0(t)exp(\gamma'\mathbf{Z}_i) \ ,$$

where $\tilde{\mu}_0(t)$ is a completely unspecified baseline mean function and γ is a $p \times 1$ vector of regression parameters. For the follow-up time C_i's, suppose that the hazard function $\lambda_i^*(t)$ for subject i is given by

$$\lambda_i^*(t) = \lambda_0^*(t)exp(\tau'\mathbf{Z}_i) \ ,$$

where $\lambda_0^*(t)$ is an unspecified baseline hazard function and τ is a $p \times 1$ vector of parameters (Cox, 1972). Furthermore, assume that N_i, H_i, C_i and $\mathbf{Z}_i$ may be dependent, but conditionally on $\mathbf{Z}_i$, N_i, H_i and C_i are independent. Sun and Wei (2000) developed estimating equations for τ, γ and β and also derived the corresponding asymptotic distributions of the estimators. Cheng and Wei (2000) and Hu *et al.* (2003) also investigated estimating equation approaches for the case when H_i and C_i are independent of N_i and the case when C_i is independent of $(H_i, N_i, \mathbf{Z}_i)$ but H_i may depend on N_i via $\mathbf{Z}_i$, respectively.

Alternative approaches to deal with the proportional means model have also been developed based on the likelihood functions. Given the covariate vector $\mathbf{Z}_i$, $i = 1, \ldots, n$, assume that the $N_i(t)$'s are non-homogeneous Poisson processes with the mean function

(1.5). Then the log likelihood function can be expressed as

$$l(\mu_0,\beta) = \sum_{i=1}^{n}\sum_{j=1}^{m_i}\Big[\{N_i(t_{i,j}) - N_i(t_{i,j-1})\}log\{\mu_0(t_{i,j}) - \mu_0(t_{i,j-1})\} + \{N_i(t_{i,j}) - N_i(t_{i,j-1})\}\beta'\mathbf{Z}_i - \{\mu_0(t_{i,j}) - \mu_0(t_{i,j-1})\}exp(\beta'\mathbf{Z}_i)\Big] \ . \quad (1.7)$$

Assuming that the dependence of the counts within a subject is negligible, the pseudo log likelihood function is given by

$$l_p(\mu_0,\beta) = \sum_{i=1}^{n}\sum_{j=1}^{m_i}\{N_i(t_{i,j})log\mu_0(t_{i,j}) + N_i(t_{i,j})\beta'\mathbf{Z}_i - \mu_0(t_{i,j})exp(\beta'\mathbf{Z}_i)\} \ . \quad (1.8)$$

Zhang (2002) studied semiparametric pseudo-likelihood methods for estimating (β, μ_0) jointly based on (1.8) and the working assumptions mentioned above. Wellner *et al.* (2004) investigated maximum likelihood estimation according to (1.7), discussed the asymptotic normality of the two estimators, and compared them under different scenarios.

A basic assumption behind all the methods discussed above is that the observation times, or the observation and censoring times, are independent of the underlying recurrent event process completely or given covariates. In practice, however, this may not be true. As mentioned in Section 1.2, some authors considered the analysis of recurrent event data with informative censoring (Wang *et al.*, 2001; Huang and Wang, 2004; Liu *et al.*, 2004; Ye *et al.*, 2007). Sinha and Maiti (2004) developed a Bayesian analysis of panel count data when the censoring time may be correlated with the underlying

counting process of interest by assuming that all the subjects have fixed observation times. However, very little further work about panel count data has been done for the case where the underlying counting process of interest and the observation process may depend on each other or the case where the censoring process may be also related with these two processes. This motivates our research in Chapters 2 and 4, respectively.

Another important issue about panel count data is multivariate analysis. Since multi-type recurrent event processes may be correlated, the corresponding dependency needs to be accommodated. Chen *et al.* (2005) proposed methods based on a mixed Poisson model with piecewise constant baseline intensities and multivariate log-normal random effects. We will present a marginal approach that avoids the assumptions of Poisson process and piecewise constant baseline intensities in Chapter 3.

1.4 Counting Processes

1.4.1 Introduction

In this section, we give a brief review of some basic concepts about counting processes, which play an essential role in the development of statistical models for event history data. For event history analysis based on counting processes, the fundamental work was done by Aalen in his 1975 Berkeley Ph.D. dissertation. He made a decisive breakthrough for the use of modern stochastic processes theory and showed how the theory of multivariate counting processes provides a general framework in event history analysis (Andersen and Borgan, 1985). Andersen and Borgan (1985)

made an extensive review of the work by Aalen and others. A detailed description and complete integration of the historical developments of the counting process approach can be found in the book by Andersen *et al.* (1993).

1.4.2 Counting Processes

Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\mathcal{T} = [0, \tau)$ be an index set of time t, where τ is a given terminal time, $0 < \tau \leq \infty$. A *stochastic process* is a family of random variables $X = \{X(t) : t \in \mathcal{T}\}$. A *filtration* or *history*, $(\mathcal{F}_t : t \in \mathcal{T})$, is defined as an increasing right-continuous family of sub-σ-algebras of $\mathcal{F}$ such that $\mathcal{F}_t = \sigma\{X(s) : 0 \leq s \leq t\}$ contains all the event information generated by the stochastic process X on $[0, t]$. A process X is *predictable* with respect to $\mathcal{F}_t$ if $X(t)$ is known given the history $\mathcal{F}_{t-}$. The process X is called *adapted* to the filtration if $X(t)$ is $\mathcal{F}_t$-measurable for every $t \in \mathcal{T}$. Therefore, any process is adapted to its history.

A *counting process* is a stochastic process $\{N(t), t \geq 0\}$ adapted to a filtration $X = \{X(t) : t \geq 0\}$ with $N(0) = 0$ and $N(t) < \infty$ *a.s.*, and whose paths are with probability one right-continuous, piecewise constant, and have only jump discontinuities, with jumps of size $+1$.

In particular, a *Poisson process* is a counting process $\{N(t), t \geq 0\}$ such that

$$P\{N(t) - N(t-h) = 1 | N(s), 0 \leq s \leq t-h\} = \lambda(t)h + o(h)$$

and

$$P\{N(t) - N(t-h) \geq 2 | N(s), 0 \leq s \leq t-h\} = o(h)$$

for small $h > 0$ and all $t > 0$, where $\lambda(t) \geq 0$ is a left continuous function satisfying $\int_0^t \lambda(s)ds = \Lambda(t) < \infty$. Here $\lambda(t)$ and $\Lambda(t)$ are the *intensity* and *cumulative intensity* functions of the Poisson process. The above conditions can be also equivalently written as

$$P\{dN(t) = 1 | \mathcal{F}_{t-}\} = P\{dN(t) = 1\} = \lambda(t)h$$

and

$$P\{dN(t) = 0 | \mathcal{F}_{t-}\} = 1 - \lambda(t)h \ .$$

The Poisson process defined above is also known as a *nonhomogeneous Poisson process*. If $\lambda(t)$ is time invariant, it is called a *homogeneous Poisson process*. For a Poisson process $\{N(t), t \geq 0\}$,

$$N(t) \sim Poisson(\Lambda(t)) \ .$$

1.4.3 Multivariate Counting Processes

The definition of the univariate counting processes can be generalized to the multivariate counting processes. A *multivariate counting process* $\mathbf{N} = [\{N_1(t), N_2(t), \ldots, N_k(t)\}, t \geq 0]$ is defined as a stochastic process with k components such that

(1) each component $N_h(t)$ is a counting process,

(2) no two component processes can jump simultaneously,

(3) each $N_h(\infty)$ is almost surely finite,

for $h = 1, \ldots, k$. In this multivariate counting process $\mathbf{N}$, each $N_h(t)$ represents the number of type h events occurring over the time interval $[0, t]$. Furthermore, the development in time of $\mathbf{N}$ is governed by its *intensity process* $\lambda = [\{\lambda_1(t), \lambda_2(t), \ldots, \lambda_k(t)\}, t \geq 0]$, where $\lambda_h(t)dt = P\{dN_h(t) = 1|\mathcal{F}_{t-}\}$ for $h = 1, \ldots, k$.

For multivariate counting processes, Aalen (1978) introduced the multiplicative intensity model. The intensity process $\lambda = [\{\lambda_1(t), \lambda_2(t), \ldots, \lambda_k(t)\}, t \geq 0]$ is defined to satisfy

$$\lambda_h(t) = \alpha_h(t)Y_h(t) \ ,$$

where $\alpha_h(t)$ is a nonnegative deterministic function, while $Y_h(t)$ is a predictable process whose value at any time t is known just before t, for $h = 1, \ldots, k$.

For example, suppose that there are n patients. We observe (T_{hi}, D_{hi}) for subject i with event type h, where T_{hi} is the survival time or the censoring time, and D_{hi} is the indicator of whether T_{hi} is a true survival time, $h = 1, \ldots, k$, $i = 1, \ldots, n$. If we assume that for event type h, all the n subjects are from a homogeneous population with the same death intensity $\alpha_h(t)$ for $h = 1, \ldots, k$, then a univariate counting process can be defined as $N_h(t) = \sum_{i=1}^{n} I(T_{hi} \leq t, D_{hi} = 1)$, and the intensity process is given by $\lambda_h(t) = \alpha_h(t)Y_h(t)$, where $Y_h(t) = \sum_{i=1}^{n} I(t \leq T_{hi})$ represents the number of subjects at risk just prior to time t for event type h.

1.4.4 Martingales and Stochastic Integrals

Consider a multivariate counting process $\mathbf{N} = [\{N_1(t), N_2(t), \ldots, N_k(t)\}, t \geq 0]$ with intensity process $\lambda = [\{\lambda_1(t), \lambda_2(t), \ldots, \lambda_k(t)\}, t \geq 0]$, where $\lambda_h(t)dt = P\{dN_h(t) =$

$1|\mathcal{F}_{t-}\}$, $h = 1, \ldots, k$. According to the definition of counting process, we have

$$E\{dN_h(t)|\mathcal{F}_{t-}\} = \lambda_h(t)dt$$

for $h = 1, \ldots, k$. For $h = 1, \ldots, k$, define

$$dM_h(t) = dN_h(t) - \lambda_h(t)dt \ ,$$

which satisfies $M_h(0) = 0$ and

$$E\{dM_h(t)|\mathcal{F}_{t-}\} = 0 \ .$$

That is, the processes

$$M_h(t) = N_h(t) - \int_0^t \lambda_h(s)ds \ , \tag{1.9}$$

$h = 1, \ldots, k$, $t \geq 0$, are *martingales*. In particular, $E\{M_h(t)\} = 0$ for $t \geq 0$.

The definition of martingales (1.9) is the key to connect the counting process approach to event history analysis. In the framework of counting processes, many estimators and test statistics can be expressed or approximated by stochastic integrals with respect to martingales. Therefore, many well studied properties of martingales can be applied to investigate various statistical inference problems (Andersen and Borgan, 1985).

One of the most important properties of martingales is the martingale central limit

theorem, which is widely used to study the large sample properties of statistical methods for counting process models. The martingale central limit theorem gives conditions under which a sequence of martingales $\{M^{(n)}(t), t \geq 0\}$, $n = 1, 2, \ldots$, converges to a continuous Gaussian martingale $M^{(\infty)}(t)$ as $n \to \infty$. The *continuous Gaussian martingale* $\{M^{(\infty)}(t), t \geq 0\}$ is a time-transformed Wiener process with independent normally distributed mean zero increments. The conditional variance of a martingale M is given by the *predictable variation process* $\langle M \rangle$, where

$$d\langle M \rangle(t) = Var\{dM(t)|\mathcal{F}_{t-}\} \ .$$

For the martingales M_h, $h = 1, \ldots, k$, defined by (1.9), it can be shown that

$$\begin{aligned} d\langle M_h \rangle(t) = & \quad Var\{dN_h(t) - \lambda_h(t)dt|\mathcal{F}_{t-}\} \\ = & \quad Var\{dN_h(t)|\mathcal{F}_{t-}\} \\ \approx & \quad \lambda_h(t)dt\{1 - \lambda_h(t)dt\} \\ \approx & \quad \lambda_h(t)dt \ , \end{aligned}$$

since $\lambda_h(t)$ is predictable and $dN_h(t)$ is a $0-1$ variable. Thus the conditional variance of $M_h(t)$ is given by

$$\langle M_h \rangle(t) = \int_0^t d\langle M_h \rangle(s)ds \approx \int_0^t \lambda_h(s)ds \ .$$

Let $H(t)$ be a predictable stochastic process, and define the processes $M_h^*(t)$ by the

stochastic integral

$$M_h^*(t) = \int_0^t H(s) dM_h(s)$$

for $h = 1, \ldots, k$. Then $M_h^*(t)$, $h = 1, \ldots, k$, are martingales too since

$$E\{dM_h^*(t)|\mathcal{F}_{t-}\} = E\{H(t)dM_h(t)|\mathcal{F}_{t-}\} = H(t)E\{dM_h(t)|\mathcal{F}_{t-}\} = 0 \ .$$

Also, because

$$Var\{H(t)dM_h(t)|\mathcal{F}_{t-}\} = H^2(t)d\langle M_h\rangle(t) \ ,$$

we have

$$\langle M_h^*\rangle(t) = \int_0^t H^2(s)d\langle M_h\rangle(s) \approx \int_0^t H^2(s)\lambda_h(s)ds \ .$$

Two martingales M_1 and M_2 are *orthogonal* if

$$\langle M_1, M_2\rangle(t) = \int_0^t d\langle M_1, M_2\rangle(s) = \int_0^t Cov\{dM_1(s), dM_2(s)|\mathcal{F}_{t-}\} = 0 \ .$$

For any two martingales $M_i(t)$ and $M_j(t)$, $i \neq j$, defined by (1.9), we have

$$d\langle M_i, M_j\rangle(t) = Cov\{dN_i(t) - \lambda_i(t)dt, dN_j(t) - \lambda_j(t)dt|\mathcal{F}_{t-}\} \approx E\{dN_i(t)dN_j(t)|\mathcal{F}_{t-}\} = 0$$

by the facts that $\lambda_i(t)$ and $\lambda_j(t)$ are predictable and that $N_i(t)$ and $N_j(t)$ do not jump at the same time. Therefore, the martingales $M_1(t), \ldots, M_k(t)$ are orthogonal (Andersen and Borgan, 1985).

1.5 Outline of the Dissertation

The rest of this dissertation contains four parts about *Semiparametric Analysis of Panel Count Data* from Chapter 2 to Chapter 5.

In Chapter 2, we consider regression analysis of panel count data with dependent observation times. For longitudinal data analysis, most of the existing methods focus on situations where observation times are independent of longitudinal response variables and therefore rely on conditional inference procedures given the observation times. In practice, however, the independence assumption may not hold. We investigate a class of semiparametric models to characterize the possible correlation between the observation process and the response process through a subject-specific latent variable or frailty. For inference, estimating equation approaches are proposed for estimation of regression parameters, and both large and finite sample properties of the methods are established. Simulation studies show that the proposed estimation procedures work well, and the methodology is applied to the bladder cancer study.

Chapter 3 discusses regression analysis of multivariate panel count data. Fields that produce such data include epidemiological studies, medical follow-up studies, reliability studies, and tumorigenicity experiments. We present a class of marginal mean models that leave the dependence structures for the related types of recurrent events completely unspecified. To estimate regression parameters, some estimating equations are developed, and the resulting estimates are shown to be consistent and asymptotically normal. Simulation studies are conducted for practical situations, and the results

are supportive. The methodology is applied to the psoriatic arthritis study discussed before.

Chapter 4 considers an approach that allows the censoring or follow-up time to be related with the response process of interest as well as the observation process. Although most available methods assume that the censoring is noninformative, this assumption may be violated in some clinical trials when there exists informative dropout such that sicker patients may tend to drop out of the study early. Therefore, we propose some shared frailty models to characterize the potential dependence among those three processes. To fit the model, estimating equations and an EM algorithm are used. The proposed estimates are consistent and have asymptotically a normal distribution. The finite sample properties of the proposed estimates are investigated through simulation, and the method is reapplied to the bladder cancer study.

This dissertation concludes with Chapter 5, which addresses several directions for future research.

Chapter 2

REGRESSION ANALYSIS OF PANEL COUNT DATA WITH DEPENDENT OBSERVATION TIMES

2.1 Introduction

This chapter discusses regression analysis of panel count data, which often occur in long term studies that concern occurrence rates of recurrent events. By panel count data, we mean that it is not feasible or realistic to keep study subjects under observation continuously and for each subject, only the numbers of occurrences of the event are known at finite distinct observation time points over the study period. Furthermore, the set of observation times may vary from subject to subject. Areas that often produce panel count data include demographical studies, epidemiological studies, medical periodic follow-up studies, and tumorgenecity experiments.

For recurrent event studies, if the occurrence times of all events are known, the ob-

served data are usually referred to as recurrent event data (Andersen *et al.*, 1993; Cook *et al.*, 1996; Lin *et al.*, 2000; Wang *et al.*, 2001). This can occur if all study subjects are under continuous observation. For the analysis of recurrent event data, it is common and convenient to treat the data as realizations of some counting processes, and many statistical methods based on this counting process formulation have been proposed. Among others, Andersen and Gill (1982) presented a Cox type intensity model for regression analysis of recurrent event data, and the model has been extensively studied in the literature. A detailed description about the model and related references can be found in the excellent book by Andersen *et al.* (1993). More recently, instead of modeling the intensity process, some authors proposed to model the mean and rate functions of the underlying counting processes (Lawless and Nadeau, 1995; Cook *et al.*, 1996). In particular, Lin *et al.* (2000) provided a rigorous formalization of the marginal means and rates models and developed corresponding inference procedures.

As mentioned in Section 1.3, several methods have been proposed for the analysis of panel count data. For example, Kalbfleisch and Lawless (1985) discussed the fitting of Markov model to panel count data, and Sun and Kalbfleisch (1995) and Wellner and Zhang (2000) considered estimation of mean function of the underlying point process that yields observed panel count data. For treatment comparison based on panel count data, Thall and Lachin (1988) presented a procedure that transforms the problem to a multivariate comparison problem, while Sun and Fang (2003) proposed a nonparametric approach. Sun and Wei (2000) and Zhang (2002) investigated regression analysis of panel count data. The former developed some estimating equation-based methods and

the latter proposed a pseudo-likelihood approach.

A fundamental assumption behind these methods is that the observation times or the counting process that characterizes the observation times is independent of occurrences of the recurrent event under study completely or given covariates. In other words, the underlying point process of interest that governs the occurrence of the event and the observation counting process are independent. In practice, however, this may not be true. For example, consider the bladder cancer study discussed in Section 1.1.2.3 and Sun and Wei (2000). In the study, the patients visited the clinical centers periodically and at each visit, the number of bladder tumors that occurred since the last visit was recorded. As commented in Sun and Wei (2000), some patients in the study had more visits than others, and the occurrence of bladder tumors and the visit may be related. To see this, we calculated the sample correlation between the first visit times and the numbers of bladder tumors observed at these times and obtained -0.1180 for the patients in the placebo group. Although the evidence is weak, it indicates that the tumor occurrence process and visit process may be negatively correlated and one should take it into account for the analysis if possible. Note that the sample correlation given above is based only on the partial information and the analysis based on all information in Section 2.5 indicates that the two processes were indeed negatively correlated. Wang *et al.* (2001) described another study of AIDS patients in which the patient observation times are their hospitalization times and it is apparent that for the analysis of any response variable related to AIDS, one may want to consider the correlation between the observation process and the response

process. The same could occur in other disease studies in which one is interested in the occurrence rate of some symptoms related to the disease under study and the observation times are hospitalization times of the patients.

In the following, we consider situations where the point process of interest and the observation process may be correlated. To begin with, we first introduce notation and models for the two processes in Section 2.2. The models allow for the possible correlation between the two processes and characterize their relationship through a subject-specific latent variable or frailty. Section 2.3 presents some estimating equation-based approaches for estimation of regression parameters, and asymptotic properties of the proposed parameter estimates are established. Section 2.4 gives some results from a simulation study conducted to assess the finite sample performance of the proposed estimation procedures, and the method is applied to the bladder cancer study in Section 2.5. Section 2.6 concludes with some discussion.

2.2 Models and Notation

Consider a study involving n subjects who may experience recurrent events. For subject i, define $N_i(t)$ to be the cumulative number of events that have occurred prior to time t, $0 \leq t \leq \tau$, where τ is a known constant time point, $i = 1, ..., n$. Let $x_i = (x_{i1}, ..., x_{ip})'$ be a $p \times 1$ covariate vector associated with subject i. For the covariate effect on $N_i(t)$, we assume that given x_i and a subject-specific unobservable positive

frailty z_i, the mean function of $N_i(t)$ has the form

$$\mu_i(t) \;=\; z_i^{\alpha}\, \mu_0(t) \exp(\beta' x_i)\;. \tag{2.1}$$

In model (2.1), $\mu_0(\cdot)$ is a completely unspecified baseline mean function, α is an unknown parameter, and β is a vector of unknown regression parameters. The goal is to make inference about β.

In this chapter, we consider the situation where only panel count data are available for the $N_i(t)$'s. Specifically, suppose that $N_i(\cdot)$ is observed only at finite time points $T_{i1} < ... < T_{iK_i}$, where K_i denotes the potential number of observation times for subject i, $i = 1, ..., n$. In other words, only values of the $N_i(t)$'s at these observation times are known. In general, not every subject can be followed until τ, and there exists a follow-up time C_i for subject i. That is, $N_i(T_{il})$ is observed only if $T_{il} \le C_i \le \tau$. Define $\tilde{N}_i(t) = H_i\{\min(t, C_i)\}$, where $H_i(t) = \sum_{l=1}^{K_i} I(T_{il} \le t)$ and $I(\cdot)$ is the indicator function, $i = 1, ..., n$. That is, $\tilde{N}_i(t)$ is a point process characterizing the ith subject's observation process and jumps only at the observation times.

In the following, we assume that given (x_i, z_i), $H_i(\cdot)$ is a non-homogeneous Poisson process with the intensity function given by

$$\lambda_i(t) \;=\; z_i\, \lambda_0(t)\, \exp(\gamma' x_i)\;. \tag{2.2}$$

In the above, $\lambda_0(t)$ is a completely unknown continuous baseline intensity function, and γ denotes the vector of regression parameters. Define $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$, the

baseline cumulative intensity function. It will be assumed that $\Lambda_0(\tau) = 1$ to avoid the identifiability issue and that the z_i's are realizations of a latent variable Z satisfying $E(Z|X) = E(Z)$ (Wang *et al.*, 2001). Note that instead of $\Lambda_0(\tau) = 1$, an alternative is to assume that $E(Z) = 1$. Furthermore, we assume that conditional on (x_i, z_i), $(N_i(\cdot), H_i, C_i)$ are mutually independent and $[\{H_i(t), N_i(t), C_i, x_i', z_i\}', 0 \leq t \leq \tau, i = 1, ..., n]$ are independent and identically distributed.

Both models (2.1) and (2.2) or their variants have been discussed separately by many authors for various situations. For example, Sun and Wei (2000) and Zhang (2002) considered model (2.1) with $\alpha = 0$ for regression analysis of panel count data. Model (2.2) with $z_i = 1$ is commonly used for regression analysis of recurrent event data. Among others, Wang *et al.* (2001) and Huang and Wang (2004) applied model (2.2) to the analysis of recurrent event data in the presence of informative censoring time.

Under the above models and assumptions, it is obvious that N_i and H_i can be correlated and their relationship is partly determined by parameter α. Specifically, $\alpha = 0$ means that the two processes are independent given x_i. For $\alpha > 0$ and subjects with the same x_i, the subject with more frequent observations would have a higher occurrence rate of the recurrent event. That is, N_i and H_i are positively correlated. On the other hand, if $\alpha < 0$, a subject with more frequent observations would have a lower occurrence rate of the event, and N_i and H_i are negatively correlated. In the following, we discuss inference about regression parameters.

2.3 Estimation Procedures

For estimation of β along with α and γ, first we consider the situation where the C_i's are independent of the N_i's and H_i's as well as the x_i's and z_i's. For this, define

$$\bar{N}_i \;=\; \sum_{l=1}^{K_i} N_i(T_{il}) I(T_{il} \le C_i) \;=\; \int_0^\tau N_i(t)\, d\tilde{N}_i(t)\,.$$

Then conditional on (x_i, z_i), we have

$$E(\bar{N}_i) \;=\; z_i^{1+\alpha}\, \exp\{(\beta+\gamma)' x_i\} \int_0^\tau \lambda_0(t)\, P(C_i \ge t)\, \mu_0(t)\, dt\,. \tag{2.3}$$

For estimation of β and α, following Sun and Wei (2000) and motivated by equation (2.3), we can consider the following estimating function

$$U_1(\beta_1 \,|\, \gamma, x_i's, z_i's) \;=\; \frac{1}{n} \sum_{i=1}^{n} w_{1i}\, x_{1i}\, \left[\bar{N}_i \;-\; \theta\, z_i^{1+\alpha}\, \exp\{(\beta+\gamma)' x_i\}\,\right]$$

$$=\; \frac{1}{n} \sum_{i=1}^{n} w_{1i}\, x_{1i}\, \left\{\bar{N}_i \;-\; \exp(\gamma' x_i)\, \exp(\beta_1' x_{1i})\,\right\}$$

for given γ and the z_i's. In the above, $\theta \;=\; \int_0^\tau \lambda_0(t)\, P(C_i \ge t)\, \mu_0(t)\, dt$, $\beta_1' \;=\; (\beta', 1+\alpha, \log(\theta))$, $x_{1i} \;=\; (x_i', \log z_i, 1)'$, and the w_{1i}'s are weights that could depend on the x_i's and C_i's. We remark that in the construction of the above estimating function, we used the $\bar{N}_i$'s instead of the N_i's. To see this, define $\bar{N}_i(t) \;=\; \int_0^t N_i(s)\, d\tilde{N}_i(s)$, yielding $\bar{N}_i \;=\; \bar{N}_i(\tau)$. Note that for the process $N_i(t)$, by assumption, only panel count data are available, and the available information is not enough to determine the whole process.

In contrast, for the process $\bar{N}_i(t)$, we have recurrent event data or complete information for the process and thus it is natural and convenient to base the estimating function on the process $\bar{N}_i(t)$.

It follows from (2.3) that $U_1(\beta_1)$ is an unbiased estimating function for β_1. Let $\tilde{\beta}_1$ denote the solution to $U_1(\beta_1 \,|\, \gamma, x_i's, z_i's) = 0$. Note that

$$\Gamma(\beta_1) = \frac{\partial U_1(\beta_1)}{\partial \beta_1} = -\frac{1}{n}\sum_{i=1}^{n} \{w_{1i}\, x_{1i}\, x_{1i}' \exp(\gamma' x_i)\, \exp(\beta_1' x_{1i})\}\,,$$

which is strictly negative definite. Thus the estimating equation given above has a unique solution. By using the same approach as that in Sun and Wei (2000), it can be easily shown that $\sqrt{n}\,(\tilde{\beta}_1 - \beta_1)$ converges in distribution to a zero mean normal random vector.

In practice, of course, γ and the z_i's are generally unknown and thus $U_1(\beta_1 \,|\, \gamma, x_i's, z_i's)$ is not available. To deal with this, we first consider estimation of γ. Define $K_i^* = \tilde{N}_i(C_i)$, the total number of observations on subject i, $i = 1, ..., n$. Let the s_j's denote the ordered and distinct time points of the observation times $\{T_{il}\}$, d_j the number of the observation times equal to s_j, and n_j the number of the observation times satisfying $T_{il} \le s_j \le C_i$, $i = 1, ..., n$. Also define $x_{2i} = (x_i', 1)'$, $\gamma_2' = (\gamma', \gamma_1) = (\gamma', \log E(Z))$. Then following Huang and Wang (2004), one can first estimate $\Lambda_0(t)$ and γ by

$$\widehat{\Lambda}_0(t) = \prod_{s_l > t} \left(1 - \frac{d_l}{n_l}\right)$$

and the estimating equation

$$\sum_{i=1}^{n} w_{2i}\, x_{2i} \left\{ K_i^* \, \widehat{\Lambda}_0^{-1}(C_i) \; - \; \exp(\gamma_2' x_{2i}) \right\} \; = \; 0 \; , \tag{2.4}$$

respectively. In equation (2.4), the w_{2i}'s are weights that could depend on x_i, C_i and Λ_0. Note that here as expected, we use only recurrent event data for model (2.2). A key fact used in the above estimation is that conditional on the observed data (x_i, C_i, z_i, K_i^*), the observation times $\{T_{i1}, ..., T_{iK_i^*}\}$ are the order statistics of a simple random sample of size K_i^* from the density function

$$\frac{z_i^{\alpha} \, \lambda_0(t) \exp(\gamma' x_i)}{z_i^{\alpha} \, \Lambda_0(C_i) \, \exp(\gamma' x_i)} \, I(0 \le t \le C_i) \; = \; \frac{\lambda_0(t)}{\Lambda_0(C_i)} \, I(0 \le t \le C_i) \; .$$

Let $\hat{\gamma}_2' \; = \; (\hat{\gamma}', \hat{\gamma}_1)$ denote the estimate of γ_2 given by equation (2.4). For the unknown z_i's, note that given (x_i, C_i, z_i), the expected value of K_i^* is equal to $z_i \, \Lambda_0(C_i) \, \exp(\gamma' x_i)$. This suggests that one can replace z_i by

$$\hat{z}_i \; = \; \frac{K_i^*}{\hat{\Lambda}_0(C_i) \, e^{\hat{\gamma}' x_i}}$$

in $U_1(\beta_1 \,|\, \gamma, x_i's, z_i's)$. Note that z_i cannot be consistently estimated and among others, Huang and Wang (2004) used the same approach for the treatment of the z_i's. Given $\hat{\gamma}$ and the $\hat{z}_i$'s, we propose to estimate β_1 by the solution to the estimating equation

$\hat{U_1}(\beta_1) = 0$, where

$$\hat{U_1}(\beta_1) = U_1(\beta_1 \,|\, \hat{\gamma}, x_i's, \hat{z}_i's) = \frac{1}{n}\sum_{i=1}^{n} w_{1i}\,\hat{x}_{1i}\,\left\{\bar{N}_i - \exp(\hat{\gamma}'x_i)\,\exp(\beta_1'\hat{x}_{1i})\,\right\} \qquad (2.5)$$

with $\hat{x}_{1i} = (x_i', \log\hat{z}_i, 1)'$, $i = 1, ..., n$.

Let $\hat{\beta}_1$ denote the estimator of β_1 defined above. Then it can be shown that $\hat{\beta}_1$ converges in probability to $\tilde{\beta}_1$ as $n \to \infty$ and is consistent. Furthermore, $\sqrt{n}\,(\hat{\beta}_1 - \beta_1)$ has asymptotically a normal distribution with mean zero and the covariance matrix $\phi^{-1}\,\Sigma\,(\phi^{-1})'$, where ϕ and Σ are given in Appendix A. The proof for the above results is sketched in Appendix A.

Now we consider the case where the C_i's may depend on the covariates x_i's. For this, following Sun and Wei (2000), we assume that for subject i, the hazard function of C_i has the form

$$\lambda_i^*(t) = \lambda_0^*(t)\,\exp(\psi'x_i)\,, \qquad (2.6)$$

where as $\lambda_0(t)$ and γ in model (2.2), $\lambda_0^*(t)$ is a completely unspecified baseline hazard function and ψ is a $p \times 1$ vector of parameters. To estimate β, motivated by $U_1(\beta_1 \,|\, \gamma, x_i's, z_i's)$ and equation (2.3), we consider the estimating function

$$U_2(\beta_2 \,|\, \gamma, \psi, S_0, x_i's, z_i's) = \frac{1}{n}\sum_{i=1}^{n} w_{1i}\,x_{1i}\,\left\{\int_0^\tau \frac{N_i(t)\,d\tilde{N}_i(t)}{[S_0(t)]^{\exp(\psi'x_i)}} - \exp(\gamma'x_i)\,\exp(\beta_2'x_{1i})\right\}$$

for given γ, ψ and the z_i's. In the above, the x_{1i}'s and w_{1i}'s are defined as before, $S_0(t) = \exp\{-\int_0^t \lambda_0^*(s)\,ds\}$ denotes the baseline survival function of the C_i's, and

$\beta_2' = (\beta', 1+\alpha, \log(\theta_1))$, where $\theta_1 = \int_0^\tau \lambda_0(t)\,\mu_0(t)\,dt$. It is easy to see that as $U_1(\beta_1)$, $U_2(\beta_2 \,|\, \gamma, \psi, S_0, x_i's, z_i's)$ has expectation zero.

As before, the parameter γ and the latent variables z_i's are unknown in practice and the same is true for ψ and S_0. It is obvious that one can deal with γ and the z_i's using the same approach discussed before. For ψ and S_0, note that they are defined in the proportional hazards model (2.6) with respect to the C_i's, for which we have complete data. It is thus natural to estimate ψ and S_0 using the partial likelihood estimate $\hat{\psi}$ defined as the solution to

$$\sum_{i=1}^{n} \int_0^\tau \left\{ x_i - \frac{\sum_{l=1}^n I(C_l \geq t)\, e^{\psi' x_l}\, x_l}{\sum_{l=1}^n I(C_l \geq t)\, e^{\psi' x_l}} \right\} d\, I(C_i \leq t) = 0 \tag{2.7}$$

and the consistent estimate

$$\hat{S}_0(t) = \exp\left\{ - \int_0^t \frac{\sum_{i=1}^n d\, I(C_i \leq s)}{\sum_{j=1}^n I(C_j \geq s)\, e^{\hat{\psi}' x_j}} \right\} \tag{2.8}$$

(Kalbfleisch and Prentice, 2002). Note that sometimes the C_i's may be right censored. In this case, $I(C_i \geq t)$ and $I(C_i \leq t)$ in (2.7) and (2.8) should be replaced by $I(C_i^* \geq t)$ and $I(C_i^* \leq t, C_i^* = C_i)$, respectively, where C_i^* denotes the smaller of C_i and the right-censoring time. Replacing γ, the z_i's, ψ and S_0 with $\hat{\gamma}$, the $\hat{z}_i$'s, $\hat{\psi}$ and $\hat{S}_0$, we propose to estimate β_2 and thus β by the solution to

$$\hat{U}_2(\beta_2) = U_2(\beta_2 \,|\, \hat{\gamma}, \hat{\psi}, \hat{S}_0, x_i's, \hat{z}_i's) = 0\,. \tag{2.9}$$

Let $\hat{\beta}_2$ denote the estimate defined above. By using similar approaches to those for $\hat{\beta}_1$, one can show that $\hat{\beta}_2$ is consistent and $\sqrt{n}\,(\hat{\beta}_2 - \beta_2)$ converges in probability to a normal random vector with mean zero and the covariance matrix $\phi^{*-1}\,\Sigma^*\,(\phi^{*-1})'$, where ϕ^* and Σ^* are given in Appendix A. The proof for these results is sketched in Appendix A.

For inference about β as well as α and γ, one needs consistent estimates of their covariance matrices. For this, following Wang *et al.* (2001), we propose to apply the simple bootstrap approach. Note that a more natural approach may be to derive some closed form estimates. However, it can be seen from Appendix A that the asymptotic covariance matrix of $\hat{\beta}_1$ involves several functions and quantities for which it is very difficult or impossible to derive consistent estimates. Furthermore, even if there exist such estimates, the resulting estimate of the asymptotic covariance matrix of $\hat{\beta}_1$ would be too complicated in computation to be useful compared to the bootstrap approach. The same is true for $\hat{\beta}_2$ and was encountered by Wang *et al.* (2001) and Huang and Wang (2004) for the analysis of recurrent event data, who also used the bootstrap procedure.

Sometimes one may be interested in estimation of $\mu_0(t)$. For this, note that for subject i, based on model (2.1), a natural estimate of the rate function $d\,\mu_0(t)$ is given by the empirical estimate

$$d\,\hat{\mu}_{i0}(t;\hat{\beta},\hat{\gamma},x_i,\hat{z}_i) \;=\; \sum_{l=1}^{K_i} \frac{N_i(T_{il}) \;-\; N_i(T_{i\,l-1})}{T_{il} \;-\; T_{i\,l-1}}\,\hat{z}_i^{-\hat{\alpha}}\,e^{-\hat{\beta}'x_i}\,I(T_{i\,l-1} < t \le T_{il} \le C_i^*)\,,$$

where $T_{i0} = 0$, $i = 1, ..., n$. A similar estimate was used in Thall and Lachin (1988). This yields an estimate of $\mu_0(t)$ given by

$$\hat{\mu}_0(t) = \int_0^t \frac{\sum_{i=1}^n d\hat{\mu}_{i0}(s; \hat{\beta}, \hat{\gamma}, x_i, \hat{z}_i)}{\sum_{i=1}^n I(s \leq C_i^*)} , \tag{2.10}$$

for $0 \leq t \leq \max\{T_{il}; T_{il} \leq C_i^*\}$. Note that for the special case where all subjects have the same observation times, $\hat{\mu}_0(t)$ is a step function with jumps at these observation times and its values at these time points equal to the sample means of the N_i's at the points adjusted by the $\hat{z}_i^{-\hat{\alpha}} e^{\hat{\beta}' x_i}$'s.

2.4 Numerical Results

We conducted a simulation study to evaluate the performance of the proposed estimation procedures under different situations with the focus on estimation of β and the case where the distribution of the C_i's is independent of the x_i's. In the study, we assumed that the covariate x_i's follow either a Bernoulli distribution with success probability 0.5 or normal distribution with mean 0 and variance 0.25. The subject-specific frailty z_i and follow-up time C_i were generated from a gamma distribution with mean 10 and variance 50 and the uniform distribution over $(\tau/2, \tau)$ with $\tau = 18$, respectively. For the observation process, we considered two situations. One is to assume that H_i is a homogeneous Poisson process with $\lambda_0(t) = 1/\tau$, and the other is to suppose that H_i is a non-homogeneous Poisson process with $\lambda_0(t) = (t + 1)/\{\tau(\tau/2 + 1)\}$. For the first set-up, we have that given x_i and z_i, K_i^*, the number

of real observation times for subject i, follows the Poisson distribution with mean

$$\Lambda(C_i|x_i, z_i) \;=\; z_i\,\Lambda_0(C_i)\,\exp(\gamma x_i) \;=\; \frac{z_i\,C_i\,\exp(\gamma x_i)}{\tau}\;,$$

$i \;=\; 1, ..., n$. Furthermore, in this case, the real observation times $(T_{i1}, ..., T_{iK_i^*})$ are the order statistics of a random sample of size K_i^* from the uniform distribution over $(0, C_i)$.

For the second set-up about the observation process, given x_i and z_i, K_i^* follows the Poisson distribution with mean

$$\Lambda(C_i|x_i, z_i) \;=\; z_i\,\Lambda_0(C_i)\,\exp(\gamma x_i) \;=\; \frac{z_i\,(C_i^2/2 + C_i)\,\exp(\gamma x_i)}{\tau\,(\tau/2 + 1)}$$

and $(T_{i1}, ..., T_{iK_i^*})$ are the order statistics of a random sample of size K_i^* from the density function

$$\frac{t^2/2 \;+\; t}{C_i^2/2 + C_i}\,I(0 \;\leq t \;\leq C_i)\;.$$

Finally we generated panel count data $N_i(T_{ij})$'s by taking $N_i(T_{ij}) \;=\; N_i^*(T_{i1}) \;+\; N_i^*(T_{i2} - T_{i1}) \;+\; ... \;+\; N_i^*(T_{ij} - T_{i\,j-1})$, where

$$N_i^*(t) \;\sim\; \mathit{Poisson}(\,z_i^{\alpha}\,t\,\exp(\beta\,x_i)\,)\;,$$

$j \;=\; 1, ..., K_i^*$, $i \;=\; 1, ..., n$. Here we took $\mu_0(t) \;=\; t$. The results given below are based on $n \;=\; 100$ and 1000 replications with the size of bootstrap samples taken to be 50.

Table 2.1 presents simulation results for situations where H_i is a homogeneous Poisson process, $\gamma = 1$, $\alpha = 0.5$, and the x_i's follow the Bernoulli or normal distribution. In the table, the true value of β was set to be -2, -1, 0, 1, or 2. The table gives the averages of estimates of β based on simulated data, the sample standard deviations of the estimates (SSD), and the means of bootstrap standard deviation estimates (BSD). The results for the case where H_i is a nonhomogeneous Poisson process are given in Table 2.2 in which other set-ups are the same as in Table 2.1. The simulation results suggest that the proposed estimation procedure based on the function given in (2.5) performs reasonably well for the situation considered here. In particular, the bootstrap variance estimates and the sample variance estimates are quite close to each other with the former tending to be smaller than the latter. The study of quantile plots of the standardized estimates of β indicates that this does not seem to be a problem.

To assess the performance of the normal approximation to the finite distribution of the estimate of β, we studied the quantile plots of the standardized estimates of β. Figures 2.1 and 2.2 display such plots corresponding to the situations where the true value of β is equal to 0, the x_i's follow the Bernoulli distribution, and H_i is homogeneous and nonhomogeneous Poisson processes, respectively. These figures indicate that the normal approximation is good, and the quantile plots for other set-ups are similar. In the simulation, we investigated the effect of the size M of bootstrap samples on variance estimation and $M = 50$ used here seems reasonable. We also considered other set-ups in the simulation study and obtained similar results.

We also considered the situation where the distribution of the C_i's depends on the

x_i's. Table 2.3 gives the results for the proposed estimate of β based on the estimating equation (2.9) obtained under the same set-up as that for Table 2.1 except that the C_i's were assumed to follow model (2.6) with $\lambda_0^*(t) = t/800$ and $\psi = 0.5$ and to be right-censored at $\tau = 18$. The table contains the same quantities as those given in Table 2.1. It can be seen that the obtained results are similar to those presented in Table 2.1 and indicate that the proposed estimation procedure based on the estimating equation (2.9) seems to perform well.

2.5 An Illustrative Example

To illustrate the estimation procedures given in the previous sections, we apply them to the bladder cancer study discussed in Section 1.1.2.3. Here we focus on the 47 bladder cancer patients in the placebo group.

To analyze the data, for patient i, define x_{i1} to be the number of initial tumors and x_{i2} to be the size of the largest initial tumor, $i = 1, ..., 47$. Assume that the occurrence process of the bladder tumors and the visit process can be described by models (2.1) and (2.2). First we investigated if the distribution of the C_i's may depend on the x_i's. For this, we assumed that they can be described by model (2.6). The application of the partial likelihood procedure based on (2.7) yielded $\hat{\psi}_1 = -0.0199$ and $\hat{\psi}_2 = -0.1115$ with estimated standard deviations of 0.128 and 0.100, respectively. These results suggest that the distribution of the C_i's does not seem to be dependent on the x_i's and thus for estimation of β, we can use the procedure based on $\hat{U}_1(\beta_1)$ defined in (2.5).

By applying the estimation procedure based on $\hat{U}_1(\beta_1)$, we obtained $\hat{\beta}_t = 0.1213$,

$\hat{\beta}_s = -0.0044$, $\hat{\gamma}_t = 0.0165$, and $\hat{\gamma}_s = 0.0139$ with the estimated standard deviations being 0.068, 0.088, 0.059 and 0.063, respectively. Here β_t and γ_t represent regression coefficients corresponding to the number of initial tumors, while β_s and γ_s corresponding to the size of the largest initial tumor. These results suggest that the number of initial tumors seems to be positively and significantly related with the tumor recurrence rate but has no significant effect on the visit process. That is, the higher number of initial tumors implies the higher tumor recurrence rate. The size of the largest initial tumor does not seem to have significant effects on both tumor recurrence rate and visit process.

For the parameter α, which represents the correlation between the tumor recurrence and visit processes, we obtained $\hat{\alpha} = -0.5067$ with estimated standard error of 0.085. The result indicates that bladder tumor recurrence and patient visit are negatively correlated, meaning that the patients who visited more often had a smaller tumor recurrence rate given other factors. This is consistent with the sample correlation obtained in Section 2.1. One possible reason for this is that the more visits a patient had, the less time the patient had for tumor recurrence and, in consequence, the smaller number of initial tumors for next visit and thus a lower tumor recurrence rate as suggested above. To further investigate this, we divided the patients into two groups with roughly equal numbers of the patients, rare visit and frequent visit groups, based on the total number of visits. In the rare visit group, every patient had at most 9 visits, while all patients in the frequent visit group had more than 9 visits. Figure 2.3 displays the separate estimates (2.10) of the baseline mean functions of the tumor recurrence

processes for the two groups and suggests that as shown above, the patients in the rare visit group had much higher tumor recurrence rate than those in the frequent visit group.

For comparison, we also analyzed the data using the approach given in Sun and Wei (2000), which assumes that the tumor recurrence and visit processes are independent given covariates. The approach yielded $\hat{\beta}_t = 0.1670$ and $\hat{\beta}_s = 0.0175$ with the estimated standard deviations being 0.172 and 0.184. It can be seen that without taking into account the correlation between the tumor recurrence and visit processes, one could get some misleading result about or actually miss the effect of the number of initial tumors on tumor recurrence rate. In other words, both the correlation and the effect exist and need to be taken into account in the analysis.

2.6 Discussion

In the preceding sections, estimating equation approaches were proposed for regression analysis of panel count data. A key feature of these approaches is that they allow for the dependence between the point process of interest and the observation process in contrast to existing methods, which assume that the two processes are independent given covariates. The simulation study suggests that the presented estimation procedures seem to perform reasonably well for practical situations.

For the association of the point process of interest and the observation process,

instead of model (2.1), a more general model is given by

$$\mu_i(t) \;=\; g(z_i;\varphi)\,\mu_0(t)\exp(\beta' x_i)\;,$$

where g is a known function depending on some unknown parameter φ and both $\mu_0(t)$ and β are defined as in model (2.1). For inference, one can develop procedures similar to those given in the previous sections. For the estimation approaches given in the preceding sections, there exist several limitations. One limitation is that H_i has been assumed to be a Poisson process. It would be useful to generalize the approach to situations where H_i is a point process with a mean function similar to (2.1). Note that for H_i, we have complete recurrent event data and thus it is relatively easy to check the Poisson assumption in practice.

Chapter 3

REGRESSION ANALYSIS OF MULTIVARIATE PANEL COUNT DATA

3.1 Introduction

As discussed before, panel count data arise in studies of recurrent events when each subject is observed only at finite discrete time points instead of continuously (Sun and Wei, 2000; Zhang, 2002). In such settings observations are taken at several distinct time points and only the number of events that occurred between observation times is known; no information is available on subjects between the observation time points. Multivariate panel count data arise if more than one type of recurrent events are of interest (Chen *et al.*, 2005).

Multivariate panel count data arise in studies involving several types of recurrent events in which patients are examined only at periodic follow-up assessments. Chen *et al.* (2005) described a study of patients with advanced cancer where the events are the development of different types of metastatic bone lesions that are only detectable by

bone scans of the entire skeleton carried out when patients visit participating clinical centers. The number of examinations varied from patient to patient, and at each examination, the number of new lesions developed since the previous examination was recorded. Three types of bone lesions arose in these patients, and the interest was in making statements about their respective rates of occurrences of these different events and related covariate effects. Another common example arises in tumorigenicity experiments when several types of tumors can occur together and are of interest. Below we consider a third example arising from a cohort study of patients with psoriatic arthritis conducted at the University of Toronto Psoriatic Arthritis Clinic where the event of interest is the development of joint damage. Clinicians are interested in damage as measured by radiographic changes as well as loss in function as detected by clinical examination, and these constitute the two types of events.

Several authors have considered the analysis of univariate panel count data. For example, Thall and Lachin (1988) and Sun and Kalbfleisch (1993) discussed the treatment comparison problem when only panel count data are available. Sun and Kalbfleisch (1995) and Wellner and Zhang (2000) investigated nonparametric estimation of the cumulative mean function of the underlying point process that generates panel count data. Sun and Wei (2000) and Zhang (2002) gave some approaches for regression analysis of panel count data. For multivariate panel count data, Chen *et al.* (2005) proposed two approaches based on a mixed Poisson model with piecewise constant baseline intensities. One approach assumes that the different types of recurrent events are related through multivariate log-normal random effects and bases inference on the

resulting full likelihood, while the other makes use of a marginal model approach. In the following, a marginal approach is presented that avoids the Poisson and piecewise constant baseline intensity assumptions.

With multivariate panel count data, one may separately apply methods for univariate panel count data to each type of event. As with multivariate failure time data (Wei *et al.*, 1989; Cai and Prentice, 1995), it is apparent that this would be less efficient than conducting a joint or multivariate analysis if the different types of recurrent events are related and associated covariate effects are the same. Multivariate analyses can, however, also be more efficient even if some covariate effects are different. Finally, separate univariate analyses, unlike multivariate analyses, cannot estimate the correlations between different covariate effects. More discussion on this is given below.

The remainder of this chapter is organized as follows. We begin with introducing some notation and models that are used throughout this chapter in Section 3.2. In particular, marginal mean models are employed for the underlying counting processes that characterize panel count data and observation times, respectively. One major advantage of these models is that they leave the dependence structures for related types of recurrent events completely arbitrary. Section 3.3 discusses estimation of covariate effects and for this, some estimating equations are proposed to give consistent and asymptotically normal estimates of regression parameters. In Section 3.4, some results from simulation studies conducted for evaluation of the proposed estimates are presented and suggest that the presented inference approach seems to work well for practical situations. Section 3.5 applies the method to the psoriatic arthritis study

discussed in Section 1.1.2.4, and Section 3.6 gives some concluding remarks.

3.2 Models and Notation

Consider a recurrent event study that involves n independent subjects, and suppose that each subject may experience K different types of events. For subject i, let $N_{ik}(t)$ denote the total number of the kth type events that have occurred up to time t, $0 \leq t \leq \tau$, where τ is a known constant representing study length, $i = 1, ..., n$, $k = 1, ..., K$. Also for each i, suppose that there exists a positive random variable C_i representing the censoring or follow-up time on subject i and a $p \times 1$ vector of covariates denoted by $x_i = (x_{i1}, ..., x_{ip})'$ that may affect the rate of occurrence of type k events. Here, for simplicity of presentation, we assume that the follow-up time or observation period and the covariates that may affect $N_{ik}(t)$ are the same for different types of recurrent events. The inference approach presented below can be easily generalized to situations where C_i and x_i may differ for different types of recurrent events. Define $Y_i(t) = I(t \leq C_i)$, indicating if subject i is at risk of experiencing the recurrent events at time t, $i = 1, ..., n$, $k = 1, ..., K$.

For the effects of covariates on $N_{ik}(t)$, we assume that given x_i, the marginal mean function of $N_{ik}(t)$ has the form

$$E\{N_{ik}(t) \,|\, x_i\} = \mu_k(t)\, g_N(x_i' \beta_0)\,. \tag{3.1}$$

In the model above, $\mu_k(t)$ is an unknown continuous baseline mean function, β_0 is a

$p \times 1$ vector of regression parameters representing the effect of x_i on $N_{ik}(t)$, and $g_N(\cdot)$ is a known, positive function that is assumed to be strictly increasing and twice differentiable. One common choice for $g_N(\cdot)$ is $g_N(x) = \exp(x)$, the exponential function. Other functions that are often used include $g_N(x) = 1 + x$ and $g_N(x) = \log(1 + e^x)$. Model (3.1) assumes that the baseline mean functions can be different for different types of recurrent events, but the effects of covariates on different types of recurrent events are the same. Some comments are given below for the situation where these effects may be different. The goal here is to estimate regression parameter β_0.

For estimation of β_0, we assume that only panel count data are available for the $N_{ik}(t)$'s. Specifically, suppose that $N_{ik}(\cdot)$ is observed only at finite time points $T_{ik,1} < ... < T_{ik,m_{ik}}$, where m_{ik} denotes the potential or scheduled number of observations on the kth type of recurrent event for subject i, $i = 1, ..., n$, $k = 1, ..., K$. That is, the observed data have the form

$$\{\, T_{ik,\ell}, N_{ik}(T_{ik,\ell}), C_i, x_i, m_{ik}\,;\, \ell = 1, ..., m_{ik}, k = 1, ..., K, n = 1, ..., n\,\}\,.$$

For each i and k, define $\tilde{N}_{ik}(t) = H_{ik}\{\min(t, C_i)\}$, where $H_{ik}(t) = \sum_{\ell=1}^{m_{ik}} I(T_{ik,\ell} \le t)$. Then $\tilde{N}_{ik}(t)$ is a point process characterizing the observation process on subject i with respect to the kth type recurrent event and jumps by one only at the observation times on N_{ik}. In the following, we assume that H_{ik} is a counting process with the marginal mean function

$$E\{H_{ik}(t) \,|\, x_i\} = \nu_k(t)\, g_H(x_i'\gamma_0) \tag{3.2}$$

given x_i. Here as with model (3.1), $\nu_k(t)$ is a completely unknown continuous baseline mean function, γ_0 denotes the effect of covariates on H_{ik}, and g_H is a known, positive function that is assumed to be strictly increasing and twice differentiable.

In the next section, some estimating equations are developed for estimation of β_0 along with γ_0. Note that for both N_{ik}, the process of interest, and H_{ik}, the observation process, only marginal mean functions are specified, and no assumption is made about the relationship among $N_{i1}, ..., N_{iK}$ or $H_{i1}, ..., H_{iK}$. In the following, we assume that the C_i's follow the same distribution function.

3.3 Estimation Procedures

For estimation of β_0, we first consider situations in which covariates have no effect on the observation process, that is, $\gamma_0 = 0$ or g_H is constant. The estimation procedure is then generalized to situations where the observation process may depend on covariates. For both situations, the focus will be on developing estimating equations for the regression parameters that do not involve the baseline mean functions μ_k and ν_k.

3.3.1 Estimation with Covariate-Independent Observation Processes

For simplicity, here we assume that $g_H(x) = 1$, and hence that the observation process is independent of the covariates. For each i and k, we define

$$\bar{N}_{ik} = \sum_{\ell=1}^{m_{ik}} N_{ik}(T_{ik,\ell})\, I(T_{ik,\ell} \leq C_i) = \int_0^\tau N_{ik}(t)\, d\tilde{N}_{ik}(t)\,,$$

$i = 1, ..., n$, $k = 1, ..., K$. Then conditional on x_i and under models (3.1) and (3.2), we have

$$E\{\bar{N}_{ik} \,|\, x_i\} \,=\, \alpha_k \, g_N(x_i'\beta_0) \,,$$

where $\alpha_k \,=\, \int_0^\tau \, \mu_k(t) \, P(C_i \geq t) \, d\nu_k(t)$. Without loss of generality, suppose that the covariates x_i's are centered. Then by following Sun and Wei (2000), a natural, unbiased estimating function for β_0 is given by

$$U_n(\beta) \,=\, \frac{1}{\sqrt{n}} \sum_{k=1}^{K} \sum_{i=1}^{n} \, x_i \, \bar{N}_{ik} \, \{g_N(x_i'\beta)\}^{-1} \,, \tag{3.3}$$

which is the same as that given in Sun and Wei (2000) if $K \,=\, 1$ and $g_N(t) \,=\, \exp(t)$.

Define the estimate $\hat{\beta}_1$ of β_0 as the solution to $U_n(\beta) \,=\, 0$. Let $g_N^{(0)} \,=\, g_N$ and $g_N^{(r)}$ denote the rth derivative of g_N, $r \,=\, 1, 2$, so that

$$\frac{\partial U_n(\beta)}{\partial \beta} \,=\, -\, \frac{1}{\sqrt{n}} \sum_{k=1}^{K} \sum_{i=1}^{n} \, x_i \, x_i' \, \bar{N}_{ik} \, g_N^{(1)}(x_i'\beta) \, \{g_N(x_i'\beta)\}^{-2} \,,$$

which is strictly negative definite since g_N is strictly increasing. Thus the equation $U_n(\beta) \,=\, 0$ has a unique solution and $\hat{\beta}_1$ is consistent.

To derive the asymptotic distribution of $\hat{\beta}_1$, we define $U_i^*(\beta) = \sum_{k=1}^{K} \, x_i \, \bar{N}_{ik} \, \{g_N(x_i'\beta)\}^{-1}$. Then $U_n(\beta) \,=\, n^{-1/2} \, \sum_{i=1}^{n} \, U_i^*(\beta)$, which is the summation of i.i.d random variables. Thus it can be easily shown that $\sqrt{n} \, (\hat{\beta}_1 \,-\, \beta_0)$ converges in distribution to a multivariate normal vector with mean zero and covariance matrix $\Sigma_1 \,=\, \phi^{-1} \, \Sigma_u \, \phi^{-1}$, where $\phi \,=\, -\, E\{\partial U_i^*(\beta_0)/\partial\beta\}$ and $\Sigma_u \,=\, Cov\{U_i^*(\beta_0)\}$. A consistent estimate of Σ_1 is given

by $\hat{\Sigma}_1 = \hat{\phi}^{-1}\,\hat{\Sigma}_u\,\hat{\phi}^{-1}$ (Wei *et al.*, 1989), where

$$\hat{\phi} = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} x_i\,x_i'\,\bar{N}_{ik}\,g_N^{(1)}(x_i'\hat{\beta}_1)\,\left\{g_N(x_i'\hat{\beta}_1)\right\}^{-2}$$

and

$$\hat{\Sigma}_u = \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{k=1}^{K} x_i\,\bar{N}_{ik}\,\left\{g_N(x_i'\hat{\beta}_1)\right\}^{-1}\right]^{\otimes 2}$$

with $a^{\otimes 2} = a\,a'$ for a vector a.

3.3.2 Estimation with Covariate-Dependent Observation Processes

Consider now the more general situation in which the observation process $H_{ik}(t)$ is affected by the covariates as well as $N_{ik}(t)$. If $\bar{N}_{ik}$ is defined as before, under models (3.1) and (3.2), we have

$$E\{\bar{N}_{ik}\,|\,x_i\} = \alpha_k\,g_N(x_i'\beta_0)\,g_H(x_i'\gamma_0)$$

given x_i, where α_k is defined earlier. If γ_0 is known, this along with $U_n(\beta)$ suggests to estimate β_0 based on the estimating equation $U_n(\beta,\gamma_0) = 0$, where

$$U_n(\beta,\gamma) = \frac{1}{\sqrt{n}}\sum_{k=1}^{K}\sum_{i=1}^{n} x_i\,\bar{N}_{ik}\,\{g_N(x_i'\beta)\}^{-1}\,\{g_H(x_i'\gamma)\}^{-1}\,. \tag{3.4}$$

In practice, of course, the parameter γ_0 is unknown. Unlike for $N_{ik}(t)$, however, here we have complete recurrent event data (Cai and Schaubel, 2004a) for the observation process $\{\tilde{N}_{ik}(s), 0 < s < C_i\}$, making estimation of γ_0 relatively easy. To see this,

define

$$S_k^{(d)}(t;\gamma) \;=\; \frac{1}{n}\sum_{i=1}^{n} Y_i(t)\, x_i^{\otimes d}\, g_H^{(d)}(x_i'\gamma)\,,$$

for $d = 0, 1, 2$, and

$$S_k^{(3)}(t;\gamma) \;=\; \frac{1}{n}\sum_{i=1}^{n} Y_i(t)\, x_i^{\otimes 2}\, \left\{g_H^{(1)}(x_i'\gamma)\right\}^2 \left\{g_H(x_i'\gamma)\right\}^{-1}\,,$$

$k = 1, ..., K$, where $Y_i(t) = I(t \le C_i)$. Also for $k = 1, ..., K$, define

$$E_k(t;\gamma) \;=\; \frac{S_k^{(1)}(t;\gamma)}{S_k^{(0)}(t;\gamma)}$$

and

$$V_k(t;\gamma) \;=\; \frac{S_k^{(3)}(t;\gamma)}{S_k^{(0)}(t;\gamma)} \;-\; E_k(t;\gamma)^{\otimes 2}\,,$$

and suppose that the limits of $S_k^{(d)}(t;\gamma)$, $E_k(t;\gamma)$ and $V_k(s;\gamma)$ exist.

Following Cai and Schaubel (2004b), we propose to estimate γ_0 by $\hat{\gamma}$, the solution to the estimating equation $H_n(\gamma) = 0$, where

$$H_n(\gamma) \;=\; \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^\tau \left\{x_i\,\frac{g_H^{(1)}(x_i'\gamma)}{g_H(x_i'\gamma)} \;-\; E_k(s;\gamma)\right\} d\tilde{N}_{ik}(s)\,. \tag{3.5}$$

Note that estimation of γ_0 depends only on the observed information on the observation process $\tilde{N}_{ik}$. It is then natural to estimate β_0 by $\hat{\beta}_2$ defined as the solution to $U_n(\beta, \hat{\gamma}) = 0$.

For the asymptotic properties of $\hat{\beta}_2$ and $\hat{\gamma}$, Cai and Schaubel (2004b) proved that

under mild regularity conditions, $\hat{\gamma}$ is unique and consistent and asymptotically follows a normal distribution. For $\hat{\beta}_2$, it can be easily shown that it is unique and consistent as $\hat{\beta}_1$. For the asymptotic distribution of $\hat{\beta}_2$, we show in Appendix B that under some regularity conditions described there, $\sqrt{n}\,(\hat{\beta}_2 - \beta_0)$ asymptotically follows a multivariate normal distribution with mean zero and variance matrix that can be consistently estimated by

$$\hat{\Sigma}_2 \;=\; \hat{F}^{-1}\,\hat{G}_1\,\hat{\Gamma}\,\hat{G}_1'\,\hat{F}'^{-1}\,. \tag{3.6}$$

In this formula, $\hat{G}_1 \;=\; (\,I_p, -\hat{D}\hat{A}^{-1}(\hat{\gamma})\,)$,

$$\hat{F} \;=\; -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\left\{g_N(x_i'\hat{\beta}_2)\right\}^{-2}\,g_N^{(1)}(x_i'\hat{\beta}_2)\;\{g_H(x_i'\hat{\gamma})\}^{-1}\;\bar{N}_{ik}x_i\,x_i'\,,$$

$$\hat{D} \;=\; -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\left\{g_N(x_i'\hat{\beta}_2)\right\}^{-1}\,g_H^{(1)}(x_i'\hat{\gamma})\;\{g_H(x_i'\hat{\gamma})\}^{-2}\;\bar{N}_{ik}\,x_i\,x_i'\,,$$

$$\hat{A}(\gamma) \;=\; -\frac{1}{n}\sum_{k=1}^{K}\int_0^{\tau} V_k(t;\gamma)\,d\tilde{N}_{\cdot k}(t)\,,$$

and $\hat{\Gamma}$ is given in Appendix B, where I_p denotes the $p\times p$ identity matrix and $\tilde{N}_{\cdot k}(t) \;=\; \sum_{i=1}^{n}\,\tilde{N}_{ik}(t)$.

One may sometimes be interested in the joint distribution of $\hat{\beta}_2$ and $\hat{\gamma}$, and estimating the baseline mean functions $\mu_k(t)$ and $\nu_k(t)$, $k = 1,\ldots,K$. In Appendix B we show that one can asymptotically approximate the joint distribution of $\sqrt{n}\,(\hat{\beta}_2 - \beta_0)$ and $\sqrt{n}\,(\hat{\gamma} - \gamma_0)$ by a multivariate normal distribution with mean zero and covariance

matrix

$$-\hat{F}^{-1}\,\hat{G}_1\,\hat{\Gamma}\,\hat{G}_0'\,, \tag{3.7}$$

where $\hat{G}_0 = (\,\mathbf{0}_p, -\hat{A}^{-1}(\hat{\gamma})\,)$ with $\mathbf{0}_p$ denoting the $p \times p$ zero matrix.

Given $\hat{\beta}_2$ and $\hat{\gamma}$, one may want to estimate the baseline mean functions $\mu_k(t)$ and $\nu_k(t)$, $k = 1, ..., K$. The estimation of $\nu_k(t)$ is relatively easy, since one has recurrent event data, and in particular, a consistent estimate of it is given by

$$\hat{\nu}_k(t;\hat{\gamma}) = \int_0^t \frac{d\tilde{N}_{\cdot\,k}(s)}{n\,S_k^{(0)}(s;\hat{\gamma})}$$

(Andersen *et al.*, 1993), $k = 1, ..., K$. For $\mu_k(t)$, first consider the estimation with respect to each subject. Given x_i and under model (3.1), a natural estimate of the rate function $d\,\mu_k(t)$ is the empirical estimate

$$d\,\hat{\mu}_{ik}(t;\hat{\beta}_2) = \sum_{\ell=1}^{m_{ik}} \frac{N_{ik}(t_{ik,\ell}) - N_{ik}(t_{ik,\ell-1})}{t_{ik,\ell} - t_{ik,\ell-1}} \left\{g_N(x_i'\hat{\beta}_2)\right\}^{-1} I(t_{ik,\ell-1} < t \le t_{ik,\ell})\,,$$

where $t_{ik,0} = 0$, $i = 1, ..., n$, $k = 1, ..., K$. Thall and Lachin (1988) considered the same estimate. This leads to an estimate of $\mu_k(t)$ given by

$$\hat{\mu}_k(t;\hat{\beta}_2) = \int_0^t d\,\hat{\mu}_k(s) = \int_0^t \frac{\sum_{i=1}^n d\hat{\mu}_{ik}(s)}{\sum_{i=1}^n I(s \le t_{ik,m_{ik}})} \tag{3.8}$$

for $0 \le t \le \max\{t_{ik,m_{ik}}\}$, $k = 1, ..., K$. Note that a similar, natural estimate of $\mu_k(t)$

is given by

$$\frac{\sum_{i=1}^{n} \sum_{\ell=1}^{m_{ik}} N_{ik}(t_{ik,\ell}) \left\{ g_N(x_i'\hat{\beta}_2) \right\}^{-1} I(t_{ik,\ell-1} < t \le t_{ik,\ell})}{\sum_{i=1}^{n} I(t \le t_{ik,m_{ik}})}$$

or

$$\frac{\sum_{i=1}^{n} \sum_{\ell=1}^{m_{ik}} N_{ik}(t_{ik,\ell-1}) \left\{ g_N(x_i'\hat{\beta}_2) \right\}^{-1} I(t_{ik,\ell-1} < t \le t_{ik,\ell})}{\sum_{i=1}^{n} I(t \le t_{ik,m_{ik}})},$$

the weighted sample mean estimates. However, unless the m_{ik}'s are large, the former can seriously overestimate $\mu_k(t)$ while the latter can underestimate $\mu_k(t)$.

3.4 Simulation Studies

This section reports some results obtained from simulation studies conducted to assess the performance of the estimation procedures proposed in the previous sections under various situations. In the simulation studies, we focused on the two sample comparison problem with $x_i = -1$ for half subjects and 1 for the others, and first generated the follow-up times C_i's from the uniform distribution over $(\tau/3, \tau)$, where τ is a positive constant. For the observation times $T_{ik,\ell}$'s, it was assumed that $H_{ik}(t)$ follows the mixed effects marginal mean model

$$E\{ H_{ik}(t) \,|\, x_i, P_i \} = P_i \, g_H(x_i\gamma_0) \, \nu_k(t) \,,$$

given x_i and random effect P_i, $i = 1, ..., n$, $k = 1, 2$. In the model above, the P_i's were generated from the gamma distribution with mean one and variance σ_P^2, $\nu_1(t) = 0.5\, t$,

$\nu_2(t) = t$, and $g_H(x)$ was set to be the identity function for the case in which the observation time is independent of covariates and $g_H(x) = e^x$ if the observation time is dependent of covariates. Then by following Cai and Schaubel (2004b), the observation times $(T_{ik,1}, ..., T_{ik,m_{ik}})$ were defined as

$$T_{ik,\ell+1} = T_{ik,\ell} - \log(U_{ik,\ell+1}) \{P_i \, g_H(x_i\gamma_0) \, d\nu_k\}^{-1},$$

where $T_{ik,0} = 0$, $U_{ik,\ell} \sim U(0,1)$, and m_{ik} is the number of the $T_{ik,\ell}$'s that fall within $(0, C_i)$.

Given m_{ik} and $(T_{ik,1}, ..., T_{ik,m_{ik}})$, we generated the panel count data $N_{ik}(T_{ik,\ell})$'s from the mixed Poisson processes as

$$N_{ik}(T_{ik,\ell}) = N^*_{ik}\left[\mu_k(T_{ik,1})\right] + N^*_{ik}\left[\mu_k(T_{ik,2}) - \mu_k(T_{ik,1})\right] + ... + N^*_{ik}\left[\mu_k(T_{ik,\ell}) - \mu_k(T_{ik,\ell-1})\right],$$

$\ell = 1, ..., m_{ik}$, $k = 1, 2$, $i = 1, ..., n$. In the above, $N^*_{ik}[\mu_k(t)]$ denotes the random number generated from the Poisson distribution with mean $Q_i \, \mu_k(t) \exp(x_i\beta_0)$, where Q_i is the random number arising from the gamma distribution with mean one and variance σ^2_Q, $\mu_1(t) = t$, and $\mu_2 = t^2$. The results given below are based on $n = 100$, or 200, $\tau = 12$, $\sigma_P = 1$, $\gamma_0 = 1$, and 1000 replications.

Table 3.1 presents the simulation results obtained for covariate-independent observation process situations with $\beta_0 = -1$, 0, or 1 and $\sigma^2_Q = 1$ or 2. The table includes the estimated bias (BIAS) given by the mean of the point estimate $\hat{\beta}_1$ minus the true value of β_0, the average of the standard deviation estimates (SEE), the sample standard

deviation (SSE), and the empirical 95% coverage probability (CP) for β_0. It can be seen that the estimated covariate effect seems unbiased and the two standard deviation estimates seem quite close to each other, suggesting that the proposed variance estimate is reasonable. The table also shows that as expected, the results become better when the sample size increases. Note that σ_Q^2 measures the correlation between the two recurrent processes N_{i1} and N_{i2}. The simulation results indicate that as expected, the covariate effect can be estimated more efficiently for smaller σ_Q^2, meaning that the two recurrent processes are less related.

The simulation results for covariate-dependent observation processes or $\hat{\beta}_2$ are given in Table 3.2, which presents the same quantities as in Table 3.1. Note that here we only give the results about β_0 because as shown in Cai and Schaubel (2004b), the estimation procedure for γ_0 seems to work as well as the estimation procedure for β_0. Table 3.2 basically gives the same results as in Table 3.1 and again suggests that as given in Section 3.3.1, the proposed estimation procedure in Section 3.3.2 also seems to work reasonably well for the situations considered here. To evaluate the normal approximation to the finite-sample distribution of $\hat{\beta}_1$ or $\hat{\beta}_2$, we studied the quantile plots of the standardized estimates against the standard normal distribution. These plots, which are not given here, indicate that the normal approximation seems good.

3.5 An Application

Now we consider the application of the proposed method to the psoriatic arthritis study described in Section 1.1.2.4. We restrict our attention to 177 female patients

having a baseline, at least one follow-up assessment and complete covariate data.

Figure 3.1 contains a timeline diagram for a sample of 10 patients. The length of the horizontal lines indicates the duration of time from clinic entry to last contact for functional assessments (solid lines) and radiological assessments (dashed lines). The vertical dashes indicate the respective assessment times and the corresponding numbers indicate the number of newly damaged joints detected since the last assessment by the corresponding method of assessment. The figure reveals that the two methods of assessment are sometimes coincident, but often not, and there can be large gaps during which no assessments are available (see individuals A, B and E). To give an idea about the relationship between the two types of events, Figure 3.2 contains a scatter plot of the crude event rates defined as the number of damaged joints detected from clinic entry to the last assessment divided by the time since clinic entry to the last assessment for all patients (the dashed line has slope 1). There is a slight tendency for the crude rates of radiological damage to be higher than those for functional damage, but it should be noted that there is considerable sampling variability in these rates.

The data on damaged joint counts as defined by functional and radiological criteria form bivariate panel count data; our interest lies in estimating covariate effects on the respective rate functions and estimating the cumulative mean number of damaged joints according to each criterion. Of course, one way for this is to apply the methods developed for univariate panel count data to each type of damaged joint counts separately. On the other hand, Figure 3.2 indicates that the two types of damaged joints are correlated and thus a joint analysis would be preferred as discussed earlier.

For the analysis, define $N_{i1}(t)$ and $N_{i2}(t)$ as the cumulative numbers of radiologically and functionally damaged joints up to time t for patient i, respectively, $i = 1, ..., 177$. Also for the ith patient, define $x_{i1} = 1$ if he or she had a family history of psoriasis and 0 otherwise, and x_{i2} and x_{i3} to be equal to the arthritis duration and the number of active joints at clinic entry, respectively. Table 3.3 gives the joint analysis results obtained by the application of the method proposed in Section 3.3.2 and includes the point estimates $\hat{\gamma}$ and $\hat{\beta}_2$, their estimated standard errors (SE) and the p-values for testing these parameters equal to zero. The first three columns of numbers pertain to the models for the observation process and the last three to the joint damage processes $N_{i1}(t)$ and $N_{i2}(t)$. For comparison, we also performed univariate analyses and included the results in Table 3.3. The univariate results involve separate modeling of covariate effects on the rate functions for radiologically damaged joint counts and functionally damaged joint counts.

The results based on the joint analysis suggest that all three covariates had significant effects on the rates of both radiological and functional damage in joints. Specifically, the patients with a family history of psoriasis seem to have lower rates of damage, suggesting that the skin component of the disease is more active in these patients. Both a longer history of psoriatic arthritis and the higher number of active joints at clinic entry imply an increased damage rate. The results also indicate that all three covariates seem to have no effects on the observation process.

The univariate analyses gave similar conclusions for most of the covariates effects except that they significantly underestimated the effect of arthritis duration on the

rate of radiologically damaged joints. It can be seen that the estimated effects of all covariates based on the joint analysis are intermediate between those based on the two univariate analyses. Also the estimated covariate effects from the two univariate analyses are reasonably close, which suggests that the assumption of the same covariate effects on the two types of joint damage seems reasonable.

Figure 3.3 presents the estimates given in (3.8) of the baseline cumulative mean functions for the numbers of joints damaged according to the radiological and functional criteria. The figure includes the estimates obtained based on both joint and univariate analyses. The close agreement between the mean functions from the univariate and bivariate models again reflects the plausibility of the assumption of common regression coefficients. It is clear that one can expect a greater number of joints to be classified as damaged by the radiological criteria than by the functional criteria. This is consistent with the notion that damage tends to be detected earlier by radiological assessment compared to functional assessment (Siannis *et al.*, 2006). Also the joint analysis suggested a larger difference between the two criteria than the univariate analysis.

3.6 Concluding Remarks

Multivariate panel count data often arise in periodic follow-up studies that concern recurrent events. In the preceding sections, some marginal mean models were presented for regression analysis, and estimating equation approaches were proposed for inference about regression parameters. One main advantage of the proposed methodology is that

it leaves the relationship of different types of recurrent events completely unspecified. Also compared to the parametric approach proposed in Chen *et al.* (2005), it does not rely on the Poisson process and piecewise constant baseline intensity assumptions and can be relatively easily implemented. The application of the proposed methodology to the psoriatic arthritis data suggests that all three covariates, a family history of psoriasis, arthritis duration, and the number of active joints at clinic entry, had significant effects on the rate of the damaged joints.

When facing the analysis of multi-type data it is often natural to assess whether joint or multivariate analyses are warranted over separate univariate analyses. This is particularly true for models which do not provide parametric estimates of association parameters. Joint methods are helpful when interest lies in the global assessment of covariate effects across two or more processes. More discussion on this aspect can be found in Wei *et al.* (1989) for multivariate failure time data and Chen *et al.* (2005) for the current setting. Multivariate methods are also helpful when interest lies in obtaining common estimates of covariate effects. In general, separate univariate analyses will give different estimates, but when it appears reasonable and the estimates are comparable, constraining the regression coefficients to be common can simplify discussion. Also under the assumption of common effects, as discussed before, the use of joint analysis will often produce more efficient estimates of the effects. That is, the common effects can be estimated more precisely based on the joint analysis than based on either of the separate analysis. In contrast, the separate analysis cannot give a direct estimation of the common effects.

The method presented in the previous sections can be generalized in several directions. One is that models (3.1) and (3.2) assume that the covariates that affect different types of recurrent events are the same. However, in practice, there may exist type-specific covariates for each particular type of recurrent events and for this, methods similar to those given above can be developed. Note that models (3.1) and (3.2) also assume that the covariate effects on different types of recurrent events are identical. Sometimes this may not be true. In this case, one can redefine a larger, type-specific covariate vector that corresponds to a new and larger vector of covariate effects that includes all different covariate effects but is the same for different types of recurrent events.

Chapter 4

SEMIPARAMETRIC ANALYSIS OF PANEL COUNT DATA WITH CORRELATED OBSERVATION AND FOLLOW-UP TIMES

4.1 Introduction

In Chapter 2, we assumed that the follow-up or censoring time is independent of the observation and response processes given covariates, but this may be violated in practice. For example, the follow-up times may be times to some terminal events related to the recurrent event of interest. Some recent references that discuss this for recurrent event data include Huang and Wang (2004), Liu *et al.* (2004) and Ye *et al.* (2007). Wang *et al.* (2001) also considered the same phenomenon and described an study of AIDS patients in which the recurrent and terminal events are hospitalization and death, respectively. Note that for recurrent event data, only the underlying counting process and the follow-up process are involved. Also one could face the same correlated

problem in general longitudinal studies and for this, an extensive literature has been developed (De Gruttola and Tu, 1994; Wulfsohn and Tsiatis, 1997; Roy and Lin, 2002; Song *et al.*, 2002).

For panel count data, as mentioned above, one could have to deal with three related processes and the same could be true for longitudinal studies. For example, Huang *et al.* (2006) considered the bladder cancer study discussed in Section 1.1.2.3 where the response process and the observation process are related. In the AIDS study discussed in Wang *et al.* (2001), suppose that one is interested in some symptoms related to AIDS such as CD4 counts or the time at which the patient's CD4 counts cross some threshold. Then the response process may be correlated with the observation process as well as the follow-up process. Lipsitz *et al.* (2002) presented a set of longitudinal data from a study of children with acute lymphoblastic leukemia which involves correlated response and observation processes. In this chapter, we consider situations where all three processes may be correlated.

The remainder of this chapter is organized as follows. Section 4.2 introduces notation and describes joint models for the three processes. To characterize the correlation, we employ some shared frailty models, a commonly used approach in both survival and longitudinal data analyses when a joint analysis is required. In Section 4.3, we consider estimation of regression parameters and for this, the estimating equation approach is applied. To implement the approach, a three-step estimation procedure is developed, and the proposed estimates of regression parameters are consistent and have asymptotically a normal distribution. Section 4.4 presents some results obtained from

a simulation study for assessing the proposed inference approach, and we will revisit the bladder cancer study in Section 4.5. Some concluding remarks are given in Section 4.6.

4.2 Models and Notation

Consider a recurrent event study that consists of n independent subjects and let $N_i(t)$ denote the number of occurrences of the recurrent event of interest before or at time t for subject i. Suppose that for each subject, there exists a vector of covariates denoted by x_i and given x_i and two latent variables u_i and v_i, the mean function of $N_i(t)$ has the form

$$E\{N_i(t)|x_i, u_i, v_i\} = \mu_N(t) \exp(x_i'\beta_1 + u_i\beta_2 + v_i\beta_3)\,. \tag{4.1}$$

Here $\mu_N(t)$ is a completely unknown continuous baseline mean function, and β_1, β_2 and β_3 are unknown regression parameters.

For subject i, suppose that $N_i(\cdot)$ is observed only at finite time points $T_{i1} < ... < T_{iK_i}$, where K_i denotes the potential number of observation times, $i = 1, ..., n$. That is, only the values of $N_i(t)$ at these observation times are known and we have panel count data on the $N_i(t)$'s. Also for subject i, suppose that there exist two follow-up times C_i^* and τ_i, where C_i^* may be related to $N_i(t)$ and the T_{il}'s, and τ_i is independent of them. Assume that one only observes $C_i = \min(C_i^*, \tau_i)$ and $\delta_i = I(C_i = C_i^*)$ and thus $N_i(t)$ is observed only at these T_{il}'s with $T_{il} \leq C_i$, $i = 1, ..., n$. Define

$\tilde{N}_i(t) = H_i\{\min(t, C_i)\}$, where $H_i(t) = \sum_{l=1}^{K_i} I(T_{il} \leq t)$, $i = 1, ..., n$. Then $\tilde{N}_i(t)$ is a point process characterizing the ith subject's observation process and jumps only at the observation times.

In the following, we assume that given (x_i, u_i), $H_i(\cdot)$ is a non-homogeneous Poisson process with the intensity function

$$\lambda_h(t) = \lambda_{0h}(t) \exp(x_i'\alpha_1 + u_i)\,. \tag{4.2}$$

In the model above, $\lambda_{0h}(t)$ is a completely unknown continuous baseline intensity function and α_1 denotes the vector of regression parameters. For the follow-up time C_i^*, it will be assumed that its hazard function is given by

$$\lambda_c(t) = \lambda_{0c}(t) \exp(x_i'\gamma_1 + u_i\gamma_2 + v_i) \tag{4.3}$$

given x_i, u_i and v_i, where $\lambda_{0c}(t)$ denotes an unknown baseline hazard function, and γ_1 and γ_2 are regression parameters.

There exists a great deal of research on each of the three models (4.1) - (4.3) and their special cases individually. For example, model (4.3) without the latent variables is the well-known proportional hazards model (Kalbfleisch and Prentice, 2002) and a number of methods have been developed for the same model with $\gamma_2 = 0$. Wang *et al.* (2001) and Huang and Wang (2004) considered a model similar to model (4.2) for recurrent event data. There also exists some limited work on the joint analysis of two of these models (Cheng and Wei, 2000). In the following, we study the joint analysis

of all three models together with the focus on estimation of regression parameters β_1 along with α_1 and γ_1. Let $\Lambda_{0h}(t) = \int_0^t \lambda_{0h}(s)ds$. We will assume that $\Lambda_{0h}(\tau) = 1$ for identifiability and $E(u_i|x_i) = E(u_i)$, where τ denotes the length of study. Also it will be assumed that $v_i \sim N(0, \sigma^2)$, where σ^2 is an unknown parameter.

4.3 Estimation of Regression Parameters

In this section, we consider estimation of β_1 along with other parameters. For this, note that if the latent effects u_i's and v_i's are known, then model (4.1) becomes the usual proportional means model and several methods such as that given in Cheng and Wei (2000) can be used. Unfortunately they are not known in practice. To deal with this, we borrow the idea used in Huang and Wang (2004) to first estimate or predict these unknown latent variables. For $i = 1, ..., n$, let $x'_{1i} = (x'_i, u_i)$, $x'_{2i} = (x'_i, u_i, v_i)$, $\beta' = (\beta'_1, \beta_2, \beta_3)$, $\alpha' = (\alpha'_1, 1, 0)$, and $\gamma' = (\gamma'_1, \gamma_2)$. The proposed estimation procedure consists of the following three steps.

4.3.1 Estimation of Model (4.2)

To estimate β_1, we first consider inference about model (4.2), for which we have recurrent event data. Let $K_i^* = \tilde{N}_i(C_i)$, the total number of observations on subject i, $i = 1, ..., n$. Also let the s_j's denote the ordered and distinct time points of all the observation times $\{T_{il}\}$, d_j the number of the observation times equal to s_j, and n_j the number of the observation times satisfying $T_{il} \leq s_j \leq C_i$ among all subjects. Define $x'_{3i} = (x'_i, 1)$, $\alpha'_* = (\alpha'_1, \alpha_2) = (\alpha'_1, E(u_i))$. Then following Huang and Wang (2004),

one can first estimate $\Lambda_{0h}(t)$ and α_* by

$$\widehat{\Lambda}_{0h}(t) \,=\, \prod_{s_l>t} \left(\, 1 \,-\, \frac{d_l}{n_l} \,\right)$$

and the estimating equation

$$\sum_{i=1}^{n} w_i \, x_{3i} \, \left\{ \, K_i^* \, \widehat{\Lambda}_{0h}^{-1}(C_i) \,-\, \exp(\alpha_*' x_{3i}) \, \right\} \,=\, 0 \,, \tag{4.4}$$

respectively. In equation (4.4), the w_i's are some weights that could depend on x_i, C_i and Λ_{0h}. A key fact used in deriving the above estimating euqation is that conditional on (x_i, C_i, u_i, K_i^*), the observation times $\{T_{i1}, ..., T_{iK_i^*}\}$ are the order statistics of a simple random sample of size K_i^* from the density function

$$\frac{\lambda_{0h}(t) \exp(\alpha_1' x_i + u_i)}{\Lambda_{0h}(C_i) \, \exp(\alpha_1' x_i + u_i)} \, I(0 \le t \le C_i) \,=\, \frac{\lambda_{0h}(t)}{\Lambda_{0h}(C_i)} \, I(0 \le t \le C_i) \,.$$

Let $\hat{\alpha}_*' \,=\, (\hat{\alpha}_1', \hat{\alpha}_2)$ denote the estimate of α_*' given by equation (4.4). Note that given (x_i, C_i, u_i), the expected value of K_i^* is equal to $\Lambda_{0h}(C_i) \, \exp(\alpha_1' x_i + u_i)$. Thus it is natural to predict u_i by

$$e^{\hat{u}_i} \,=\, \frac{K_i^*}{\hat{\Lambda}_{0h}(C_i) \, e^{\hat{\alpha}_1' x_i}} \,. \tag{4.5}$$

4.3.2 Estimation of Model (4.3)

In this subsection, we discuss estimation of model (4.3). For this, let $O \,=\, (O_1, ..., O_n)$, where $O_i' \,=\, (C_i, \delta_i, x_i', u_i)$ denotes the observed data on subject i assuming that u_i is

known. Also let $c_1 < \cdots < c_k$ denote the ordered observed failure times and assume that we can write $\Lambda_{0c}(t)$ as

$$\Lambda_{0c}(t) \;=\; \sum_{j=1}^{k} a_j \, I(t \geq c_j)\,,$$

where $a' \;=\; (a_1, ..., a_k)$ is a vector of unknown parameters. Define $\theta \;=\; (a', \gamma', \sigma^2)'$. Then the full likelihood function has the form

$$L(\theta) \;=\; \prod_{i=1}^{n} \{\lambda_{0c}(C_i) \exp(x_{1i}'\gamma + v_i)\}^{\delta_i} \exp\{-\Lambda_{0c}(C_i) \exp(x_{1i}'\gamma + v_i)\}\, \phi(v_i; \sigma)$$

based on the pseudo complete data O and the v_i's, where $\phi(\cdot; \sigma)$ denotes the density function of $N(0, \sigma^2)$.

To maximize $L(\theta)$ with respect to θ, we propose to replace u_i in $L(\theta)$ by its prediction given in (4.5) and then use the EM algorithm to deal with the latent variables v_i's as usual. To implement the EM algorithm, we first consider the E-step, which computes the conditional expectation of the log likelihood function given the current estimate of θ and the observed data O. To this end, note that the log likelihood function can be written as

$$l(\theta) \;=\; \sum_{i=1}^{n} \Big\{ \delta_i \Big[\log\{\lambda_{0c}(C_i)\} + x_{1i}'\gamma + v_i \Big] \;-\; \Lambda_{0c}(C_i) \exp(x_{1i}'\gamma + v_i) + \log \phi(v_i; \sigma) \Big\}$$

$$= \sum_{i=1}^{n} \delta_i \Big[\log\{\lambda_{0c}(C_i)\} + x_{1i}'\gamma \Big] \;+\; \sum_{i=1}^{n} g(v_i; \theta)\,,$$

where

$$g(v_i;\theta) \;=\; \delta_i v_i - \Lambda_{0c}(C_i)\exp(x_{1i}'\gamma + v_i) + \log\phi(v_i;\sigma)\,.$$

To calculate $E\{l(\theta)|O,\theta^{(m)}\}$, one needs to calculate

$$E_i\{g(v_i;\theta)|O_i,\theta^{(m)}\} \;=\; \int g(v_i;\theta)\, f(v_i|O_i,\theta^{(m)})\, dv_i\,,$$

where $\theta^{(m)}$ denotes the current estimate of θ and

$$f(v_i|O_i,\theta) \;=\; \frac{\exp(\delta_i v_i)\exp\{-\Lambda_{0c}(C_i)\exp(x_{1i}'\gamma + v_i)\}\phi(v_i;\sigma)}{\displaystyle\int \exp(\delta_i v_i)\exp\{-\Lambda_{0c}(C_i)\exp(x_{1i}'\gamma + v_i)\}\phi(v_i;\sigma)dv_i}\,,$$

the conditional density of v_i given O_i and θ. It is apparent that this integration has no closed form. For this, with $\theta \;=\; \theta^{(m)}$, let $\{\, v_i^{(l)}; i = 1,...,n, l = 1,...,L\,\}$ be L i.i.d. samples from $N(0,\{\sigma^{(m)}\}^2)$. Then one can approximate $E_i\{g(v_i;\theta)|O_i,\theta^{(m)}\}$ by

$$\hat{E}\{g(v_i;\theta)|O_i,\theta^{(m)}\} \;=\; \frac{\sum_{l=1}^{L} g(v_i^{(l)};\theta)\exp(\delta_i v_i^{(l)})\exp\{-\Lambda_{0c}^{(m)}(C_i)\exp(x_{1i}'\gamma^{(m)} + v_i^{(l)})\}}{\sum_{l=1}^{L}\exp(\delta_i v_i^{(l)})\exp\{-\Lambda_{0c}^{(m)}(C_i)\exp(x_{1i}'\gamma^{(m)} + v_i^{(l)})\}}\,. \tag{4.6}$$

Now we consider the M-step of the EM algorithm, which maximizes $E\{l(\theta)|O,\theta^{(m)}\}$ with respect to θ. For this, by taking its derivatives with respect to θ and setting the derivatives equal to zero, we obtain the following equations

$$a_j^{(m+1)} = \Big[\sum_{i=1}^{n} E_i\left\{\exp(x_{1i}'\gamma + v_i)I(C_i \geq c_j)\right\}\Big]^{-1}, \tag{4.7}$$

for $j = 1, .., k$, $\sigma^{(m+1)} = \{n^{-1} \sum_{i=1}^{n} E_i(v_i^2)\}^{1/2}$, and

$$\sum_{i=1}^{n} E_i\Big[x_{1i}\left\{\delta_i - \Lambda_{0c}(C_i)\exp(x_{1i}'\gamma + v_i)\right\}\Big] = 0 \tag{4.8}$$

for the updated estimate $\theta^{(m+1)}$ of θ. In practice, we propose to obtain the $a_j^{(m+1)}$ and thus $\Lambda_{0c}^{(m+1)}$ first by using (4.7) with letting $\theta = \theta^{(m)}$. Then by replacing Λ_{0c} with $\Lambda_{0c}^{(m+1)}$, one can solve (4.8) to get $\gamma^{(m+1)}$ and $\{\sigma^{(m+1)}\}^2$. Finally, given the estimate $\hat{\theta}$ of θ, one could calculate the conditional expectation of v_i given O_i as or predict v_i by

$$\hat{v}_i = E_i(v_i|O_i, \hat{\theta}), \tag{4.9}$$

which can be approximated by (4.6).

4.3.3 Estimation of Model (4.1)

Now we are ready to estimate β_1 or β in model (4.1). For this, define $Y_i(t) = I(t \leq C_i)$ and

$$S_j(\beta; t) = \frac{1}{n}\sum_{i=1}^{n} Y_i(t)\,\exp\{x_{2i}'(\beta + \alpha)\}\, x_{2i}^{\otimes j},$$

for $j = 0, 1, 2$, where $a^{\otimes 0} = 1$, $a^{\otimes 1} = a$ and $a^{\otimes 2} = a\,a'$ for a vector a. Note that if all the u_i's and v_i's are known, following Cheng and Wei (2000), one can estimate β using the estimating function

$$U(\beta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^{\tau}\left\{x_{2i} - \frac{S_1(\beta; t)}{S_0(\beta; t)}\right\} N_i(t)\, d\,\tilde{N}_i(t).$$

Motivated by this, we propose to estimate β based on the following estimating function

$$\hat{U}(\beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^\tau \left\{ \hat{x}_{2i} - \frac{\hat{S}_1(\beta;t)}{\hat{S}_0(\beta;t)} \right\} N_i(t) d\,\tilde{N}_i(t)\,. \tag{4.10}$$

Here $\hat{x}_{2i} = (x_i', \hat{u}_i, \hat{v}_i)'$ with the $\hat{u}_i$ and $\hat{v}_i$ given by (4.5) and (4.9), respectively, and $\hat{S}_j(\beta;t)$ denotes $S_j(\beta;t)$ with the x_{2i}'s and α replaced by the $\hat{x}_{2i}$'s and $\hat{\alpha} = (\hat{\alpha}_1', 1, 0)'$, respectively.

Define the estimate $\hat{\beta}$ of β as the solution to $\hat{U}(\beta) = 0$. Note that $\hat{x}_{2i}$ is independent of β and we have

$$\frac{\partial \hat{U}(\beta)}{\partial \beta} = -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^\tau \left\{ \frac{\hat{S}_2(\beta;t)\hat{S}_0(\beta;t) - \hat{S}_1(\beta;t)\hat{S}_1(\beta;t)'}{\hat{S}_0^2(\beta;t)} \right\} N_i(t)\, d\tilde{N}_i(t)\,,$$

which is strictly negative. Thus $\hat{\beta}$ is unique. Also it is expected that $\hat{\beta}$ is consistent and its distribution can be approximated by a normal distribution when n is large.

For estimation of the covariance matrix of $\hat{\beta}_1$ or $\hat{\beta}$, we propose to use the following simple bootstrap procedure. Let B denote a prespecified positive integer. For each b, where $1 \le b \le B$, draw a simple random sample of size n,

$$D^{(b)} = \left\{ T_{i1}^{(b)}, \ldots, T_{iK_i^*}^{(b)}, N_i^{(b)}(T_{i1}^{(b)}), \ldots, N_i^{(b)}(T_{iK_i^*}^{(b)}), C_i^{(b)}, \delta_i^{(b)}, x_i^{(b)\prime};\ i = 1, \ldots, n \right\},$$

with replacement from the observed data

$$D = \left\{ T_{i1}, \ldots, T_{iK_i^*}, N_i(T_{i1}), \ldots, N_i(T_{iK_i^*}), C_i, \delta_i, x_i';\ i = 1, \ldots, n \right\}.$$

Let $\hat{\beta}^{(b)}$ be the proposed estimate of β based on the data set $D^{(b)}$ defined above. Then a natural estimate of the covariance matrix of $\hat{\beta}$ is given by

$$\hat{\Sigma} = \frac{1}{B-1}\sum_{b=1}^{B}\left\{\hat{\beta}^{(b)} - \frac{1}{B}\sum_{b=1}^{B}\hat{\beta}^{(b)}\right\}^2 .$$

4.4 Numerical Results

We conducted a simulation study to assess the performance of the estimation procedure proposed in the previous sections under different situations with the focus on estimation of β_1. In the study, the covariate x_i's were assumed to follow a Bernoulli distribution with success probability 0.5. To generate the simulated data, we first generated the $u_i^* = \exp(u_i)$ and v_i from the gamma distribution with mean 10 and variance 50 and the normal distribution with mean 0 and variance $\sigma^2 = 0.25$, respectively. Given x_i, u_i and v_i, C_i^* was generated under model (4.3) with $\lambda_{0c}(t) = 0.002$ and it was assumed that $\tau = 18$.

For the observation process, we assumed that H_i follows the homogeneous Poisson process with $\lambda_{0h}(t) = \tau^{-1}$. Then given x_i and u_i, K_i^*, the number of real observation times for subject i, follows the Poisson distribution with mean

$$\Lambda_h(C_i|x_i,u_i) = \Lambda_{0h}(C_i)\exp(x_i\alpha_1 + u_i) = \frac{C_i\exp(x_i\alpha_1+u_i)}{\tau},$$

$i = 1, 2, \ldots, n$. Furthermore, the observation times $(T_{i1}, ..., T_{iK_i^*})$ are the order statistics of a random sample of size K_i^* from the uniform distribution over $(0, C_i)$. Given K_i^*

and $(T_{i1}, \ldots, T_{iK_i^*})$, we generated $N_i(T_{ij})$ using the formula

$$N_i(T_{ij}) \;=\; N_i^*[\lambda_N(T_{i1})] + N_i^*[\lambda_N(T_{i2}) - \lambda_N(T_{i1})] + \ldots + N_i^*[\lambda_N(T_{ij}) - \lambda_N(T_{i\,j-1})]$$

for $j \;=\; 1, ..., K_i^*$ and $i \;=\; 1, .., n$. Here $N_i^*[\lambda_N(t)]$ denotes the random number generated from the Poisson distribution with mean

$$t\, \exp(x_i\beta_1 + u_i\beta_2 + v_i\beta_3)\,.$$

The results given below are based on $n = 100$ or 200, $L = 50$ in equation (4.6), $B = 20$ in the bootstrap procedure for variance estimation, and 1000 replications.

Table 4.1 presents the simulation results on estimation of β_1 for the situations where $\beta_1 = -2, -1, 0, 1, 2$ along with $\beta_2 = \beta_3 = 0$ and $\alpha_1 = \gamma_1 = \gamma_2 = 1$. The table includes the averages of proposed estimates of β_1 based on the simulated data, the sample standard deviations of the estimates (SSD), the means of the bootstrap standard deviation estimates (BSD), and the empirical 95% coverage probabilities (CP) for β_1. Table 4.2 gives the estimation results for the same situations as in Table 1 except that $\beta_2 = \beta_3 = 0.2$. These results indicate that the estimate $\hat{\beta}_1$ seems to be unbiased and the bootstrap variance estimation procedure provides reasonable estimates. Also the results on the empirical coverage probabilities indicate that the normal approximation seems to be appropriate.

To assess the performance of the normal approximation to the finite-sample distribution of the estimate of β_1, we studied the quantile plots of the standardized estimates

of β_1. Figures 4.1 and 4.2 display such plots corresponding to the situations where $n = 100$, $\beta_1 = 0$ or 1 with $\beta_2 = \beta_3 = 0$. These figures indicate that the normal approximation seems reasonable. Similar plots were obtained for other setups.

4.5 An Application

In this section, we illustrate the proposed methodology by revisiting the data set from the bladder cancer study discussed in Chapters 1 and 2. Following Sun and Wei (2000), we restrict our attention to the patients in the placebo (47) and thiotepa (38) groups.

For the analysis, define the first component of x_i to be equal to 1 if the ith patient was given the thiotepa treatment and 0 otherwise. Also define the second and third components of x_i to be the number of initial tumors and the size of the largest initial tumor of the patient, respectively. Assume that the occurrence process of the bladder tumors, the clinical visit process and the follow-up process can be described by models (4.1), (4.2) and (4.3), respectively. The application of the estimation procedure proposed in the previous sections gave $\hat{\beta}_1 = (-1.8483, 0.1996, 0.0015)'$ with the estimated standard errors of $(0.6879, 0.3181, 0.3562)'$. These results suggest that the thiotepa treatment significantly reduced the occurrence rate of the bladder tumors. However, the occurrence rate of the bladder tumors does not seem to be significantly related with the number of initial tumors and the size of the largest initial tumor.

For comparison, we noticed that Sun and Wei (2000) assumed that the three processes involved are independent of each other given the covariates and estimated

the effects of three covariates as $(-2.0249, 0.6620, -0.1229)'$ with the estimated standard errors being $(0.4500, 0.2133, 0.2035)'$. It can be seen that the results from the two methods are similar but the approach that took into account the possible correlation among the three processes gave smaller estimated effects. In other words, without taking into account the correlation, one could overestimate the treatment or covariate effects. One possible reason for this is that part of the estimated effect given by the approach assuming the independence may be due to the correlation of the three processes. Huang *et al.* (2006) studied a similar problem and gave a similar conclusion.

4.6 Discussion

In this chapter, we considered regression analysis of panel count data when all three processes involved may be related and for the purpose, some shared frailty models were proposed. For inference, an estimating equation approach and an EM algorithm were developed for estimation of regression parameters representing covariate effects. A key advantage of the proposed approach over existing methods for panel count data is that it allows both the observation process and the follow-up process to be related with the response process of interest. In general, it may be hard to have enough evidence to verify or assess the independence assumption or the existence of the correlation. But as mentioned before, this could happen quite often in practice.

In the preceding sections, our focus has been on estimation of regression parameter β_1. Sometimes, one may be interested in estimating the baseline mean function $\mu_N(t)$ in model (4.1). For this, one could develop an estimate following the one given in Thall

and Lachin (1988) or others by treating u_i and v_i to be known and replaced with $\hat{u}_i$ and $\hat{v}_i$ given in (4.5) and (4.9), respectively.

Chapter 5

FUTURE RESEARCH

In this chapter, we discuss several potential directions for future research that are related to the analysis of panel count data.

5.1 More Efficient Estimation for Regression Parameters

In this dissertation, estimating equation-based approaches were used to estimate regression parameters in semiparametric models. For example, the estimating equations (2.4) and (2.5) involve weight functions w_{2i} and w_{1i}, respectively, which could depend on x_i, C_i and Λ_0 in Chapter 2. In Chapter 4, the weight function w_i in equation (4.4) could depend on x_i, C_i and Λ_{0h}. As future research, it would be useful to investigate how to choose these weight functions to obtain efficient or optimal estimates. Similarly, one could apply some weight functions to the estimating functions defined in (3.3), (3.4), (3.5) and the equation (4.10) and ask the same question.

Another approach for more efficient estimates is to develop new estimating equations

for the problems considered in the previous chapters. For example, in the construction of the estimating equation (2.5), we used $\bar{N}_i = \bar{N}_i(\tau) = \int_0^\tau N_i(t)\, d\tilde{N}_i(t)$. Although this equation is natural and intuitive, a more efficient estimating equation may be derived directly based on the two processes $N_i(t)$ and $\tilde{N}_i(t)$ at all the observation times (Cheng and Wei, 2000). A similar idea was applied in Chapter 4.

5.2 Regression Analysis of Multivariate Panel Count Data with Time-Dependent Covariates

In Chapter 3, we considered regression analysis of multivariate panel count data with time-invariant covariates. However, the values taken by the covariates may vary over time in many longitudinal studies (e.g, age since clinic entry, blood pressure, and current smoking status). It is useful to study how to generalize the proposed methods to the situation where the covariates of interest may be time-dependent. Some new inference procedures would be needed.

5.3 Likelihood-Based Approach to the Analysis of Panel Count Data with Dependent Observation and Follow-up Times

As discussed in Chapters 2 and 4, both marginal and random effect approaches can be used in the analysis of panel count data. However, there seems to have few likelihood-based methods for panel count data. Recently, Zeng and Lin (2007) proposed a class of semiparametric transformation models with random effects for recurrent

event data by accounting for the dependence of the recurrent event times within the same subject and developed nonparametric maximum likelihood estimators (NPMLEs) for regression parameters. A generalization of their method to the analysis of panel count data for the situation, where the censoring mechanism is dependent of both the response and observation processes, will be an interesting and possible direction for future research.

APPENDIX

Appendix A:
Proof of the Asymptotic Properties of $\hat{\beta}_1$ and $\hat{\beta}_2$ in Chapter 2

In this appendix, we prove the consistency and normality of $\hat{\beta}_1$ and $\hat{\beta}_2$. First we consider $\hat{\beta}_1$. Let N_i, $\tilde{N}_i$, T_{ij}, x_i, x_{1i}, x_{2i}, z_i, K_i^*, w_{1i}, w_{2i}, C_i, $\hat{U}_1$ and $\hat{U}_2$ be defined as in Sections 2.2 and 2.3, and X, X_1, X_2, Z, K^*, W_1, W_2 and C denote the underlying random variables of the x_i's, x_{1i}'s, x_{2i}'s, z_i's, K_i^*'s, w_{1i}'s, w_{2i}'s and C_i's, respectively. For the proof, we need the following regular conditions that are similar to those given in Huang and Wang (2004):

(a) $P(C \geq \tau, Z > 0) > 0$;

(b) X is uniformly bounded;

(c) the variance of Z is bounded and there exists a positive small constant $\epsilon > 0$ such that $Z > \epsilon$ almost surely;

(d) $G(u) = E\{ZI(C \geq u)\}$ is continuous for $u \in [0, \tau]$.

Let V_1 and V_2 denote the joint distributions of $(C\,,\,Z)$ and $(W_2\,,\,X_2\,,\,K^*\,,\,C)$, re-

spectively. Define $Q(u) = \int_0^u G(v)\, d\Lambda_0(v)$, $R(u) = G(u)\Lambda_0(u)$,

$$b_i(t) \;=\; \sum_{j=1}^{K_i^*} \Big\{ \int_t^\tau \frac{I(T_{ij} \le u \le C_i)\, dQ(u)}{R^2(u)} - \frac{I(t \le T_{ij} \le \tau)}{R(T_{ij})} \Big\},$$

$$f_{i2} \;=\; \int \frac{w_2 x_2 k^* b_i(c)}{\Lambda_0(c)} dV_2(w_2, x_2, k^*, c) + w_{2i} x_{2i} \left\{ K_i^* \Lambda_0^{-1}(C_i) - \exp(\gamma_2' x_{2i}) \right\},$$

and f_i to be the vector function $E\{-\partial f_{i2}/\partial \gamma_2\}^{-1} f_{i2}$ without the last entry. Then we have $G(u) = E\{ZI(C \ge u)\} = \int zI(c \ge u)\, dV_1(c, z)$,

$$\widehat{\gamma} - \gamma \;=\; \frac{1}{n} \sum_{i=1}^{n} f_i + o_p(n^{-1/2}) \qquad (A.1)$$

and

$$\widehat{\Lambda}_0(t) - \Lambda_0(t) \;=\; \frac{1}{n} \Lambda_0(t) \sum_{i=1}^{n} b_i(t) + o_p(n^{-1/2}), \quad t \le \tau \qquad (A.2)$$

(Wang *et al.*, 2001; Huang and Wang, 2004).

For the consistency of $\widehat{\beta}_1$, let

$$m_i = E\{z_i | K_i^*, x_i, C_i\} = \frac{K_i^*}{\Lambda_0(C_i) \exp(\gamma' x_i)}$$

and β_1^* denote the solution to the equation

$$U_1^*(\beta_1) = \frac{1}{n} \sum_{i=1}^{n} w_{1i} x_{3i} \left\{ \bar{N}_i - \exp(\gamma' x_i) \exp(\beta_1' x_{3i}) \right\} = 0\,,$$

where $x_{3i} = (x_i', \log(m_i), 1)'$. Let X_3 denote the underlying random variable of the

x_{3i}'s. Then it can be easily shown that β_1^* is a consistent estimate of the true value β_1 if Λ_0 and γ are known. Thus for the consistency of $\widehat{\beta}_1$, it is sufficient to show that $\widehat{\beta}_1 - \beta_1^*$ converges to zero in probability. To this end, define

$$\begin{aligned} A_n(b) = & \frac{1}{n}\sum_{i=1}^{n} w_{1i}\Big[\bar{N}_i(b-\beta_1^*)'\widehat{x}_{1i} + \exp(\widehat{\gamma}'x_i)\{\exp(\beta_1^{*\prime}\widehat{x}_{1i}) - \exp(b'\widehat{x}_{1i})\}\Big], \\ A(b) = & \frac{1}{n}\sum_{i=1}^{n} w_{1i}\Big[\bar{N}_i(b-\beta_1^*)'x_{3i} + \exp(\gamma'x_i)\{\exp(\beta_1^{*\prime}x_{3i}) - \exp(b'x_{3i})\}\Big]. \end{aligned}$$

It is easy to see that $\hat{U}_1(b)$ and $U_1^*(b)$ are derivatives of $A_n(b)$ and $A(b)$ and $\widehat{\beta}_1$ and β_1^* are the unique maximums of A_n and A, respectively. Thus the consistency follows from the facts that $\widehat{\gamma}$ and $\widehat{\Lambda}_0(C_i)$ converge with $\sqrt{n}$-convergence rate (Wang *et al.*, 2001) and both $A_n(b)$ and $A(b)$ are concave functions.

For the asymptotic normality of $\hat{\beta}_1$, note that using some algebra, we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} w_{1i}\hat{x}_{1i}\left\{\bar{N}_i - \exp(\hat{\gamma}'x_i)\exp(\beta_1'\hat{x}_{1i})\right\} = H_{1n} - H_{2n} - H_{3n} + H_{4n} + o_p(1),$$

where

$$H_{1n} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_{1i}(\beta_1) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} w_{1i}x_{3i}\left\{\bar{N}_i - \exp(\gamma'x_i)\exp(\beta_1'x_{3i})\right\},$$

$$H_{2n} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} w_{1i}x_{3i}\exp(\beta_1'x_{3i})\left\{\exp(\hat{\gamma}'x_i) - \exp(\gamma'x_i)\right\},$$

$$H_{3n} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} w_{1i}x_{3i}\exp(\gamma'x_i)\left\{\exp(\beta_1'\hat{x}_{1i}) - \exp(\beta_1'x_{3i})\right\},$$

and

$$H_{4n} \;=\; \frac{1}{\sqrt{n}} \sum_{i=1}^{n} w_{1i}(\hat{x}_{1i} - x_{3i}) \left\{ \bar{N}_i - \exp(\gamma' x_i) \exp(\beta_1' x_{3i}) \right\} .$$

Define $\mathbf{e}_{p+2}$ be a $(p+2)$-dimensional vector whose elements are all zero except that the $(p+1)$th element is equal to 1. Then it can be easily shown that

$$\begin{aligned} H_{2n} \;&=\; \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_{2i}(\beta_1) \;+\; o_p(1) \\ &=\; \frac{1}{\sqrt{n}} \sum_{i=1}^{n} E_{W_1,X_3}\{W_1 X_3 \exp(\beta_1' X_3 + \gamma' X) X'\} f_i \;+\; o_p(1)\,, \end{aligned}$$

$$\begin{aligned} H_{3n} \;&=\; \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_{3i}(\beta_1) \;+\; o_p(1) \\ &=\; -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} E_{W_1,X_3,C}\Big[W_1 X_3 \exp(\beta_1' X_3 + \gamma' X) \beta_1' \mathbf{e}_{p+2} \left\{ b_i(C) \;+\; X' f_i \right\} \Big] \;+\; o_p(1)\,, \end{aligned}$$

$$\begin{aligned} H_{4n} \;&=\; \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_{4i}(\beta_1) \;+\; o_p(1) \\ &=\; -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{e}_{p+2} E_{W_1,X_3,\bar{N},C}\Big[W_1 \left\{ \bar{N} - \exp(\gamma' X + \beta_1' X_3) \right\} \left\{ b_i(C) + X' f_i \right\} \Big] \;+\; o_p(1)\,, \end{aligned}$$

where E_{W_1,X_3}, $E_{W_1,X_3,C}$ and $E_{W_1,X_3,\bar{N},C}$ denote the expectations with respect to the joint distribution of (W_1, X_3), (W_1, X_3, C) and $(W_1, X_3, \bar{N}, C)$, respectively.

Let $g_i(\beta_1) = \sum_{j=1}^{4} g_{ji}(\beta_1)$. It follows from the above results that

$$\hat{U}_1(\beta_1) \;=\; \frac{1}{n} \sum_{i=1}^{n} g_i(\beta_1) \;+\; o_p(n^{-1/2})\,.$$

This along with the Taylor series expansion yields that

$$\sqrt{n}\,(\hat{\beta}_1 \,-\, \beta_1) \,=\, \frac{1}{\sqrt{n}}\,\phi^{-1} \sum_{i=1}^{n} g_i(\beta_1) \,+\, o_p(1)\,,$$

where $\phi \,=\, E(-\partial g_i/\partial\beta_1) = E_{W_1,X_3}\{W_1 X_3 X_3' \exp(\gamma' X + \beta_1' X_3)\}$. Thus $\sqrt{n}\,(\hat{\beta}_1 - \beta_1)$ has an asymptotic normal distribution with mean zero and covariance matrix $\phi^{-1}\,\Sigma\,(\phi^{-1})'$, where $\Sigma \,=\, E(g_i\, g_i')$.

Now we study $\hat{\beta}_2$ in Chapter 2. The consistency follows the same arguments as those for the consistency of $\hat{\beta}_1$ and thus we will only prove the asymptotical normality. Define $\Lambda_0^*(t) \,=\, \int_0^t \lambda_0^*(s)\,ds$, the baseline cumulative hazard function of the C_i's, and let C_i^* be defined as in Section 2.3. Also define the counting process $N_i^*(t) \,=\, I(C_i^* \le t, C_i^* = C_i)$ and the martingale $M_i^*(t) \,=\, N_i^*(t) \,-\, \int_0^t I(C_i^* \ge s)\,\lambda_0^*(s)\,\exp(\psi' x_i) ds$, $i \,=\, 1, ..., n$. It then follows from Andersen *et al.* (1982) and Sun and Wei (2000) so that we have

$$\hat{\psi} \,-\, \psi \,=\, \frac{1}{n}\sum_{i=1}^{n} e_i + o_p(n^{-1/2}) \tag{A.3}$$

and

$$\hat{\Lambda}_0^*(t) \,-\, \Lambda_0^*(t) \,=\, \frac{1}{n}\,\sum_{i=1}^{n} h_i(t) \,+\, o_p(n^{-1/2})\,,\; t \in [0,\tau]\,, \tag{A.4}$$

where $h_i(t) \,=\, \int_0^t S_0(s)^{-1} dM_i^*(s)$ and

$$e_i \,=\, \left[\int_0^\tau \left\{\frac{a_2(s)}{a_0(s)} - \frac{a_1(s)a_1'(s)}{a_0^2(s)}\right\} dN^*(s)\right]^{-1} \int_0^\tau \left\{x_i - \frac{a_1(s)}{a_0(s)}\right\} dN_i^*(s)$$

with $a_0(s) = \sum_{i=1}^n I(C_i^* \geq s)\exp(\psi' x_i)$, $a_1(s) = \sum_{i=1}^n I(C_i^* \geq s)\exp(\psi' x_i)x_i$ and $a_2(s) = \sum_{i=1}^n I(C_i^* \geq s)\exp(\psi' x_i)x_i x_i'$.

For $i = 1, \cdots, n$, define

$$\bar{N}_i = \int_0^\tau \frac{N_i(t)d\tilde{N}_i(t)}{\{S_0(t)\}^{\exp(\psi' x_i)}}\ ,\ \hat{\bar{N}}_i = \int_0^\tau \frac{N_i(t)d\tilde{N}_i(t)}{\{\hat{S}_0(t)\}^{\exp(\hat{\psi}' x_i)}}$$

and

$$H_{5n}(\beta_2) = \frac{1}{\sqrt{n}}\sum_{i=1}^n w_{1i}\, x_{3i}\,(\hat{\bar{N}}_i - \bar{N}_i)\,.$$

Using Taylor series expansion and the properties (A.3) and (A.4), one can show that

$$H_{5n}(\beta_2) = \frac{1}{\sqrt{n}}\sum g_{5i}(\beta_2) + o_p(1)\,,$$

where

$$\begin{aligned} & g_{5i}(\beta_2) \\ = \ & E_{W_1,X_3,N,\tilde{N},C}\left[W_1 X_3 \textstyle\int_0^\tau N(t)\exp\{\Lambda_0^*(t)\exp(\psi' X) + \psi' X\}\,\{h_i(t) + \Lambda_0^*(t)X' e_i\}\,d\tilde{N}(t)\right], \end{aligned}$$

the expectation with respect to the joint distribution of $(W_1, X_3, N, \tilde{N}, C)$. Let $g_i^*(\beta_2) = \sum_{j=1}^5 g_{ji}(\beta_2)$. Then it follows from the proof of the asymptotical normality for $\hat{\beta}_1$ that

$$\begin{aligned} \hat{U}_2(\beta_2) &= \hat{U}_1(\beta_2) + \tfrac{1}{n}\textstyle\sum_{i=1}^n w_{1i}x_{3i}(\hat{\bar{N}}_i - \bar{N}_i) \\ &= \tfrac{1}{n}\textstyle\sum_{i=1}^n g_i^*(\beta_2) + o_p(n^{-1/2})\,. \end{aligned}$$

This implies that $\sqrt{n}(\hat{\beta}_2 - \beta_2)$ has an asymptotically normal distribution with mean

zero and covariance matrix $\phi^{*-1}\,\Sigma^*\,(\phi^{*-1})'$, where $\Sigma^* = E(g_i^*\,g_i^{*'})$ and $\phi^* = E(-\partial g_i^*/\partial\beta_2) = \phi$.

Appendix B:
Proof of the Asymptotic Normality of $\hat{\beta}_2$ and $\hat{\gamma}$ in Chapter 3

Let $N_{ik}(t)$, $\tilde{N}_{ik}(t)$, $Y_i(t)$, $S_k^{(d)}(t;\gamma)$, $E_k(t;\gamma)$ and $V_k(s;\gamma)$ be defined as in Chapter 3 and $s_k^{(d)}(t;\gamma)$, $e_k(t;\gamma)$ and $v_k(s;\gamma)$ denote the limit processes of $S_k^{(d)}(t;\gamma)$, $E_k(t;\gamma)$ and $V_k(s;\gamma)$, respectively. Define

$$A(\gamma) \;=\; \sum_{k=1}^{K} \int_0^{\tau} v_k(t;\gamma)\, s_k^{(0)}(t;\gamma)\, d\nu_k(t)$$

and

$$F(\beta,\gamma) \;=\; E\left[\sum_{k=1}^{K} \{g_N(x_i'\beta)\}^{-2}\, g_N^{(1)}(x_i'\beta)\, \{g_H(x_i'\gamma)\}^{-1}\, \bar{N}_{ik}\, x_i\, x_i'\right].$$

For the asymptotic normality of $\hat{\beta}_2$ and $\hat{\gamma}$, we assume that the regularity conditions given in Cai and Schaubel (2004b) hold. Also we need the following conditions:

(a) $\{N_{ik}(\cdot), Y_i(\cdot), x_i', H_{ik}(\cdot)\}_{k=1}^{K}$ are i.i.d. for $i = 1, \cdots, n$;

(b) $P(Y_i(\tau) = 1) > 0$;

(c) $|x_{il}| < c_1$ almost surely for all i and l, where c_1 is a positive constant and x_{il} denotes the lth component of x_i;

(d) both matrices $A(\gamma)$ and $F(\beta,\gamma)$ are positive definite;

(e) $N_{ik}(\tau) < c_2$ and $H_{ik}(\tau) < c_2$ almost surely for all $i = 1, ..., n$ and $k = 1, \cdots, K$,

where c_2 is a positive constant;

(f) there exist some neighborhoods of β_0 and γ_0, respectively, within which $g_N(x_i'\beta) \geq c_3$ and $g_H(x_i'\gamma) \geq c_3$, and $s_k^{(d)}(t;\gamma)$ is uniformly continuous with respect to γ and $t \in [0,\tau]$, where again c_3 is a positive constant, $d = 0, 1, 2$.

For each i and k and given γ, define

$$d\hat{M}_{ik}(t;\gamma) = d\tilde{N}_{ik}(t) - Y_{ik}(t)\, g_H(x_i'\gamma)\, d\hat{\nu}_k(t;\gamma)\,,$$

which is a zero-mean process under model (3.2). Also define

$$\hat{\Sigma}_U = \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{k=1}^{K} x_i\, \bar{N}_{ik}\left\{g_N(x_i'\hat{\beta}_2)\right\}^{-1}\left\{g_H(x_i'\hat{\gamma})\right\}^{-1}\right]^{\otimes 2}, \qquad (B.1)$$

$$\hat{\Sigma}_H = \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{k=1}^{K}\int_0^{\tau}\left\{x_i\,\frac{g_H^{(1)}(x_i'\hat{\gamma})}{g_H(x_i'\hat{\gamma})} - E_k(t;\hat{\gamma})\right\} d\hat{M}_{ik}(t;\hat{\gamma})\right]^{\otimes 2}, \qquad (B.2)$$

and

$$\hat{\Sigma}_\gamma = A^{-1}(\hat{\gamma})\,\hat{\Sigma}_H\, A^{-1}(\hat{\gamma})\,.$$

The asymptotic normality of $\hat{\gamma}$ is given in Theorem 1 in Cai and Schaubel (2004b). In particular, they showed that $\sqrt{n}\,(\hat{\gamma} - \gamma_0)$ converges in distribution to a multivariate normal vector with mean zero and covariance matrix that can be consistently estimated by $\hat{\Sigma}_\gamma$. For the asymptotic normality of $\hat{\beta}_2$, using the Taylor series expansions of $U_n(\hat{\beta}_2, \hat{\gamma})$ and $H_n(\hat{\gamma})$ around β_0 and γ_0 and based on the regularity conditions given

above, one can easily show that $\sqrt{n}\,(\hat{\beta}_2 - \beta_0)$ has the same asymptotic distribution as

$$-\,F^{-1}\left\{U_n(\beta_0,\gamma_0) \;-\; D\,A^{-1}(\gamma_0)\,H_n(\gamma_0)\right\}\,, \tag{B.3}$$

where

$$F \;=\; F(\beta_0,\gamma_0) \;=\; E\left\{\frac{\partial U_i^*(\beta,\gamma_0)}{\partial\beta}\right\}_{\beta=\beta_0}\,,$$

$$D \;=\; D(\beta_0,\gamma_0) \;=\; E\left\{\frac{\partial U_i^*(\beta_0,\gamma)}{\partial\gamma}\right\}_{\gamma=\gamma_0}\,,$$

and

$$U_i^*(\beta,\gamma) \;=\; \sum_{k=1}^{K} x_i\,\bar{N}_{ik}\left\{g_N(x_i'\beta)\right\}^{-1}\left\{g_H(x_i'\gamma)\right\}^{-1}.$$

For $U_n(\beta_0,\gamma_0)$ and $H_n(\gamma_0)$, as in Sun and Wei (2000), it can be easily shown that they asymptotically have a joint multivariate normal distribution with mean zero and covariance matrix that can be approximated by

$$\hat{\Gamma} \;=\; \begin{pmatrix} \hat{\Sigma}_U & \hat{\Sigma}_{UH} \\ \hat{\Sigma}_{UH}' & \hat{\Sigma}_H \end{pmatrix}\,,$$

where $\hat{\Sigma}_U$ are $\hat{\Sigma}_H$ are defined in (B.1) and (B.2), respectively, and

$$\hat{\Sigma}_{UH} \;=\; \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{k=1}^{K} x_i\,\bar{N}_{ik}\left\{g_N(x_i'\hat{\beta}_2)\right\}^{-1}\left\{g_H(x_i'\hat{\gamma})\right\}^{-1}\right]$$

$$\times\left[\sum_{k=1}^{K}\int_0^{\tau}\left\{x_i\,\frac{g_H^{(1)}(x_i'\hat{\gamma})}{g_H(x_i'\hat{\gamma})} \;-\; E_k(t;\hat{\gamma})\right\}\,d\hat{M}_{ik}(t;\hat{\gamma})\right]'\,.$$

Thus it follows from (B.3) that $\sqrt{n}\,(\hat{\beta}_2 - \beta_0)$ has an asymptotic normal distribution with mean zero and covariance matrix that can be consistently estimated by the quantity given in (3.6).

The joint asymptotic normality of $\hat{\beta}_2$ and $\hat{\gamma}$ directly follows from (B.3) and the fact that $\sqrt{n}\,(\hat{\gamma} - \gamma_0)$ has the same asymptotic distribution as

$$G_0\,(\,U_n(\beta_0, \gamma_0)', H_n(\gamma_0)'\,)'\,.$$

These also prove the consistency of the estimate given in (3.7) for their covariance matrix.

BIBLIOGRAPHY

Aalen, O. O. (1975). *Statistical Inference for a Family of Counting Processes*. PhD Thesis, University of California, Berkeley.

Andersen, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. New York: Springer.

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

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

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

Breslow, N. E. (1984). Extra-Poisson variation in log-linear models. *Applied Statistics* **33**, 38-44.

Byar, D. P. (1980). The veterans administration study of chemoprophylaxis for recurrent stage I bladder tumors: comparisons of placebo, pyridoxine, and topical thiotepa. In *Bladder Tumors and Other Topics in Urological Oncology*, M. Pavane-Macaluso, P. H. Smith, and F. Edsmyr (eds), 363-370. New York: Plenum.

Cai, J. and Prentice, R. L. (1995). Estimating equations for hazard ratio parameters based on correlated failure time data. *Biometrika* **82**, 151-164.

Cai, J. and Schaubel, D. E. (2004a). Analysis of recurrent event data. *Handbook of Statistics* **23**, 603-623.

Cai, J. and Schaubel, D. E. (2004b). Marginal means/rates models for multiple type recurrent event data. *Lifetime Data Analysis* **10**, 121-138.

Chen, B. E., Cook, R. J., Lawless, J. F., and Zhan, M. (2005). Statistical methods for multivariate interval-censored recurrent events. *Statistics in Medicine* **24**, 671-691.

Chen, Y. Q., Wang, M.-C., and Huang, Y. (2004). Semiparametric regression analysis on longitudinal pattern of recurrent gap times. *Biostatistics* **5**, 277-290.

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

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

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

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B* **34**, 187-220.

Cox, D. R. (1975). Partial likelihood. *Biomerika* **62**, 269-276.

De Gruttola, V. and Tu, X. (1994). Modelling progression of CD4-lymphocyte count and its relationship to survival time. *Biometrics* **50**, 1003-1014.

Diamond, I. D. and McDonald, J. W. (1991). The analysis of current status data. In *Demographic Applications of Event History Analysis*, J. Trussel, R. Hankinson, and J. Tilton (eds), 231-252. Oxford: Oxford University Press.

Diamond, I. D., McDonald, J. W., and Shah, I. H. (1986). Proportional hazards models for recurrent status data: application to the study of differentials in age at weaning in Pakistan. *Demography* **23**, 607-620.

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

Dinse, G. E. and Lagakos, S. W. (1983). Regression analysis of tumor prevalence data. *Applied Statistics* **32**, 236-248.

Fleming, T. R. and Harrington, D. P. (1991). *Counting Process and Survival Analysis*. New York: Wiley.

Gladman, D. D., Farewell, V. T., and Nadeau, C. (1995). Clinical indicators of progression in psoriatic arthritis (PsA): multivariate relative risk model. *Journal of Rheumatology* **22**, 675-679.

Hinde, J. (1982). Compound Poisson regression models. In *GLIM 82: Proceedings of the International Conference in Generalized Linear Models*, R. Gilchrist (ed), 109-121. Berlin: Springer-Verlag.

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

Huang, C.-Y. and Wang, M.-C. (2004). Joint modeling and estimation for recurrent event processes and failure time data. *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.

Huang, Y. and Chen, Y. Q. (2003). Marginal regression of gaps between recurrent events. *Lifetime Data Analysis* **9**, 293-303.

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*. Second edition, New York: Wiley.

Lawless, J. F. (1987). Regression models for Poisson process data. *Journal of the American Statistical Association* **82**, 808-815.

Lawless, J. F. and Nadeau, J. C. (1995). Some simple robust methods for the analysis of recurrent events. *Technometrics* **37**, 158-168.

Lawless, J. F. and Zhan, M. (1998). Analysis of interval-grouped recurrent-event data using piecewise constant rate functions. *The Canadian Journal of Statistics* **26**, 549-565.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.

Lin, D. Y., Sun, W., and Ying, Z. (1999). Nonparametric estimation of the gap time distributions for serial events with censored data. *Biometrika* **86**, 59-70.

Lin, D. Y., Wei, L. J., and Ying, Z. (1998). Accelerated failure time models for counting processes. *Biometrika* **85**, 605-618.

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.

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

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

Pepe, M. S. and Cai, J. (1993). Some graphical displays and marginal regression analyses for recurrent failure times and time dependent covariates. *Journal of the American Statistical Association* **88**, 811-820.

Prentice, R. L., Williams, B. J., and Peterson, A. V. (1981). On the regression analysis of multivariate failure time data. *Biometrika* **68**, 373-379.

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

Roy, J. and Lin, X. (2002). Analysis of multivariate longitudinal outcomes with non-ignorable dropouts and missing covariates: changes in methadone treatment practices. *Journal of the American Statistical Association* **97**, 40-52.

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 Study. *Annals of Internal Medicine* **95**, 257-282.

Siannis, F., Farewell, V. T., Cook, R. J., Schentag, C. T., and Gladman, D. D. (2006). Clinical and radiological damage in psoriatic arthritis. *Annals of Rheumatic Disease* **65**, 478-481.

Sinha, D. and Maiti, T. (2004). A Bayesian approach for the analysis of panel-count data with dependent termination. *Biometrics* **60**, 34-40.

Song, X., Davidian, M., and Tsiatis, A. A. (2002). A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. *Biometrics* **58**, 742-753.

Staniswalls, J. G., Thall, P. F., and Salch, J. (1997). Semiparametric regression analysis for recurrent event interval counts. *Biometrics* **53**, 1334-1353.

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 the American Statistical Association* **88**, 1449-1454.

Sun, J. and Kalbfleisch, J. D. (1995). Estimation of the mean function of point processes based on panel count data. *Statistica Sinica* **5**, 279-290.

Sun, J. and Matthews, D. E. (1997). A random-effect regression model for medical follow-up studies. *The Canadian Journal of Statistics* **25**, 101-111.

Sun, J. and Rai, S. N. (2001). Non-parametric tests for the comparison of point processes based on incomplete data. *Scandinavian Journal of Statistics* **28**, 725-732.

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.

Sun, L., Park, D., and Sun, J. (2006). The additive hazards model for recurrent gap times. *Statistica Sinica* **16**, 919-932.

Thall, P. F. (1988). Mixed Poisson likelihood regression models for longitudinal interval count data. *Biometrics* **44**, 197-209; correction, **45**, 1039.

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

Thall, P. F. and Vail, S. C. (1990). Some covariance models for longitudinal count data with overdispersion. *Biometrics* **46**, 657-671.

Wang, M.-C. and Chang, S.-H. (1999). Nonparametric estimation of a recurrent survival function. *Journal of the American Statistical Association* **94**, 146-153.

Wang, M.-C., Qin, J., and Chiang, C.-T. (2001). Analyzing recurrent event data with informative censoring. *Journal of the American Statistical Association* **96**, 1057-1065.

Wei, L. J. and Lachin, J. M. (1984). Two-sample asymptotically distribution-free tests for incomplete multivariate observations. *Journal of the American Statistical Association* **79**, 653-661.

Wei, L. J., Lin, D. Y., and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association* **84**, 1065-1073.

Wellner, J. A. and Zhang Y. (1998). Large sample theory for an estimator of the mean of a counting process with panel count data. *Technical Report*. Department of Statistics, University of Washington, Seattle.

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

Wellner, J. A., Zhang, Y., and Liu, H. (2004). A semiparametric regression model for panel count data: when do pseudo-likelihood estimators become badly inefficient? *Proceedings of the Second Seattle Symposium in Biostatistics*, New York: Springer, 143-174.

Wulfsohn, M. S. and Tsiatis, A. A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics* **53**, 330-339.

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

Zeng, D. and Lin, D. Y. (2007). Semiparametric transformation models with random effects for recurrent events. *Journal of the American Statistical Association* **102**, 167-180.

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

Zhang, Y. and Jamshidian, M. (2003). The gamma-frailty Poisson model for the nonparametric estimation of panel count data. *Biometrics* **59**, 1099-1106.

Table 2.1: Estimation of β with a Homogeneous Observation Process

	True β				
	-2	-1	0	1	2
	$x_i \sim Bernoulli\,(0.5)$				
$\hat{\beta}$	-1.9890	-0.9818	0.0241	1.0160	2.0236
SSD	0.1717	0.1528	0.1434	0.1442	0.1452
BSD	0.1597	0.1451	0.1402	0.1391	0.1417
	$x_i \sim N\,(0\,,\,0.25)$				
$\hat{\beta}$	-1.9638	-0.9845	0.0115	1.0062	2.0043
SSD	0.3082	0.3106	0.3198	0.3187	0.3640
BSD	0.3025	0.2957	0.2926	0.3080	0.3361

Table 2.2: Estimation of β with a Nonhomogeneous Observation Process

	True β				
	-2	-1	0	1	2
	$x_i \sim Bernoulli\,(0.5)$				
$\hat{\beta}$	-1.9670	-0.9755	0.0329	1.0330	2.0340
SSD	0.1987	0.1876	0.1806	0.1842	0.1840
BSD	0.1886	0.1770	0.1748	0.1767	0.1755
	$x_i \sim N\,(0\,,\,0.25)$				
$\hat{\beta}$	-1.9391	-0.9472	0.0334	1.0268	2.0137
SSD	0.3799	0.3761	0.3737	0.3995	0.4468
BSD	0.3645	0.3557	0.3558	0.3783	0.4131

Table 2.3: Estimation of β with Covariate-Dependent Follow-up Times

	True β				
	-2	-1	0	1	2
	$x_i \sim Bernoulli\,(0.5)$				
$\hat{\beta}$	-1.9815	-0.9903	0.0200	1.0127	2.0227
SSD	0.1447	0.1216	0.1305	0.1216	0.1273
BSD	0.1340	0.1216	0.1167	0.1154	0.1176
	$x_i \sim N\,(0\,,\,0.25)$				
$\hat{\beta}$	-1.9816	-0.9881	0.0265	1.0184	2.0211
SSD	0.2558	0.2426	0.2530	0.2865	0.3133
BSD	0.2447	0.2364	0.2410	0.2540	0.2831

Table 3.1: Simulation Results for Covariate-Independent Observation Processes

Sample Size	σ_Q^2	β_0	BIAS	SEE	SSE	CP
	1	-1	-0.0025	0.2276	0.2330	0.945
		0	0.0045	0.2273	0.2331	0.946
$n = 100$		1	0.0020	0.2271	0.2378	0.935
	2	-1	0.0058	0.2600	0.2796	0.930
		0	0.0133	0.2618	0.2849	0.923
		1	-0.0042	0.2619	0.2853	0.927
	1	-1	-0.0095	0.1683	0.1616	0.957
		0	0.0009	0.1698	0.1621	0.952
$n = 200$		1	-0.0042	0.1690	0.1666	0.946
	2	-1	-0.0054	0.1993	0.2012	0.939
		0	0.0045	0.1963	0.1996	0.941
		1	-0.0032	0.1995	0.2084	0.947

Table 3.2: Simulation Results for Covariate-Dependent Observation Processes

Sample Size	σ_Q^2	β_0	BIAS	SEE	SSE	CP
	1	-1	-0.0056	0.2042	0.2059	0.949
		0	0.0072	0.2045	0.2015	0.952
$n = 100$		1	0.0111	0.2045	0.2111	0.940
	2	-1	0.0097	0.2409	0.2628	0.933
		0	0.0103	0.2408	0.2630	0.922
		1	-0.0033	0.2413	0.2669	0.921
	1	-1	-0.0004	0.1518	0.1467	0.959
		0	0.0029	0.1520	0.1441	0.955
$n = 200$		1	0.0116	0.1538	0.1449	0.956
	2	-1	0.0149	0.1825	0.1889	0.940
		0	0.0093	0.1827	0.1878	0.928
		1	0.0120	0.1824	0.1880	0.951

Table 3.3: Results of Joint and Univariate Analyses of Radiological and Functional Joint Damage Data from the University of Toronto Psoriatic Arthritis Clinic

	Observation Process			Joint Damage Process		
Covariate	$\hat{\gamma}$	SE($\hat{\gamma}$)	p-value	$\hat{\beta}_2$	SE($\hat{\beta}_2$)	p-value
			Multivariate Analysis			
Family history of psoriasis	0.1689	0.1165	0.1470	-1.4111	0.3913	0.0003
Duration of PsA in years	-0.0015	0.0057	0.7936	0.0587	0.0197	0.0029
Number of active joints	0.0030	0.0060	0.6106	0.0669	0.0194	0.0006
			Univariate Analysis - Radiological Damage			
Family history of psoriasis	0.1375	0.1403	0.3271	-0.9376	0.3653	0.0103
Duration of PsA in years	-0.0043	0.0063	0.4940	0.0340	0.0210	0.1057
Number of active joints	-0.0079	0.0075	0.2957	0.0751	0.0182	< 0.0001
			Univariate Analysis - Functional Damage			
Family history of psoriasis	0.1609	0.1200	0.1799	-1.5467	0.4397	0.0004
Duration of PsA in years	-0.0007	0.0058	0.9024	0.0646	0.0206	0.0017
Number of active joints	0.0048	0.0059	0.4154	0.0662	0.0211	0.0017

Table 4.1: Estimation of β_1 with $\beta_2 = \beta_3 = 0$

	True β_1				
	-2	-1	0	1	2
			$n = 100$		
$\hat{\beta}_1$	-2.0159	-1.0019	0.0027	0.9966	2.0071
SSD	0.1843	0.1224	0.0967	0.0807	0.0778
BSD	0.1847	0.1287	0.1001	0.0876	0.0837
CP	0.930	0.943	0.937	0.946	0.945
			$n = 200$		
$\hat{\beta}_1$	-2.0143	-1.0045	-0.0013	1.0018	2.0009
SSD	0.1273	0.0877	0.0660	0.0601	0.0539
BSD	0.1282	0.0900	0.0708	0.0624	0.0583
CP	0.930	0.942	0.957	0.948	0.947

Table 4.2: Estimation of β_1 with $\beta_2 = \beta_3 = 0.2$

	True β_1				
	-2	-1	0	1	2
			$n = 100$		
$\hat{\beta}_1$	-2.0135	-1.0037	-0.0097	0.9936	1.9925
SSD	0.1700	0.1264	0.1107	0.1043	0.1076
BSD	0.1644	0.1248	0.1096	0.1023	0.1036
CP	0.920	0.933	0.944	0.934	0.942
			$n = 200$		
$\hat{\beta}_1$	-2.0097	-1.0051	-0.0054	0.9938	1.9980
SSD	0.1334	0.0941	0.0895	0.0861	0.0890
BSD	0.1242	0.0950	0.0847	0.0838	0.0827
CP	0.938	0.946	0.940	0.939	0.940

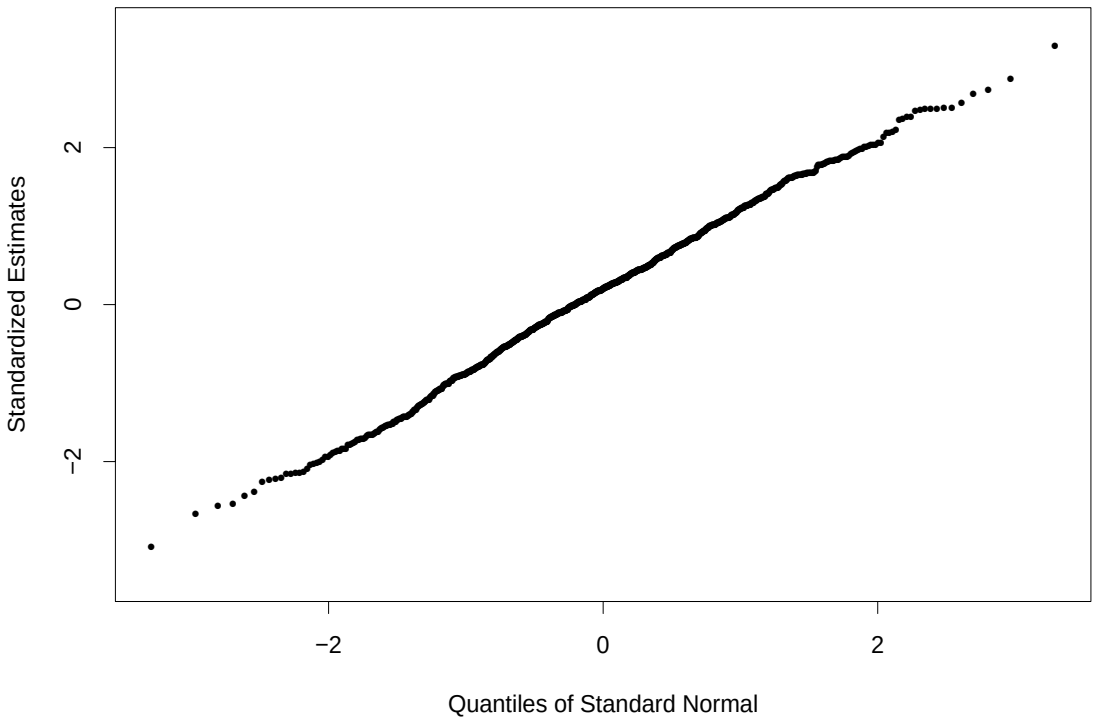

Figure 2.1: Quantile Plot with a Homogeneous Observation Process

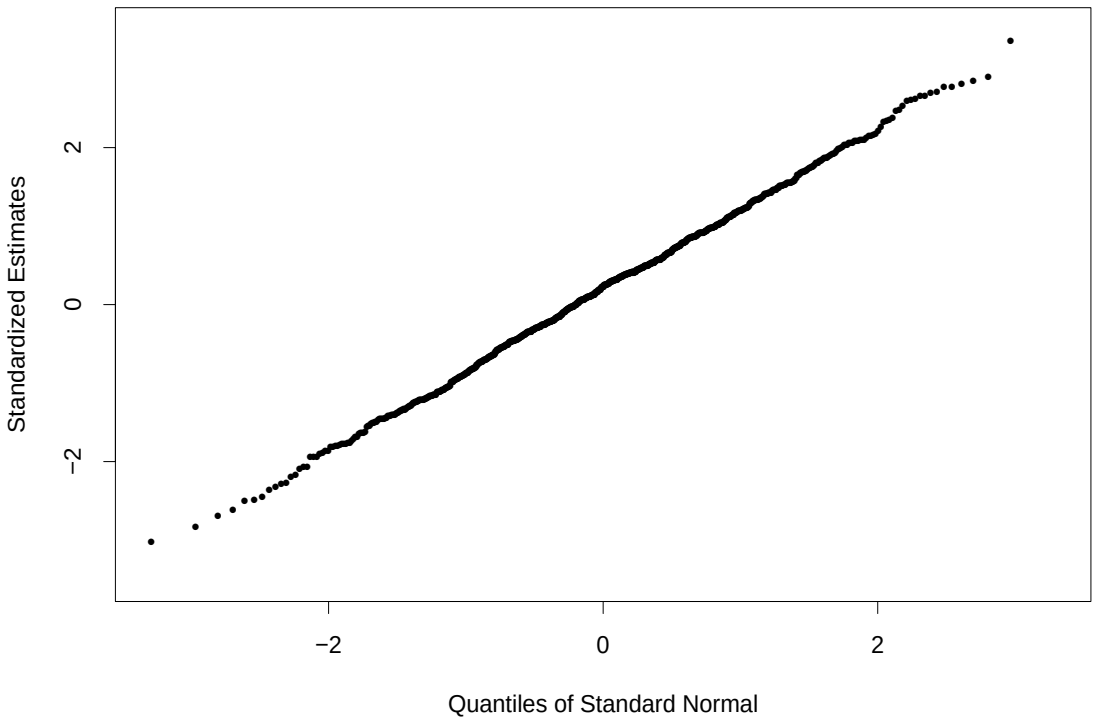

Figure 2.2: Quantile Plot with a Nonhomogeneous Observation Process

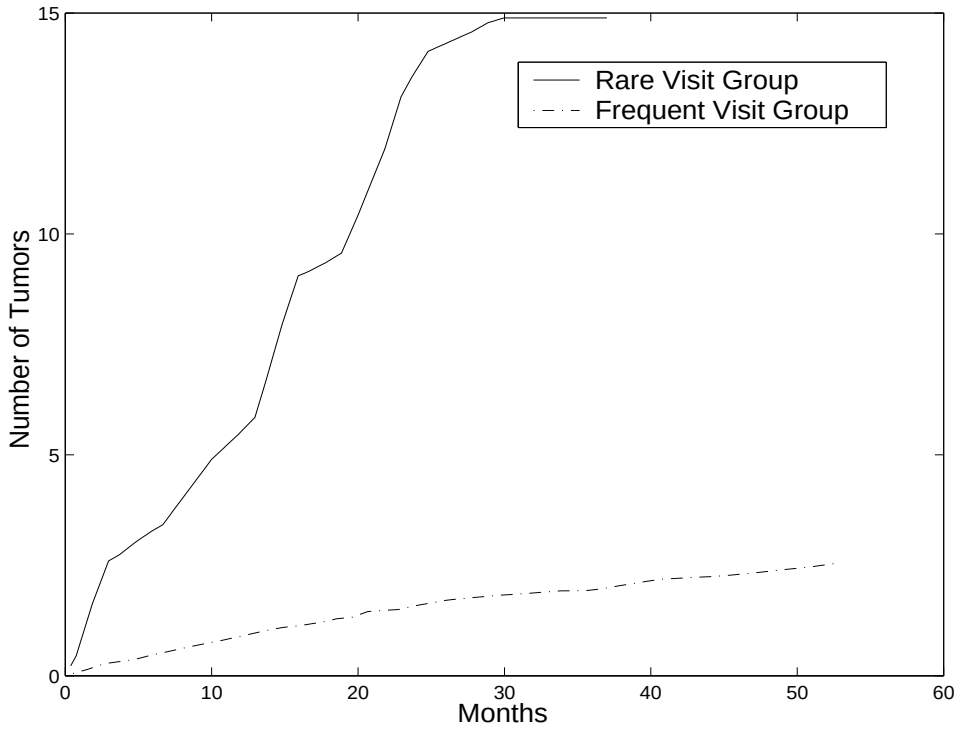

Figure 2.3: Estimates of the Baseline Mean Functions

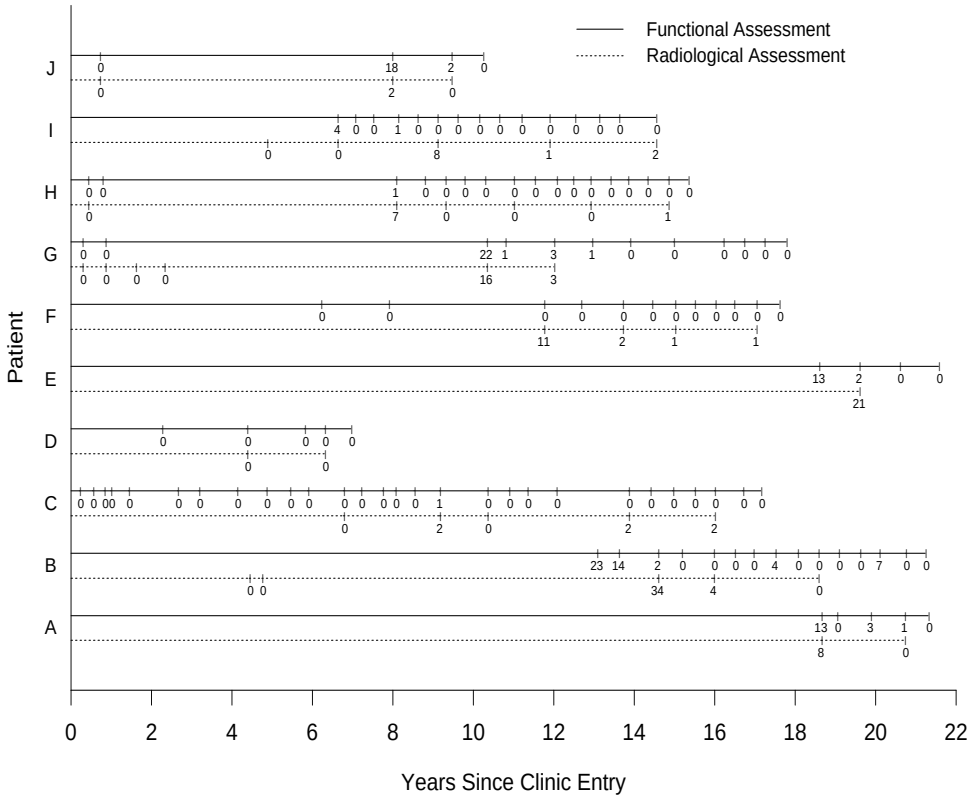

Figure 3.1: Timeline Diagram of Visits, Event Counts and Follow-up Durations for Sample of Patients in the University of Toronto Psoriatic Arthritis Clinic

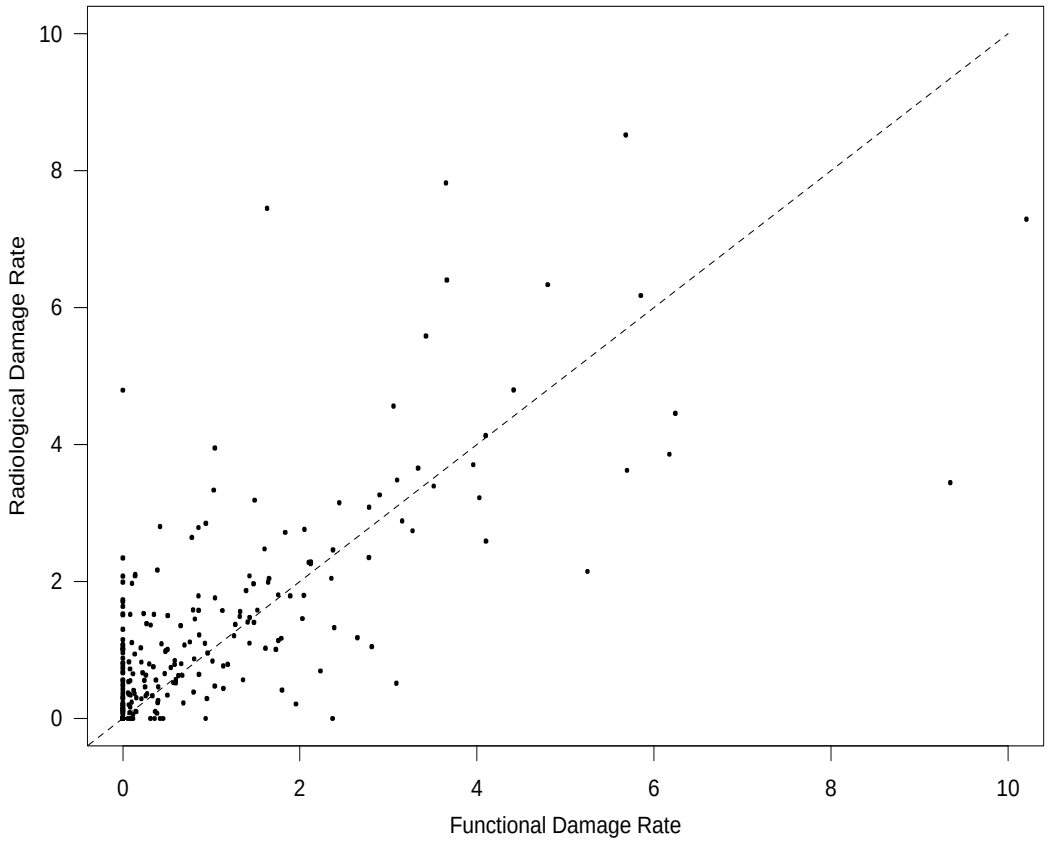

Figure 3.2: Crude Event Rates for Radiological Damage Against Functional Damage

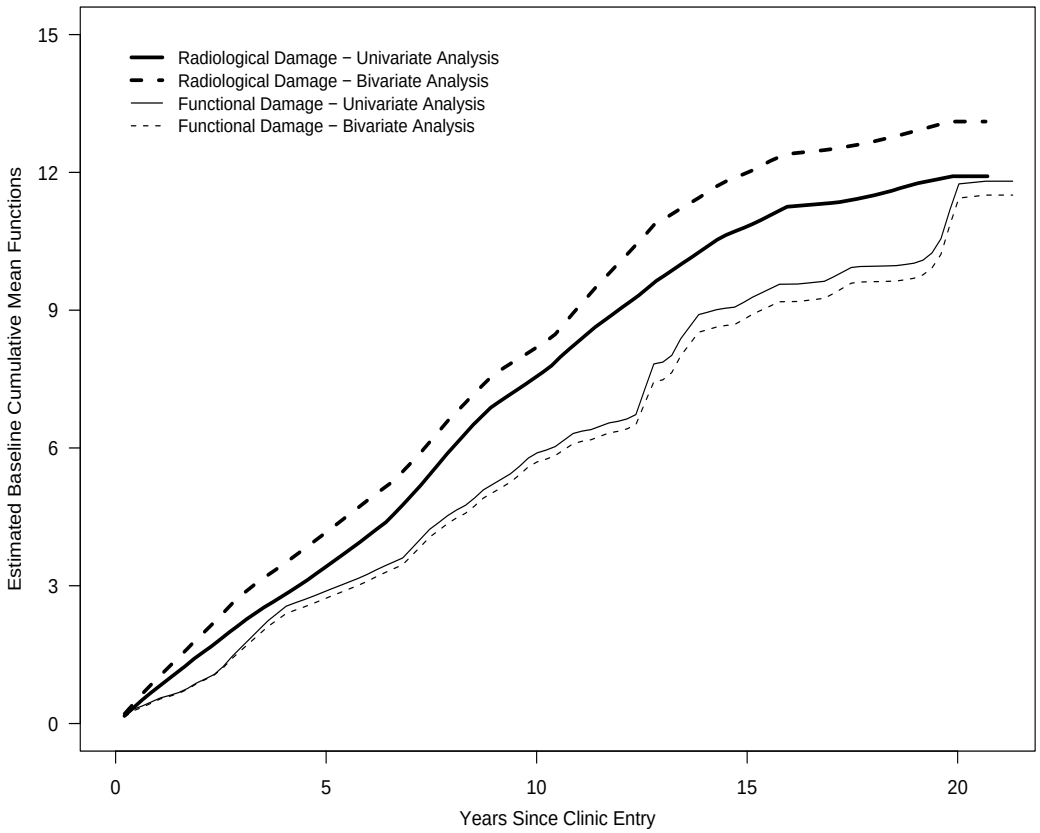

Figure 3.3: Estimated Baseline Cumulative Mean Functions from Univariate and Bivariate Regression Models Applied to Data from the University of Toronto Psoriatic Arthritis Clinic

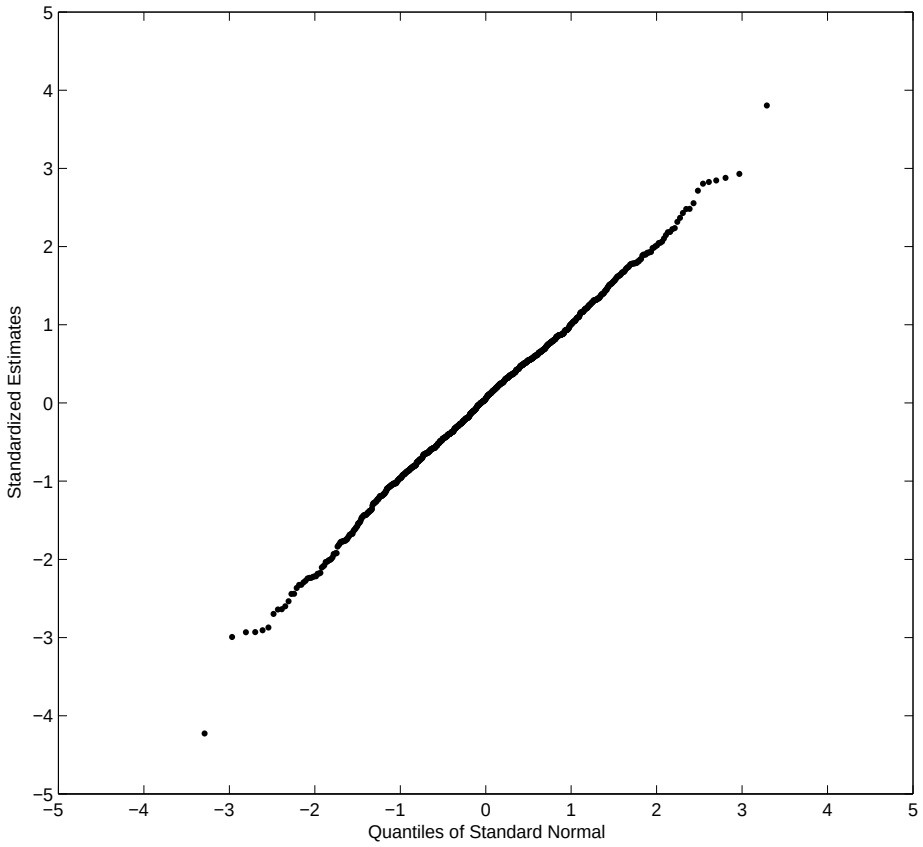

Figure 4.1: Quantile Plot with $\beta_1 = 0$, $\beta_2 = 0$, and $\beta_3 = 0$

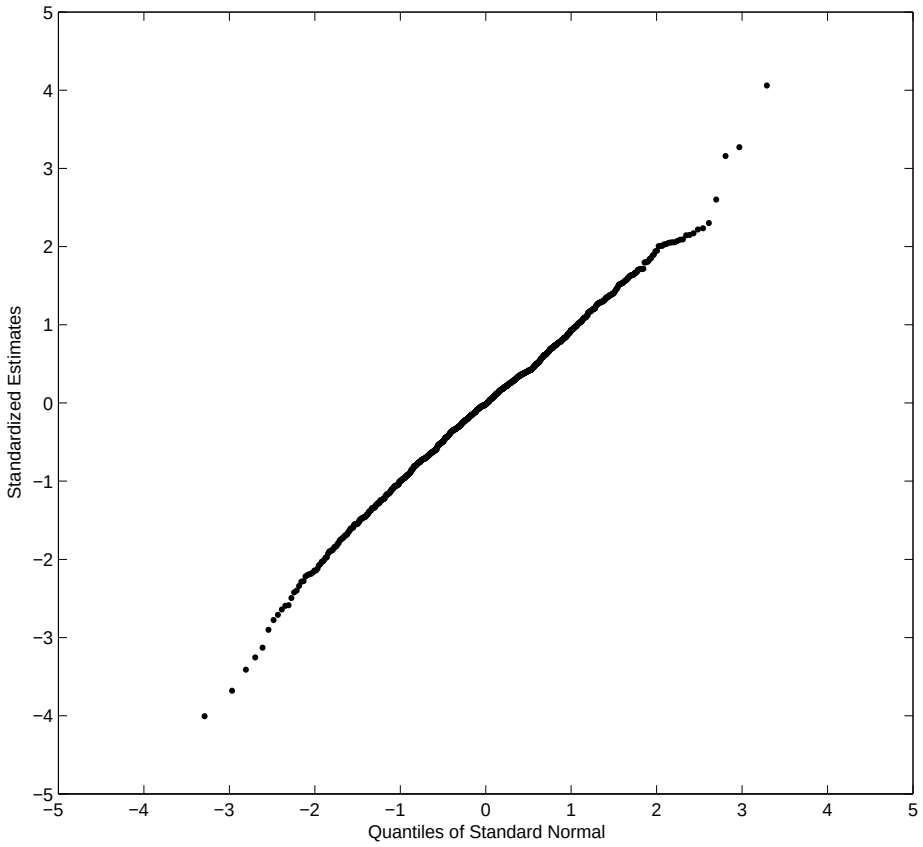

Figure 4.2: Quantile Plot with $\beta_1 = 1$, $\beta_2 = 0$, and $\beta_3 = 0$

VITA

Xin He was born on January 26, 1981, in Shijiazhuang, Hebei Providence, People's Republic of China. After attending public schools in Beijing, he received his B.S. in Statistics and B.A. in Economics from Peking University in 2003. He joined the graduate program in the Department of Statistics at the University of Missouri-Columbia in August 2003. He will receive his Ph.D. in Statistics in August 2007. As of September 2007, he will be serving as an Assistant Professor in the College of Public Health at The Ohio State University in Columbus, Ohio.