ABSTRACT

Title of Dissertation: SEMIPARAMETRIC ANALYSIS OF MULTIVARIATE PANEL COUNT DATA WITH AN INFORMATIVE OBSERVATION PROCESS

Chang Chen
Doctor of Philosophy, 2023

Dissertation Directed by: Professor Xin He
Department of Epidemiology and Biostatistics

Panel count data and recurrent event data often arise in event history studies. Unlike recurrent event data which are collected from studies that monitor subjects continuously, panel count data are encountered when subjects are observed only at discrete time points. In such case, the exact occurrence times of the events are unknown, but only the numbers of occurrences of the events between subsequent observation time points are recorded. Statistical analysis of panel count data have been studied based on two stochastic processes: an observation process and a response process that characterizes the occurrences of the events of interest.

The first part of the dissertation will present a likelihood-based joint modeling procedure for the regression analysis of univariate panel count data with dependent observation equations and time processes. The inference procedure involves estimating equations and an EM algorithm for the estimation of all involved parameters.

In the second part, we will extend the proposed methods to multivariate panel count data,

which occurs when a recurrent event study involves several related types of recurrent events. In particular, we will present three types of multivariate modeling scenarios and the corresponding inference procedures. A model checking procedure is developed for the proposed univariate models and all three types of multivariate models.

Simulation studies indicate that the proposed inference procedures have a good and consistent performance across various situations. The proposed methods are applied to a skin cancer study with bivariate panel count data on the occurrences of two types of related non-melanoma skin cancers.

SEMIPARAMETRIC ANALYSIS OF MULTIVARIATE PANEL COUNT DATA WITH AN INFORMATIVE OBSERVATION PROCESS

by

Chang Chen

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2023

Advisory Committee:
Professor Xin He, Chair/Advisor
Professor Mei-Ling Ting Lee
Professor Joan Jian-Jian Ren
Professor Paul J. Smith
Professor Yan Li, Dean's Representative

Acknowledgments

First of all, I would like to express my most sincere gratitude and appreciation to my dissertation advisor, Dr. Xin He, for his guidance, patience and dedication throughout the development of this very interesting project. Dr. He devotes an incredible amount of time into my dissertation research. Without his encouragement and support, it would be impossible for me to get progress and overcome all of the challenges along the journey. I also wish to express my deep appreciation to all of my committee members, Professor Paul J. Smith, Professor Joan Jian-Jian Ren, Professor Mei-Ling Ting Lee and Professor Yan Li, for their insightful perspectives and valuable comments that greatly improved my dissertation.

I would also like to give my loving gratitude to every member of my family, especially my wife Vicky, my parents and my grandparents, for their love, understanding and endless support during the years of my dissertation research. It is their love ultimately makes everything possible for me.

Finally, I acknowledge help and support from all of the staff members in the Department of Mathematics and the Department of Epidemiology and Biostatistics, and highly grateful to every faculty from both departments who helped launching this fantastic PhD program.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction and Literature Reviews

1.1 Event History Studies

Event history study plays an important role in the area of biomedical research. One of the most essential topics in event history studies is survival analysis, which attempts to investigate the time to the first occurrence of the event of interest. However, in some research fields, the events may occur repeatedly and interest often lies in characterizing the underlying processes over time. Some examples include the occurrences of asthma attacks in respirology trials and epileptic seizures in neurology studies.

In general, the recurrent events concerned in event history studies can also be classified into two types. The first type is referred to as recurrent event data, which are collected from studies that monitor or observe subjects continuously and the information on the times of all occurrences of the events are collected. The other type is called panel count data, the focus of this dissertation, which arise when subjects can only be observed at discrete time points. In such case, the exact occurrence times of the events are unknown, but only the numbers of occurrences of the events between subsequent observation time points are recorded.

Panel count data and recurrent event data are different in the amount of relevant information available for statistical inference. However, most methods for the analysis of panel count data developed in the past extended from the analysis of recurrent event data, and it is worth reviewing literature before discussing the methods for panel count data analysis. The following section provides a brief review of some commonly used approaches for the analysis of recurrent event data.

1.2 Statistical Analysis of Recurrent Events

In many biomedical studies, the interests are often to examine the recurrent event process where the events of interest are repeatedly generated over time. The resulting recurrent event data may vary across different areas of application. In some areas interest may lie in a small number of processes but generating a large number of events; while in other areas, the interest is only for a relatively small number of recurrent events but lies in a relatively large number of processes. The latter is more common in biomedical studies, where a group of patients may experience multiple clinical events repeatedly over the study period.

1.2.1 Example of Recurrent Event Data

An example of recurrent event data was obtained on 35 players from a professional rugby league club competing in the 2008 Australian NRL competition (Ullah et al., 2014). Injury and participation data were collected from 29 matches (including all trial, fixture and finals matches). Injuries were defined as conditions associated with pain or disability that occurred during match participation, irrespective of the need for first aid, medical attention or time loss. Figure 1.1 shows the timing of the incidence of each injury event in relation to the total number of matches (each of 80 min or 1.33 h duration). Censoring sometimes occurred when a player (a) had not experienced the relevant outcome, by the end of the season; (b) was lost to follow-up; or (c) experienced a different event that made further follow-up impossible. The data structure shows the complex nature of the recurrent injuries in that, in some players, several injuries occurred and the time between injuries also differed across players.

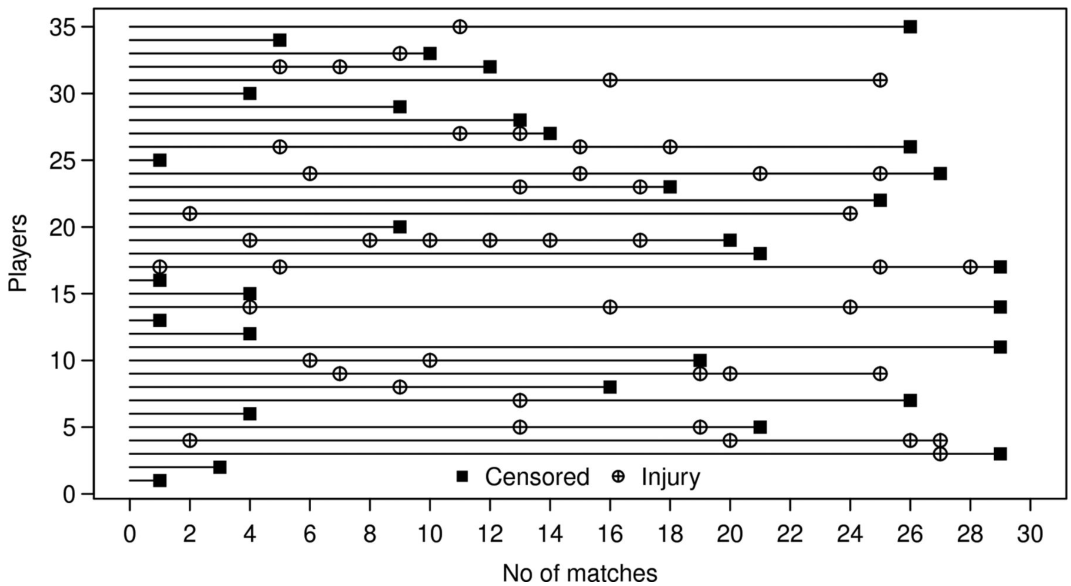


Figure 1.1: Recurrent injury history of 35 professional rugby league players (Ullah et al., 2014).

1.2.2 Poisson Processes

Modeling of recurrent events can be approached by the definition of intensity functions and the concept of counting process. For a single recurrent event process that starts at $t = 0$ with the time of the kth event denoted as T_k, we have that $0 \leq T_1 < T_2 < \cdots < T_k < \cdots$. Then the cumulative number of events can be generated by the corresponding counting process $\{N(t), t \geq 0\}$, where $N(t) = \sum_{k=1}^{\infty} I(T_k \leq t)$ is the number of events occurring over the time interval $[0, t]$. The number of events occurring over the interval $(s, t]$ is generally represented by $N(s, t) = N(t) - N(s)$.

Following this framework, we can then consider the probability distribution for the number of events in very short intervals denoted by $[t, t+\Delta t]$ given the history of event occurrence before time t. We further denote the number of events in the interval $[t, t+\Delta t]$ by $\Delta N(t) = N(t+\Delta t) - N(t)$, and the history of the process at time t can be represented by $H(t) = \{N(s) : 0 \leq s < t\}$. The event intensity function is then defined as

$$\lambda(t|H(t)) = \lim_{\Delta t \to 0} \frac{Pr\{\Delta N(t) = 1|H(t)\}}{\Delta t}, \tag{1.1}$$

which gives the instantaneous probability of an event occurring at t, conditional on the event history with the assumption that two events cannot occur simultaneously for events occurring in continuous time.

However, the mean function and variance function for the counting process $\{N(t), 0 \leq t\}$ are difficult to determine without assuming distributions of event counts or specifying the underlying recurrent event processes. Among all types of processes for recurrent events, Poisson process is considered as two of the canonical types of processes for recurrent events. The other canonical type, which we will not discuss in the dissertation, is the renewal process, in which the gap times between successive events are statistically independent. Unlike the renewal process, the Poisson process is based on counts and describes situations where the numbers of events between any two non-overlapping time intervals are assumed independent, which is considered more appropriate to model incidental events or events that are influenced by random external factors.

Poisson processes can be defined via the intensity function below

$$\lambda(t|H(t)) = \rho(t), \tag{1.2}$$

where $\rho(t)$ is a non-negative integrable function for $t > 0$, and $H(t)$ is the event history. Then the cumulative intensity function defined below

$$\Lambda(t) = \int_0^t \lambda(s)ds \tag{1.3}$$

is assumed to be continuous and finite for all $t > 0$. The key property of a Poisson process is that for $0 \leq s < t$, $N(s,t)$ has a Poisson distribution with mean $\Lambda(s,t) = \Lambda(t) - \Lambda(s)$, which can be proved by showing that the marginal probability of n events is the Poisson probability mass function,

$$\Pr(n \text{ events in } (s,t]) = \frac{\Lambda(s,t)^n}{n!} \exp\{-\Lambda(s,t)\} \quad (n = 0, 1, ...) \tag{1.4}$$

Poisson processes can be further differentiated as being homogeneous if $\lambda(t)$ is a constant ($\lambda(t) = \lambda$) or nonhomogeneous otherwise. The Poisson process is a Markov process, for which the probability of an event in $(t, t + \Delta t)$ may depend on t but is independent of the event history $H(t)$. This implies that the random variables $N(s,t)$ and $N(s',t')$ are assumed to be independent if $(s,t]$ and $(s',t']$ are non-overlapping intervals.

Fixed covariates can be incorporated in a Poisson process by specifying the intensity as a function of t with the following form

$$\lambda(t) = \lambda_0(t) \exp(\beta' X), \tag{1.5}$$

where β is a vector of regression parameters and the baseline intensity function $\lambda_0(t)$ is always positive. The model is fully parametric if the baseline intensity function $\lambda_0(t)$ is specified with a parametric distribution; otherwise, the model is semiparametric. The semiparametric model with an arbitrary positive baseline intensity function is sometimes called the Andersen-Gill model (Andersen et al. 1982).

Considering the situation when all of the covariates are fixed, the process $\{N(t), 0 \leq t\}$ is Poisson with mean function

$$E\{N(t)|X\} = \Lambda_0(t) \exp(\beta' X), \tag{1.6}$$

where $\Lambda_0(t) = \int_0^t \lambda(s)ds$ is the baseline mean function.

For some cases, one may notice that for individuals $i = 1, ..., n$, the variance of $N_i(t)$ appears to be substantially larger than the expectation of $N_i(t)$. However, under a Poisson model, the variance of $N_i(t)$ is assumed to be identical with the expectation of $N_i(t)$. If the counts are of interest but there is more interindividual variation in event occurrence to be accounted for by a Poisson process, we can consider incorporating unobservable random effects z_i for individuals $i = 1, ..., n$. Then the process $\{N_i(t), 0 \leq t\}$ given covariates x_i and z_i is a Poisson process with intensity function

$$\lambda(t|x_i, z_i, H_i(t)) = z_i \lambda_0(t) \exp(x_i' \beta), \tag{1.7}$$

where $x_i = (x_{i1}, x_{i2}, ..., x_{ip})$, and $z_1, ..., z_n$ are taken to be i.i.d. with finite mean and distribution function $G(z)$. Without loss of generality, the expectation of z_i is assumed to be 1. Random effects can be incorporated in multiple ways, and we will further discuss this topic in later sections.

1.2.3 Nonparametric Methods

Considering the situations when parametric assumptions are likely to violated, for such cases the baseline intensity functions are not assumed to be parametric. To make nonparametric inferences, we first consider $d\Lambda(t) = \lambda(t)dt$ as the expected value of $N(t, t+dt)$. Without adding covariates, the estimating equation for $d\Lambda(s)$ is given by Cook and Lawless (2007):

$$\sum_{i=1}^{n} Y_i(s)\{dN_i(s) - d\Lambda(s)\} = 0. \tag{1.8}$$

The solution for $d\hat{\Lambda}(s)$ from the above estimating equation is then given by $d\hat{\Lambda}(s) = d\bar{N}_{\cdot}(s)/Y_{\cdot}(s)$, where $d\bar{N}_{\cdot}(s) = \sum_{i=1}^{n} Y_i(s)dN_i(s)$ and $Y_{\cdot}(s) = \sum_{i=1}^{n} Y_i(s)$ represents the total number of observed events and the total number of subjects at risk within the interval $[s, s+ds)$ respectively. In addition, the estimator obtained from the estimating equation above is unbiased based on the fact that given $E\{d\bar{N}_i(s)|H_i(s), Y_i(s) = 1\} = d\Lambda(s)$,

$$E\{d\hat{\Lambda}(s)\} = E[E\{d\bar{N}_{\cdot}(s)/Y_{\cdot}(s)|Y_1(s), ..., Y_n(s)\}] = d\Lambda(s).$$

Since $\Lambda(t) = \int_0^t d\Lambda(s)$, an estimator for $\Lambda(t)$ can be obtained as follows

$$\hat{\Lambda}(t) = \int_0^t d\hat{\Lambda}(s) = \int_0^t \frac{d\bar{N}_{\cdot}(s)}{Y_{\cdot}(s)} = \sum_{m:t_m \le t} \frac{d\bar{N}_{\cdot}(t_m)}{Y_{\cdot}(t_m)}, \tag{1.9}$$

where $t_1 < t_2 < ... < t_M$ denotes M distinct event times among all individuals. It is worth noting that the estimator obtained above is constructed as a nonparametric maximum likelihood estimator of the mean function for Poisson processes (Cook and Lawless, 2007).

1.2.4 Semiparametric Regression

The semiparametric regression model with the profile likelihood estimates has been discussed in Cook and Lawless (2007), following Anderson and Gill (1982). Considering a regression-based model with the intensity function

$$\lambda_i(t) = \lambda_0(t)\exp(X_i'\beta), \tag{1.10}$$

where $\lambda_0(t)$, the baseline intensity function, is not assumed to have any specific parametric form. This is the semiparametric regression model and it is sometimes called the $Andersen - Gill$ model. We can explore this type of estimation and inference by first considering the concept of profile likelihood.

The profile likelihood estimates can be derived based on the following estimating equation

$$\sum_{i=1}^{n} Y_i(s)\{dN_i(s) - \exp(X_i'\beta)d\lambda_0(s)\} = 0 \quad s \geq 0. \tag{1.11}$$

The profile likelihood estimates for $d\lambda(s)$ can then be solved as

$$d\widetilde{\lambda}_0(s;\beta) = \frac{d\bar{N}.(s)}{\sum_{i=1}^{n} Y_i(s)\exp(X_i'\beta)}. \tag{1.12}$$

After substitution, a $p \times 1$ system of equations can be constructed, which yields

$$U_\beta(\beta) = \sum_{i=1}^{n} \int_0^\tau Y_i(s)G_i(s;\beta)dN_i(s), \tag{1.13}$$

where

$$G_i(s;\beta) = x_i(s) - \frac{\sum_{j=1}^{n} Y_j(s)\exp(x_j'\beta)x_j}{\sum_{j=1}^{n} Y_j(s)\exp(x_j'\beta)}. \tag{1.14}$$

The estimate is the solution of the above system of equations $U_\beta(\beta) = 0$ (Cook and Lawless, 2007).

1.2.5 Bibliographic Notes

Recurrent event data are frequently encountered in event history studies or longitudinal follow-up studies. Among those well-developed models, Poisson models have a long history of use in the analysis of recurrent events. In 1987, Lawless considered parametric and semiparametric methods for regression analysis based on Poisson process. Andersen and Gill (1982) proposed the semiparametric regression models that used counting process theory to derive asymptotic properties, which were generalized based on methods in the field of survival analysis (Cox, 1972). In 1995, Nelson discussed nonparametric estimation with the use of the population Mean Cumulative Function (MCF) for general processes. In 2001, Wang et al. considered the situation that the censoring mechanism is not independent of the recurrent event process. To deal with this case, the occurrence of recurrent events was modeled by a subject-specific nonstationary Poisson process via a latent variable, and nonparametric and semiparametric methods for estimating the cumulative rate function and regression parameters based on recurrent event data in informative censoring models were developed.

Besides Poisson models, smooth estimates of a common rate function have also been studied. In 1983, Ramlau–Hansen et al. developed a method to estimate counting process intensities using kernel functions to smooth the nonparametric Nelson estimator for the cumulative intensity. In addition, Andersen et al. (1993) discussed the choice of bandwidth and variance estimation for the smooth nonparametric estimates.

1.3 Statistical Analysis of Panel Count Data

In many biomedical studies, study subjects are not examined in a continuous way, instead, they are observed only at discrete time points. Under that situation, only the numbers of occurrences of the recurrent events of interest between subsequent observation times can be recorded. In other words, the exact occurrence times of the events are unknown. In the following sections, we will first discuss two examples of panel count data and then review multiple approaches developed for the analysis of panel count data.

1.3.1 Examples of Panel Count Data

For panel count data, the first example comes from a bladder cancer study conducted by the Veterans Administration Cooperative Urological Research Group (Byar, 1980; Wei et al., 1989; Wellner and Zhang, 1998), which produced univariate panel count data. In this study, all patients had superficial bladder tumours when they entered the trial. At baseline, bladder tumors were removed transurethrally and patients were randomly allocated into three treatments: placebo, thiotepa, and pyridoxine. Many patients had multiple recurrences of tumors during the study. These recurrent tumors were removed at the patient's clinic visits. For illustration, Figure 1.2 gives the numbers of new tumors discovered at on 36 patients from the place. In Figure 1.2, dot means no visit and the number represents the number of bladder tumors that occurred between the previous and current visits. The second column gives the size of the largest initial tumor, and

Patients' identification	Size†	Numbers of new tumours at the following months: 0 10 20 30 40 50 53
1	3	1 0 . . . .
2	1	2 0 . . 0 . . . .
3	1	1 . . . . . . 0 . . . .
4	1	5 . . 0 . . . . . 0 0 . . . .
5	1	4 0 . . 0 . 1 0 . . 0 . . . .
6	1	1 . . 0 . . . . . . 0 . . . 0 . . . .
7	1	1 . 0 . . . . 0 . 2 . . . 3 . 0 . . . .
8	1	1 . . 0 . . . . 0 . . . 0 . . . .
9	3	1 . . . . 2 . . . . 0 . 0 . . . . . 0 . . . .
10	3	1 . . 0 . . . 0 . . 6 . . . . 3 . . . 0 . . . 0 . . . .
11	1	1 0 . 8 . . . . 0 . 0 . . 0 . . 8 . . 0 . . . 8 . . . .
12	1	3 . . 1 . . 0 . . 1 0 . . 0 . 0 . 0 . . 0 8 . 0 . . . .
13	3	3 . . 0 . . 0 . . 0 . 0 . . 0 . . . 0 . . . . 0 . . . .
14	3	2 . . 0 . . . 8 . . 7 . . 0 . . 5 . . . . 7 . . . .
15	1	1 . . 1 . . 0 . . 0 . . 0 . . 1 . 0 . . 0 . 0 . . 3 . . . .
16	1	8 8 . . 0 . . 0 . . . 0 . . 0 . . 0 . . 0 . . . . 0 0 . . . .
17	4	1 . 4 . . . 0 . . . . 8 . . . .
18	2	1 . . 0 . . 0 . . . . . . 0 . . 0 . . . . . 0 0 . . 0 . . . .
19	2	1 . . . . . 0 . . . . 3 . . 0 . . . .
20	4	1 . . . . 0 . . 0 . . . . 0 . . . . 0 . . . .
21	2	1 . . 0 . . 0 . . . . 0 . . . . 0 . . . .
22	1	4 . 0 . . 0 . . 0 . . 0 . . . . . . 0 . . . . . 0 . . . . 0 . . . .
23	5	1 . 4 . . . . 0 . . . . . . 2 . . . . 4 0 . . . . 0 . 0 . . . .
24	1	2 . . 1 . . 3 . 3 . . . 3 0 . 0 0 0 0 . . . 0 0 . . 3 . . . 0 . . . .
25	6	1 . 0 . . 0 0 . . 0 . . . 0 . . . . 0 . . . . . 2 . 1 . . . .
26	3	1 0 . 0 0 . 0 . . 0 . . 2 . . 3 . . 0 . . . . . 1 . . . 0 . . 0 . . . .
27	2	1 . . 0 . . 0 . . 0 . . 0 . 0 . . 0 . . 0 . . . . . 0 . . . . . 0 . . . .
28	1	2 . . . 0 . . . 0 . 0 . . . 0 . . . . 0 . . . . 0 0 . 0 0 . . . .
29	1	2 . . 0 . . 0 . . 0 . 0 . . 0 . . 0 . . 0 . . 0 . . . . . . 0 . . . . . 0 . . . .
30	1	3 . . . . 0 . . . . 0 . . . . 8 . . . . . . 0 . . . .
31	2	1 0 . . 0 . . . . . 0 . . . . . . 0 . . . . . . 0 . . . . . . 0 . . . . . 0 . . . .
32	1	4 . 0 . . . . . . 8 . . . . 0 . . 2 . . . . 5 . 1 . . . 0 . . 0 . . . . 0 . . . 0 . . . .
33	1	5 . 0 . 0 0 . . 0 . . 0 . . . 0 1 . . 8 . . . 1 . . 0 . . 2 . . 0 . 1 . . 0 . . 3 . . . .
34	2	1 . . 0 . . 0 . . 0 . 0 . . 0 . . 0 . . 0 . . 0 . . 0 . . 0 . . . . . 0 0 . . . . 0 . . . .
35	1	1 . . 3 . 0 . 0 . . 0 . . . 0 . . . . . 0 . . . . . . 0 . . 0 . . . 0 . . 0 . . . . . 0 . . . .
36	6	2 . . 0 . . 1 0 0 . . 0 . . 0 . 0 0 . . 0 . . 0 . . . . 0 . . 0 . . . . 0 . . . .

Figure 1.2: Numbers of new tumors at each visit for 36 subjects in the placebo group

the number of initial tumors (at month 0) is given in column 3.

ID	Covariates		Observation number											
			1	2	3	4	5	6	7	8	9	10	11	12
			DFMO group											
1	(2, 56, M)	t	180	350	538	742	924	1,100	1,287	1,498	1,680	1,778		
		N_1	0	0	0	0	0	0	1	0	0	2		
		N_2	0	0	0	0	0	0	0	0	0	0		
2	(9, 76, M)	t	180	370	412	543	606	747	873	1,337				
		N_1	0	0	0	0	0	0	0	0				
		N_2	1	4	0	6	0	0	0	0				
3	(7, 76, F)	t	264	362	633	721	994	1,357	1,440	1,788				
		N_1	0	0	0	0	0	0	0	0				
		N_2	0	0	1	0	1	1	1	0				
4	(1, 49, F)	t	99	131	188	342	523	722	910	1,085	1,275	1,457	1,793	
		N_1	0	0	0	0	0	0	0	0	0	0	0	
		N_2	0	0	0	0	0	0	0	0	0	0	0	
5	(5, 64, F)	t	154	352	532	632	820	1,213	1,409	1,453	1,621	1,795		
		N_1	0	0	0	0	1	0	0	1	0	0		
		N_2	0	0	0	1	0	0	0	0	1	0		
6	(1, 82, M)	t	44	179	371	511	840							
		N_1	0	0	0	0	0							
		N_2	0	0	0	0	0							
7	(3, 53, M)	t	179	364	378	544	728	908	1,118	1,309	1,489	1,670	1,770	
		N_1	0	0	0	0	0	0	1	0	0	0	0	
		N_2	0	0	0	0	0	0	0	0	0	0	0	
8	(3, 50, M)	t	151	350	515	718	900	1,068	1,257	1,460	1,656	1,797		
		N_1	0	0	0	0	2	0	0	0	0	0		
		N_2	0	0	0	0	1	0	0	0	0	0		
9	(2, 80, F)	t	182	229	264	462	550	1,280	1,698					
		N_1	0	0	0	0	0	0	0					
		N_2	0	0	0	0	0	0	0					
10	(4, 60, F)	t	176	233	393	575	759	939	1,120	1,304	1,367	1,493	1,688	1,759
		N_1	0	0	0	0	0	0	0	2	0	0	0	0
		N_2	0	0	0	0	0	0	0	0	0	0	0	0

Figure 1.3: The Skin Cancer Chemoprevention Trial: observation times in days (t) and numbers of new skin cancers (N1 and N2 for basal cell carcinoma and squamous cell carcinoma respectively)

Figure 1.3 presents another example of bivariate panel count data on the occurrences of two types of related non-melanoma skin cancers, such as basal cell carcinoma and squamous cell carcinoma from the skin cancer chemoprevention trial (Sun, 2013). Figure 1.3 only contains data from the first 10 patients from the treatment group, but the study involved 291 skin cancer patients in total with one removal, and the remaining 290 patients were randomized into a treatment group with 143 patients and a placebo group with 147 patients. For each patient, a sequence of observation times and the numbers of occurrences of both basal cell carcinoma and squamous cell carcinoma between the observation times were recorded. It is worth noticing that in this study, both types of skin cancers were evaluated and recorded during each patient visit, and thus the observation times for each patient were the same across both event types. In addition to the treatment indicator, information was also available on three baseline covariates including patient's gender, age at the diagnosis, and the number of prior skin cancers from the first diagnosis to randomization. This data example will be further discussed and analyzed in Chapter 6.

Figure 1.4 presents another example of multivariate panel count data when the observation processes for various event types are different. The information was collected on the functional and radiological courses of disease for patients with psoriatic arthritis (Gladman et al., 1995). Figure 1.4 contains the timeline diagrams for a sample of 10 patients, where the duration of time from clinic entry to last contact for functional assessments (solid lines) and radiological assessments (dashed lines) were represented based on the length of the horizontal lines. As shown in the figure, the number of visits and the duration between two subsequent visits were largely different between the two types of assessments, there was no overlap in observation time between assessment types.

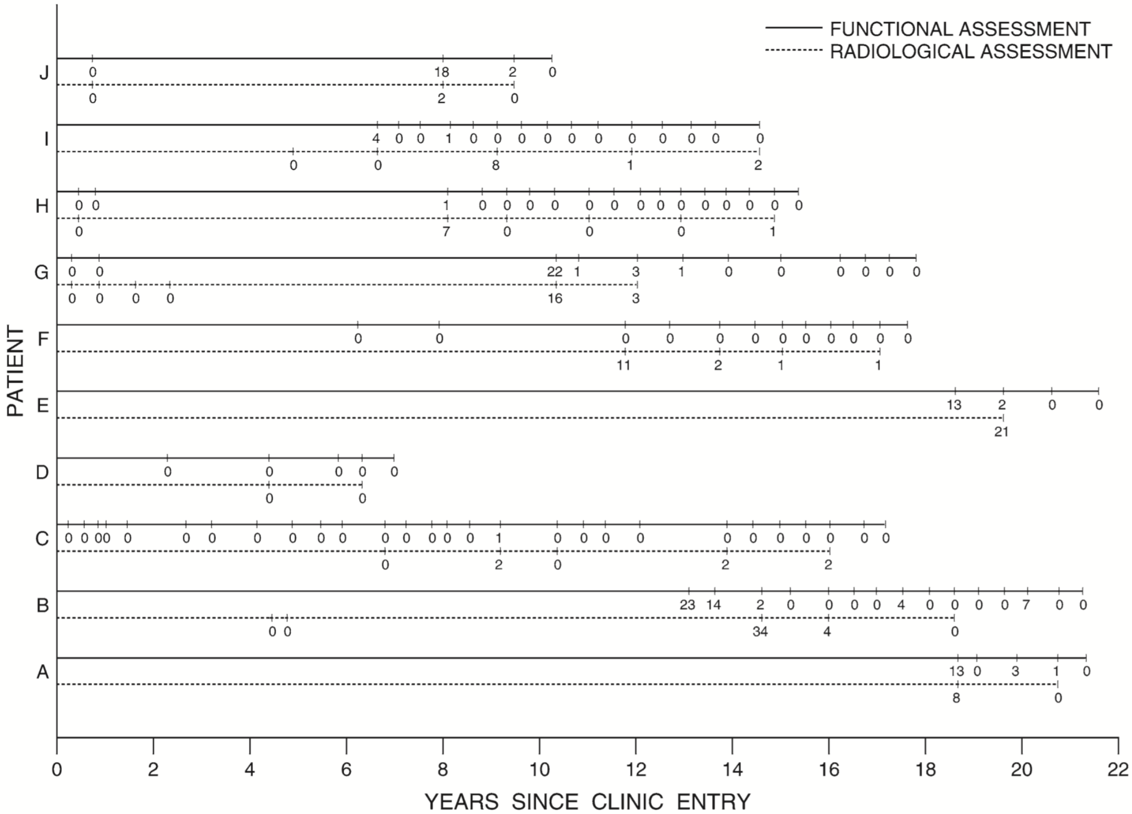


Figure 1.4: Timeline diagram for sample of patients in the University of Toronto Psoriatic Arthritis Clinic

1.3.2 Counting Processes and Martingales

Before introducing the methodology for the analysis of panel count data, it is worth introducing some notation and reviewing some concepts and models about counting processes, which are the foundation of the development of many statistical approaches and inferential procedures for event history analysis. Among many models developed previously, the Cox type intensity model for counting processes proposed by Andersen and Gill (1982) has been playing an essential role. Andersen and Gill developed a partial likelihood estimation procedure for regression parameters and established the large sample theory for the resulting estimators.

A counting process is a type of stochastic process $\{N(t); t \geq 0\}$ with $N(0) = 0$ and $N(t) < \infty$ almost surely. The process is right-continuous with probability one and has only jump discontinuities with size 1. An intensity process is defined to model the counting process

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P\{N(t+\Delta t^-) - N(t^-) = 1|H(t)\}}{\Delta(t)}, \tag{1.15}$$

where $H(t)$ is the event history prior to time t. Given $\lambda(t)$, the cumulative intensity process can be obtained as $\Lambda(t) = \int_0^t \lambda(s)ds$.

Counting processes have played even more fundamental roles for the analysis of event history studies when connected with martingale methods. Let $M(t)$ be an integrable stochastic process with $E\{|M(t)|\} < \infty$ for all t, and let $H(t)$ be the event history up to time t. Then $M(t)$ is a martingale if

$$E\{M(t)|H(s)\} = M(s) \tag{1.16}$$

for all $s \leq t$. If we further define

$$dM(t) = dN(t) - \lambda(t)dt, \tag{1.17}$$

Then the process

$$M(t) = N(t) - \int_0^t \lambda(s)ds \tag{1.18}$$

is a martingale with $E\{M(t)\} = 0$.

1.3.3 Likelihood-based Approach

Considering a recurrent event study that only observes the number of events that have occurred over some time interval for each of the n independent subjects. Let N_i be the observed event count for each subject i, and suppose that a vector of covariates $\mathbf{X}_i, i = 1, ..., n$ exists in the study. To make inferences about the effects of covariates on the occurrence rate of the event of interest, we first assume that given $\mathbf{X}_i$, the conditional mean of N_i has the following form

$$E(N_i|\mathbf{X}_i) = \exp(\beta^T \mathbf{X}_i), \tag{1.19}$$

where β denotes a vector of regression parameters. For statistical inference on β, some likelihood-based procedures can be employed if the N_i are assumed to follow a Poisson distribution. In addition, some estimating equation-based procedures can be used without the Poisson assumption. In this section we first discuss the likelihood-based procedures. Assume that $N_i's$ follow Poisson distributions with the mean given by

$$E(N_i|\mathbf{X}_i) = \exp(\beta^T \mathbf{X}_i), \tag{1.20}$$

then the log-likelihood function of β is proportional to

$$l(\beta) = \sum_{i=1}^{n} \Big\{ N_i \beta^T \mathbf{X_i} - \exp(\beta^T \mathbf{X}_i) \Big\}. \tag{1.21}$$

Then the maximum likelihood estimator $\hat{\beta}$ of β can be solved using the following score estimating equation

$$\sum_{i=1}^{n}\mathbf{X}_i\Bigg\{N_i - \exp(\beta^T\mathbf{X}_i)\Bigg\} = 0. \tag{1.22}$$

Since $\hat{\beta}$ is the MLE, it is easy to show its consistency. The covariance matrix of $\hat{\beta}$ can be expressed as

$$V(\hat{\beta}) = \frac{1}{n\Bigg\{\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^T\exp(\hat{\beta}^T\mathbf{X}_i)\Bigg\}}. \tag{1.23}$$

It is worth noting that the Poisson assumption is rather strong and it often does not hold in practice. In other words, $Var(N_i|\mathbf{X_i})$ may not be the same as $E(N_i|\mathbf{X_i})$. One common choice is to include a latent variable z_i such that given z_i, N_i follows a Poisson distribution with the mean

$$E(N_i|\mathbf{X}_i, z_i) = z_i\exp(\beta^T\mathbf{X}_i), \quad i = 1, ..., n \tag{1.24}$$

More details on the latent variable will be further discussed in section 1.3.4.3.

1.3.3.1 Likelihood-based Semiparametric Estimation Methods

In 2007, Wellner and Zhang proposed two types of likelihood-based approaches (i.e., profile likelihood and pseudo-likelihood methods) for the semiparametric estimation methods for the regression analysis of panel count data. Here, we focus on the profile likelihood approach. In Wellner and Zhang (2007), the following proportional mean regression model was considered

$$\Lambda(t|X) \equiv E\{N(t)|X\} = \exp(\beta_0^T X)\Lambda_0(t), \tag{1.25}$$

where Λ_0 is a monotone increasing baseline mean function, and β_0 and Λ_0 are the parameters of primary interest. In the model presented above, $\{N(t) : t \geq 0\}$ is a univariate counting process, $E\{N(t)|X\}$ is the expected number of events of interests.

Under the assumption that $N(t)$ is a nonhomogeneous Poisson process conditionally on X, the likelihood can be obtained using the conditional independence of the increments of $N(t)$, which are defined as $\Delta N(s,t] \equiv N(t) - N(s)$. The Poisson distribution of these increments is given by

$$P(\Delta N(s,t] = k|X) = \frac{[\Delta\Lambda((s,t]|X)]^k}{k!} \exp\{-\Delta\Lambda((s,t]|X)\}. \tag{1.26}$$

The log-likelihood can then be expressed as

$$l_n(\beta, \Lambda) = \sum_{i=1}^{n} \sum_{j=1}^{K_i} \left\{ \Delta N_{ij} \log \Delta\Lambda_{ij} + \Delta N_{ij} \beta^T X_i - \exp(\beta^T X_i) \Delta\Lambda_{ij} \right\}, \tag{1.27}$$

where K_i is the number of distinct time points for subject i, $\Delta N_{ij} \equiv N(T_{i,j}) - N(T_{i,j-1})$ and $\Delta\Lambda_{ij} \equiv \Lambda(T_{i,j}) - \Lambda(T_{i,j-1})$ for $i = 1, ..., n$ and $j = 1, ..., K_i$. Then the parameters $(\hat{\beta}_n, \hat{\Lambda}_n)$ can be obtained by maximizing the log-likelihood function above, which can be carried out in two steps via profile likelihood.

A doubly iterative algorithm is employed to update the estimates alternately. The algorithm and iteration are conducted by first choosing the initial $\beta^{(0)}$, and then for a given $\beta^{(k)}$, the updated estimate of $\Lambda^{(k)}$ is computed by the modified iterative convex minorant algorithm proposed by Jongbloed (1998) on the likelihood $l_n(\beta^k, \Lambda)$. The iteration is stopped when the relative change of the log-likelihood is small enough.

The second part is to obtain the updated estimate of $\beta^{(k+1)}$ by optimizing $l_n(\beta, \Lambda^{(k)})$ using the Newton-Raphson method. The iteration is stopped when the change of β is small enough. The algorithm is to repeat those two steps above until an updated convergence criterion is satisfied. It is also worth noting that the log-likelihood $l_n(\beta, \Lambda)$ is a concave function of the regression parameter β with negative definite Hessian matrix for any given monotone nondecreasing function Λ. In addition, the iterative algorithms proposed via both approaches converge very well.

One of the limitation for this profile likelihood approach is that the algorithm is not efficient. The pseudo-likelihood estimator, on the other hand, is computationally less intensive.

1.3.4 Estimating Equation Approach

In Sun and Wei (2000), the authors proposed a model for the mean function of counting process without assuming the event count process as a homogeneous or non-homogeneous Poisson process. Consider a study with n participants and for subject i, let $N_i(t)$ be the cumulative number of events that have occurred before time t, $0 \leq t \leq \tau$, where τ is a known fixed time point. Suppose that there exists a $p \times 1$ covariate vector $\mathbf{X}_i = (X_{i1}, ..., X_{ip})^T$. And the expected value of $\mathbf{X}_i$ is set to 0 without loss of generality. Then given $\mathbf{X}_i$, the mean function of $N_i(t)$ is of the following form

$$\mu_i(t) = \mu_0(t) \exp(\beta^T \mathbf{X}_i), \tag{1.28}$$

where $\mu_0(\cdot)$ is completely unspecified and β denotes a vector of unknown regression parameters.

Let C_i be the follow-up time for the ith subject, since not every subject can be followed until τ. In this case, $N_i(T_{i,j}), j = 1, ..., K_i$ is not observable when $C_i < T_{i,j} < \tau$. Given this fact, A new point process $\widetilde{N}_i(t)$ is defined for the ith subject's observation times

$$\widetilde{N}_i(t) = H_i\{\min(t, C_i)\}, \tag{1.29}$$

where

$$H_i(t) = \sum_{j=1}^{K_i} I(T_{i,j} \leq t), \tag{1.30}$$

and $I(\cdot)$ is an indicator function.

1.3.4.1 Observation and Follow-up Times Independent of Covariates

Following the discussion from the previous section, H_i and C_i are assumed to be mutually independent and independent of N_i and $\mathbf{X}_i$. In addition, $\mathbf{X}_i$ and $\exp(-\beta^T\mathbf{X}_i)N_i(T_{i,j})$ are uncorrelated at each observation time point $T_{i,j}$ under model (1.28). Then the following estimating function can be constructed

$$G(\beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1}^{K_i} \frac{\mathbf{X}_i N_i(T_{i,j}) I(T_{i,j} \leq C_i)}{\exp(\beta^T \mathbf{X}_i)} \tag{1.31}$$

Here $G(\beta)$ is an unbiased estimating function given that the expected value of $G(\beta)$ is 0. To facilitate the estimation procedure, we define the following quantity

$$\bar{N}_i = \sum_{j=1}^{K_i} N_i(T_{i,j}) I(T_{i,j} \leq C_i) = \int_0^\tau N_i(t) d\widetilde{N}_i(t)$$

Then the estimating function 1.31 can be expressed as

$$G(\beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\mathbf{X}_i \bar{N}_i}{\exp(\beta^T \mathbf{X}_i)} \tag{1.32}$$

To solve for β, note that

$$\frac{\partial G(\beta)}{\partial \beta} = -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\bar{N}_i \mathbf{X}_i \mathbf{X}_i^T}{\exp(\beta^T \mathbf{X}_i)}$$

is strictly negative definite, and thus the estimating equation $G(\beta) = 0$ has a unique solution.

1.3.4.2 Observation Times Dependent on Covariates

Some nonparametric and pseudolikelihood approaches have been proposed for the analysis of panel count data. Those methods assume that the counting process that characterizes the observation times or the observation process is independent of the point process that characterizes the occurrence of the event. This assumption may not be always true, as the observation process may be correlated with the point process of interest. In 2000, Sun and Wei proposed a regression-based method that considered a situation where observation times may depend on covariates.

We first make the assumption that the censoring time C_i is independent of the two processes H_i, N_i as well as the covariates $\mathbf{X}_i$. In addition, H_i is independent of N_i conditionally on $\mathbf{X}_i$. Furthermore, assume that the mean function of the counting process $H_i(t)$ is

$$\widetilde{\mu}_i(t) = \widetilde{\mu}_0(t) \exp(\gamma' \mathbf{X}_i).$$

The estimating equation for β can be simplified as

$$G(\beta, \gamma) = G(\beta + \gamma) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\mathbf{X}_i \bar{N}_i}{\exp((\beta + \gamma)^T \mathbf{X}_i)}$$

where

$$\bar{N}_i = \sum_{j=1}^{K_i} N_i(T_{i,j}) I(T_{i,j} \leq C_i).$$

1.3.4.3 Regression Analysis with Unobservable Frailty

In 2007, Sun et al. extended the method by assuming that

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

where z_i is a subject-specific unobservable positive frailty and α is an unknown parameter. It is also assumed that given (x_i, z_i), $H_i(\cdot)$ is a nonhomogeneous Poisson process with the intensity function

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

Under the above models, N_i and H_i can be correlated and their relationship is partly determined by the parameter α. With the same definition of $\bar{N}_i$, following Sun and Wei (2000), the estimation function below can be considered

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

where $\theta = \int_0^{\tau} \lambda_0(t) P(C_i \geq t) \mu_0(t) dt$, $\beta_1' = (\beta', 1 + \alpha, \log(\theta))$, $x_{1i}' = (x_i', \log(z_i), 1)$, and the

w_{1i}'s are weights that could depend on the x_i's and C_i's.

The estimating equation given above has a unique solution as the second derivative of β_1 is strictly negative definite. The estimation procedure is to first estimate γ by following Huang and Wang (2004) and first obtaining $\hat{\Lambda}_0(t)$ by

$$\hat{\Lambda}_0(t) = \prod_{s_l > t}(1 - \frac{d_l}{n_l}),$$

and estimating γ with the estimating equation

$$\sum_{i=1}^{n} w_{2i} x_{2i} \{K_i^* \hat{\Lambda}_0^{-1}(C_i) - \exp(\gamma_2' x_{2i})\} = 0,$$

where K_i^* is the number of distinct observation times for subject i. For the unknown z_i's, note that given (x_i, C_i, z_i), the expected value of K_i^* is equal to $z_i \Lambda_0(C_i) \exp(\gamma' x_i)$, which suggests that z_i can be replaced by

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

Given $\hat{\gamma}$ and the z_i's, β_1 can then be estimated by the solution to the estimating equation $\hat{U}_1(\beta_1) = 0$.

1.3.5 Bibliographic Notes

The Poisson process plays a major role in the parametric and semiparametric inference procedures. For the regression analysis of panel count data, two situations are usually considered. One is that the underlying recurrent event processes of interest are non-homogeneous Poisson processes, and the other is that the recurrent event processes are mixed Poisson processes, as discussed in Lawless (1987). In 1997, Staniswalls et al. also considered the situation where the underlying recurrent event process of interest are mixed Poisson processes and the baseline rate functions are completely unspecified. Using the mixed Poisson model is one of the two ways to relax the Poisson assumption. In 1993, Wedel et al. proposed an alternative by developing a latent class-based Poisson model.

Nonparametric estimation can be studied with or without assuming an underlying Poisson process. In 2000, Wellner and Zhang considered the non-parametric maximum likelihood estimator (NPMLE) and the non-parametric maximum pseudo likelihood estimator (NPMPLE) as two likelihood-based estimators of the mean function without assuming a non-homogeneous Poisson process model. In a recent paper in 2021, Zeng and Lin employed nonparametric maximum likelihood estimation and formulated the effects of potentially time-dependent covariates on multiple types of recurrent events through nonhomogeneous Poisson processes with random effects. Another type of nonparametric estimation is to consider the smooth estimation of the rate function of the recurrent event processes with the focus on kernel estimation, which was developed by Wand and Jones (1995). Bean and Tsokos (1980) discussed the methods commonly used for kernel estimation of density functions for the selection of the bandwidth parameter. However, more research remains to be done for nonparametric estimation with panel count data without assuming that the observation process is independent of the underlying recurrent event process of interest.

1.4 Outline

The analysis of recurrent event data and the analysis of panel count data are closely related, and the literature reviewed above provide a comprehensive coverage of well developed approaches for both data types. This dissertation will focus on two major topics on the analysis of panel count data. In Chapter 2, we start with univariate panel count data and consider the situation where the data can be described by non-homogeneous Poisson processes using semi-parametric inference procedures. A key advantage of the assumption is that the correlation between the observation process and the point process that characterizes the occurrence of the event are modeled with a covariance matrix under a bivariate normal distribution, which allows the inferences on both processes to be conducted simultaneously without multi-step procedures. Chapter 3 extends the proposed methods to multivariate panel count data. Three types of multivariate modeling assumptions and the corresponding inference procedures will be discussed. Chapter 4 will present model checking procedures for both univariate and multivariate models. The procedures will use martingale-based residuals to assess the objectivity of the proposed models by visually assessing how unusual the observed pattern is and by the supremum tests. In Chapter 5, we conduct the simulation studies and present the simulation results for all of the proposed models. The key idea is to show the unbiasedness of the proposed estimators and the performance of the proposed variance estimation procedures. In Chapter 6, the proposed methodology will be applied to analyze the bivariate panel count data on the occurrences of two types of related non-melanoma skin cancers, including basal cell carcinoma and squamous cell carcinoma, from the skin cancer chemoprevention trial. The results will be compared with the findings obtained from other literature. Chapter 7 concludes with some final remarks for the dissertation, including potential directions for future research and possible software package for the analysis of multivariate panel count data.

Chapter 2: Semiparametric Analysis of Panel Count Data

2.1 Introduction

This chapter presents a novel statistical approach for the analysis of univariate panel count data. As discussed in chapter one, panel count data involve an extra observation process compared to recurrent event data. In this chapter, we consider the situation where the underlying recurrent event process of interest is not independent of the observation process. More specifically, the goal is to model the correlation between the two processes.

Regression ananlysis of panel count data may start with modeling the intensity process with the intensity function of the observation process and the mean function of the underlying recurrent event process of interest. The intensity function and the mean function can be modeled parametrically or in a nonparametric way. However, it is often difficult to properly determine the parametric models. In our proposed method, we employ semiparametric approaches without assuming any parametric forms of the baseline intensity function or the baseline mean function.

We first discuss model assumptions in Section 2.2. For regression analysis of panel count data with informative or dependent observation processes, we then describe a joint modeling procedure that allows the underlying recurrent event process of interest and the observation process to be correlated, even conditional on covariates, in Section 2.3. To achieve this approach, a key assumption has to be made such that the relationship between the two processes can be characterized through some frailty terms or latent variables. The estimation procedure involves the use of an EM algorithm for the estimation of all involved parameters, which are presented in Section 2.4. Section 2.5 discusses the advantages of our proposed methods compared to other existing methods.

2.2 Models and Notations

Assuming that given X_i and a subject-specific unobservable frailty Z_i, $N_i(t)$ follows a nonhomogeneous Poisson process with the mean function

$$\mu_i(t) = \mu_0(t)\exp\{X_i'\beta + Z_i\}, \tag{2.1}$$

where $\mu_0(.)$ is a completely unspecified baseline mean function, and β is a vector of unknown regression parameters.

For the observation process, it is assumed that given X_i and a subject-specific unobservable frailty U_i, the observation process $H_i(t)$ follows a nonhomogeneous Poisson process with the intensity function

$$\lambda_i(t) = \lambda_0(t)\exp\{X_i'\gamma + U_i\}, \tag{2.2}$$

where $\lambda_0(.)$ is a completely unspecified baseline intensity function, and γ is a vector of unknown regression parameters.

To model the correlation between the two processes $N_i(t)$ and $H_i(t)$, it is assumed that (Z_i,

U_i) follows a bivariate normal distribution with mean (0, 0) and variance-covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_z^2 & \sigma_{zu} \\ \sigma_{zu} & \sigma_u^2 \end{pmatrix} \tag{2.3}$$

Under the above models and assumptions, there are 5 sets of unknown parameters. The first and second sets are $\mu_0(.)'s$ and $\lambda_0(.)'s$. Let M be the number of distinct time points of $\{T_{i,j}\}$. Then we have M unknown $\mu_0(.)'s$ to be estimated. A similar idea apply to $\lambda_0(.)'s$, which contains M unknown parameters.

The third and fourth sets are $\beta's$ and $\gamma's$. Let p be the number of available covariates, then we have p unknown $\beta's$ and p unknown $\gamma's$ to be estimated. The last set is the unknown parameters in the variance-covariance matrix. The variance-covariance matrix includes the variance for Z_i, the variance for U_i and their corresponding covariance. The detailed estimation procedure will be discussed in the following sections.

2.3 Joint Likelihood

We first set $T_{i0} = 0$ and $N_{iK_i} = N_i(T_{i,K_i})$, $\Delta N_{ij} = N_i(T_{i,j}) - N_i(T_{i,j-1})$ for $j = 1, ..., K_i$, $i = 1, ...n$. Assuming that the observations from the n subjects are i.i.d. copies, given the observed data $O_i = \{X_i, \boldsymbol{T_i}, \boldsymbol{N_i}, C_i\}$, $i = 1, ..., n$ with $\boldsymbol{T_i} = \{T_{i,0}, T_{i,1}, ..., T_{i,K_i}\}$ and $\boldsymbol{N_i} = \{N_i(T_{i,1}), ..., N_i(T_{i,K_i}))\}$, the likelihood function of $\theta = (\beta, \gamma, \Sigma, \mu, \Lambda)$ is proportional to

$$\prod_{i=1}^{n} \int \int f(\boldsymbol{T_i}, \boldsymbol{N_i} | X_i, C_i, u_i, z_i, \beta, \gamma, \mu, \Lambda) \phi(u_i, z_i; \Sigma) du_i dz_i, \tag{2.4}$$

where

$$f(\boldsymbol{T_i}, \boldsymbol{N_i}|X_i, C_i, u_i, z_i, \beta, \gamma, \mu, \Lambda)$$

$$= \frac{\exp(-\Lambda(C_i)e^{X_i'\gamma+u_i})[\Lambda(C_i)e^{X_i'\gamma+u_i}]^{K_i}}{K_i!} \prod_{j=1}^{K_i} \left(\frac{\lambda(T_{i,j})}{\Lambda(C_i)} \{\mu(T_{i,j}) - \mu(T_{i,j-1})\}^{\Delta N_{ij}} \right.$$
$$\left. \times \exp((X_i'\beta + z_i)\Delta N_{ij}) \exp\{-(\mu(T_{i,j}) - \mu(T_{i,j-1})) \exp(X_i'\beta + z_i)\}/(\Delta N_{ij}!) \right),$$

and $\phi(.;\Sigma)$ is the density function of a bivariate normal distribution with zero mean and variance-covariance matrix Σ. The expression above can be simplified as follows:

$$f(\boldsymbol{T_i}, \boldsymbol{N_i}|X_i, C_i, u_i, z_i, \beta, \gamma, \mu, \lambda)$$

$$= \frac{\exp(-\Lambda(C_i)e^{X_i'\gamma+u_i})[e^{X_i'\gamma+u_i}]^{K_i}}{K_i!}$$
$$\times \exp((X_i'\beta + z_i)N_{iK_i}) \exp\{-(\mu(T_{i,K_i}) \exp(X_i'\beta + z_i)\}$$
$$\times \prod_{j=1}^{K_i} \left(\lambda(T_{i,j})\{\mu(T_{i,j}) - \mu(T_{i,j-1})\}^{\Delta N_{ij}}/(\Delta N_{ij}!) \right).$$

Then the likelihood function can be obtained by removing the terms that are not related to the parameters, which can be written as

$$L_n(\theta|O_i's) = \left(\prod_{i=1}^{n} \prod_{j=1}^{K_i} \lambda(T_{i,j}) \right) \times \prod_{i=1}^{n} \int\int \exp(-\Lambda(C_i) \exp(X_i'\gamma + u_i))[\exp(X_i'\gamma + u_i)]^{K_i}$$
$$\times \exp((X_i'\beta + z_i)N_{iK_i}) \exp\{-\mu(T_{i,K_i}) \exp(X_i'\beta + z_i)\}$$
$$\times \prod_{j=1}^{K_i} \{\mu(T_{i,j}) - \mu(T_{i,j-1})\}^{\Delta N_{ij}} \phi(u_i, z_i; \Sigma) du_i dz_i.$$

Let the true value of θ be $\theta_0 = (\beta_0, \gamma_0, \Sigma_0, \mu_0, \Lambda_0)$ and $0 < t_1 < \ldots < t_M$ be the ordered distinction time points of $\{T_{i,j}\}$. Then the maximum likelihood estimation of θ_0 can be obtained by maximizing the above likelihood function subject to the monotonicity of $\Lambda(.)$ (e.g., non-decreasing). It seems natural to calculate the maximum likelihood estimators. However, this maximum of the foregoing likelihood function is infinity, because we can always choose some function $\Lambda(.)$ with fixed values at $T'_{i,j}s$ while letting $\lambda(t) = d\Lambda(t)/dt$ go to infinity. Therefore, we relax $\Lambda(.)$ to be right-continuous and allow it to have jumps at $T_{i,j}$. Then the likelihood function can be rewritten as

$$\begin{aligned} L_n(\theta|O'_i s) = & \left(\prod_{i=1}^{n} \prod_{j=1}^{K_i} \Delta\Lambda(T_{i,j}) \right) \times \prod_{i=1}^{n} \int \int \exp(-\Lambda(C_i) \exp(X'_i \gamma + u_i)) [\exp(X'_i \gamma + u_i)]^{K_i} \\ & \times \exp((X'_i \beta + z_i) N_{iK_i}) \exp\{-\mu(T_{i,K_i}) \exp(X'_i \beta + z_i)\} \\ & \times \prod_{j=1}^{K_i} \{\mu(T_{i,j}) - \mu(T_{i,j-1})\}^{\Delta N_{ij}} \phi(u_i, z_i; \Sigma) du_i dz_i. \end{aligned}$$

where $\Delta\Lambda(t) = \Lambda(t) - \Lambda(t^-)$ and $\Lambda(t) = \sum_{m=1}^{M} \Delta\Lambda(t_m) I(t \geq t_m)$. The resulting estimators are referred to as maximum likelihood estimators, denoted by $\hat{\theta}_n = (\hat{\beta}_n, \hat{\gamma}_n, \hat{\Sigma}_n, \hat{\mu}_n, \hat{\Lambda}_n)$.

2.4 Estimation Procedures

2.4.1 Maximum Likelihood Estimation and Algorithm

Some difficulties are involved in this maximization procedure. The first is that the number of estimated parameters $(2M+2p+3)$ is large due to the large M, which is the number of distinct time points $T'_{i,j}s$. Secondly, the integral has no closed form. Monte Carlo approximation of this integral may be applied, but we will implement an EM algorithm instead.

For $m = 1, ..., M$, set $N_m = \sum_{i=1}^{n} \sum_{j=1}^{K_i} I(T_{i,j} = t_m)$, which represents the total number of observations whose observation time is exactly t_m. Based on this setup, the $u_i's$ and $z_i's$ were treated as missing data. Then the complete likelihood function is given by

$$\begin{aligned} L_n(\theta|O_i's, u_i's, z_i's) = & \prod_{m=1}^{M} [\Delta\Lambda(t_m)]^{N_m} \times \prod_{i=1}^{n} \exp(-\Lambda(C_i)\exp(X_i'\gamma + u_i))[\exp(X_i'\gamma + u_i)]^{K_i} \\ & \times \exp((X_i'\beta + z_i)N_{iK_i})\exp\{-\mu(T_{i,K_i})\exp(X_i'\beta + z_i)\} \\ & \times \prod_{j=1}^{K_i} \{\mu(T_{i,j}) - \mu(T_{i,j-1})\}^{\Delta N_{ij}} \phi(u_i, z_i; \Sigma). \end{aligned}$$

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

$$\begin{aligned} l_n(\theta|O_i's, u_i's, z_i's) = & \sum_{i=1}^{n} l_{ni}(\theta|O_i, u_i) \\ = & \sum_{m=1}^{M} N_m \log \Delta\Lambda(t_m) + \sum_{i=1}^{n} \Big\{ -\Lambda(C_i)\exp(X_i'\gamma + u_i) \\ & + K_i(X_i'\gamma + u_i) + N_{iK_i}(X_i'\beta + z_i) - \mu(T_{i,K_i})\exp(X_i'\beta + z_i) \\ & + \log\phi(u_i, z_i; \Sigma) + \sum_{j=1}^{K_i} \Delta N_{ij} \log[\mu(T_{i,j}) - \mu(T_{i,j-1})] \Big\} \end{aligned}$$

Now we consider the maximization of the above expectation, which is the M-step. Suppose that at kth step, the current parameter estimate is $\theta^{(k)} = (\beta^{(k)}, \gamma^{(k)}, \Sigma^{(k)}, \mu^{(k)}, \Lambda^{(k)})$, then the conditional expectation of the "complete" log-likelihood function is

$$l_n(\theta|O_i's, \theta^{(k)}) = l_n^{(1)}(\gamma, \Lambda|O_i's, \theta^{(k)}) + l_n^{(2)}(\beta, \mu|O_i's, \theta^{(k)}) + l_n^{(3)}(\Sigma|O_i's, \theta^{(k)}), \qquad (2.5)$$

where

$$
\begin{aligned}
l_n^{(1)}(\gamma, \Lambda | O_i's, \theta^{(k)}) = &\sum_{m=1}^{M} N_m \log \Delta\Lambda(t_m) \\
&+ \sum_{i=1}^{n} \Big[- \Lambda(C_i) e^{X_i'\gamma} E\{e^{u_i} | O_i, \theta^{(k)}\} \\
&+ K_i(X_i'\gamma + E\{u_i | O_i, \theta^{(k)}\}) \Big],
\end{aligned}
$$

$$
\begin{aligned}
l_n^{(2)}(\beta, \mu | O_i's, \theta^{(k)}) = &\sum_{i=1}^{n} \Big[\sum_{j=1}^{K_i} \Delta N_{ij} \log[\mu(T_{i,j}) - \mu(T_{i,j-1})] \\
&+ N_{iK_i}(X_i'\beta + E\{z_i | O_i, \theta^{(k)}\}) \\
&- \mu(T_{i,K_i}) \exp(X_i'\beta) E\{e^{z_i} | O_i, \theta^{(k)}\} \Big],
\end{aligned}
$$

$$
l_n^{(3)}(\Sigma | O_i's, \theta^{(k)}) = E\{\log \phi(u_i, z_i; \Sigma) | O_i, \theta^{(k)}\}.
$$

Therefore, the maximization of the log-likelihood function can be achieved by maximizing the above three log-likelihood functions separately, and then we can obtain the next estimate $\theta^{(k+1)}$. In particular, the estimate of Λ in step (k+1) is

$$
[\Delta\Lambda(t_m)]^{(k+1)} = N_m \Big[\sum_{i=1}^{n} I(C_i \geq t_m) e^{X_i'\gamma^{(k)}} E(e^{u_i} | O_i, \theta^{(k)}) \Big]^{-1}, \tag{2.6}
$$

for $m = 1, ..., M$ and $\gamma^{(k+1)}$ can be obtained by solving the following equation

$$\sum_{i=1}^{n} X_i \bigg[K_i - [\Lambda(C_i)]^{(k)} \exp(X_i'\gamma) E(e^{u_i} | O_i, \theta^{(k)}) \bigg] = 0. \tag{2.7}$$

2.4.2 A Self-Consistent Algorithm

By maximizing the following log-likelihood function

$$\begin{aligned} l_n^{(2)}(\beta, \mu | O_i's, \theta^{(k)}) = & \sum_{i=1}^{n} \bigg[\sum_{j=1}^{K_i} \Delta N_{ij} \log[\mu(T_{i,j}) - \mu(T_{i,j-1})] \\ & + N_{iK_i}(X_i'\beta + E\{z_i | O_i, \theta^{(k)}\}) \\ & - \mu(T_{i,K_i}) \exp(X_i'\beta) E\{e^{z_i} | O_i, \theta^{(k)}\} \bigg], \end{aligned}$$

one obtains the $(k+1)$th step estimates of μ and β. This algorithm was developed by Wellner and Zhang (2007). Here, we propose a new algorithm to generate a nondecreasing estimator by using the self-consistent algorithm (Hu et al., 2009), which has been demonstrated to be consistent without the Poisson assumption. Let

$$\mu_m = \mu(t_m) - \mu(t_{m-1}),$$

with $t_0 = 0$ and $\mu(0) = 0$. Then it is obvious that the estimator can be written as

$$\hat{\mu}(t) = \sum_{m=1}^{M} \mu_m I(t \geq t_m).$$

The following algorithm will be used to estimate the $(k+1)$th step estimators:

$$\mu_m^{(k+1)} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{K_i} I(t_m \in (T_{i,j-1}, T_{i,j}])\mu_m^{(k)}\Delta N_{ij}/\{\mu^{(k)}(T_{i,j}) - \mu^{(k)}(T_{i,j-1})\}}{\sum_{i=1}^{n} I(T_{i,K_i} \geq t_m)\exp(X_i'\beta^{(k)})E\{e^{z_i}|O_i, \theta^{(k)}\}}, \tag{2.8}$$

and $\beta^{(k+1)}$ can be calculated by solving the following equation

$$\sum_{i=1}^{n} X_i\Big[N_{iK_i} - \mu(T_{i,K_i})^{(k)}\exp(X_i'\beta)E\{e^{z_i}|O_i, \theta^{(k)}\}\Big] = 0. \tag{2.9}$$

2.4.3 Estimation of Σ

The following log-likelihood function

$$l_n^{(3)}(\Sigma|O_i's, \theta^{(k)}) = E\{\log\phi(u_i, z_i; \Sigma)|O_i, \theta^{(k)}\}.$$

can be maximized by obtaining the $(k+1)$th step estimates of Σ as

$$\Sigma^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n} E\Bigg[\begin{pmatrix} z_i^2 & z_i u_i \\ z_i u_i & u_i^2 \end{pmatrix}|O_i, \theta^{(k)}\Bigg]. \tag{2.10}$$

Then the maximum likelihood estimate of θ can be obtained by the following step:

Step 1. Start with initial value $\theta^{(0)}$;

Step 2. At the kth step, we calculate the $(k+1)$th step $\theta^{(k+1)}$ through (2.6 - 2.10); and then iterate this step until the desired convergence achieves.

It is clear that to calculate 2.10, one needs to first calculate $E[z_i^2|O_i, \theta^{(k)}]$, $E[u_i z_i|O_i, \theta^{(k)}]$ and $E[u_i^2|O_i, \theta^{(k)}]$.

We let

$$f_i^*(u_i, z_i; \theta) = \exp(-\Lambda(C_i)\exp(X_i'\gamma + u_i) + K_i u_i + z_i N_{iK_i} - \mu(T_{i,K_i})\exp(X_i'\beta + z_i)),$$

for $i = 1, ..., n$. Generally, for some function $h(u_i, z_i)$,

$$E\{h(u_i, z_i)|O_i, \theta^{(k)}\} = \int\int h(u_i, z_i) f(u_i, z_i|O_i, \theta^{(k)}) du_i dz_i,$$

where

$$f(u_i, z_i|O_i, \theta) = \frac{f_i^*(u_i, z_i; \theta)\phi(u_i, z_i; \Sigma)}{\int\int f_i^*(u_i, z_i; \theta)\phi(u_i, z_i; \Sigma) du_i z_i} \tag{2.11}$$

is the conditional density function of U_i and Z_i given O_i and θ. It is apparent that this integration has no closed form. For this situation, with $\theta = \theta^{(k)}$, let $\{u_{iq}, z_{iq}; i = 1, ..., n; q = 1, ..., Q\}$ be Q i.i.d. samples from $N(0, \Sigma^{(k)})$, where Q is sufficiently large. Then one can approximate $E\{h(u_i, z_i)|O_i, \theta^{(k)}\}$ by

$$\hat{E}\{h(u_i, z_i)|O_i, \theta^{(k)}\} = \frac{\sum_{q=1}^{Q} h(u_{iq}, z_{iq}) f_i^*(u_{iq}, z_{iq}; \theta^{(k)})}{\sum_{q=1}^{Q} f_i^*(u_{iq}, z_{iq}; \theta^{(k)})}. \tag{2.12}$$

Therefore, $E[z_i^2|O_i, \theta^{(k)}]$ can be approximated by

$$\hat{E}\{z_i^2|O_i, \theta^{(k)}\} = \frac{\sum_{q=1}^{Q} z_{iq}^2 f_i^*(u_{iq}, z_{iq}; \theta^{(k)})}{\sum_{q=1}^{Q} f_i^*(u_{iq}, z_{iq}; \theta^{(k)})}, \tag{2.13}$$

$E[u_i z_i | O_i, \theta^{(k)}]$ can be approximated by

$$\hat{E}\{u_i z_i | O_i, \theta^{(k)}\} = \frac{\sum_{q=1}^{Q} u_{iq} z_{iq} f_i^*(u_{iq}, z_{iq}; \theta^{(k)})}{\sum_{q=1}^{Q} f_i^*(u_{iq}, z_{iq}; \theta^{(k)})}, \tag{2.14}$$

and $E[u_i^2 | O_i, \theta^{(k)}]$ can be approximated by

$$\hat{E}\{u_i^2 | O_i, \theta^{(k)}\} = \frac{\sum_{q=1}^{Q} u_{iq}^2 f_i^*(u_{iq}, z_{iq}; \theta^{(k)})}{\sum_{q=1}^{Q} f_i^*(u_{iq}, z_{iq}; \theta^{(k)})}. \tag{2.15}$$

2.5 Discussion

A main feature of the methodology discussed in this chapter is that the correlation between the observation process and the point process that characterizes the occurrence of the events is modeled using a variance-covariance matrix from a bivariate normal distribution of two frailty terms Z_i and U_i. This allows us to conduct the estimation procedure with the joint likelihood. In this case, we can develop a complete likelihood approach without using some two-step or three-step procedures that are commonly implemented in previous literature. Furthermore, the EM algorithm is used for determination of parameter estimators, and a straightforward bootstrap procedure is implemented for the variance estimation. In Chapter 3, we discuss an extension of this approach and the algorithm to the situation in which multiple related types of recurrent events are presented.

Chapter 3: Semiparametric Analysis of Multivariate Panel Count Data

3.1 Introduction

This chapter discusses the extension of the proposed methodology to the statistical analysis of multivariate panel count data, which arise when the study involves multiple related types of recurrent events, and the study subjects are examined only at discrete time points. For multivariate panel count data, the correlation between different types of events may exist and should be taken into consideration. To deal with it, we propose three types of modeling assumptions, as listed in the beginning of Section 3.2. The first type assumes common covariate effects across all event types with an identical correlation between the observation process and the point process that characterizes the occurrence of the events. The second type also assumes common covariate effects across all event types, but allows the correlation between the two processes to be event-type-specific. The third one assumes different covariate effects for multiple event types with an identical correlation between the two processes. Those three multivariate modeling approaches will cover a wide variety of multivariate panel count data in real-world data analysis. Regarding the estimation procedures, we adopt similar joint modeling techniques but the complete likelihood function and the corresponding estimation equations for each of the three multivariate models are quite different, which are described in details in Section 3.3. Finally, Section 3.4 discusses the advantages of our proposed methods as well as possible directions for future research.

3.2 Models and Notations

In the multivariate modeling procedure, we propose three types of modeling structures, which are presented in the following table:

Table 3.1: Features of Multivariate Models

	Type of β	Type of Σ
Multivariate Model A	Constant	Constant
Multivariate Model B	Constant	Event-type-specific
Multivariate Model C	Event-type-specific	Constant

For the proposed model A, it is assumed that given X_i and a subject-specific unobservable frailty z_i, $N_{i(l)}(t), (l = 1, ..., L)$, which is the total number of type l events that have occurred up to time t, follows a nonhomogeneous Poisson process with the mean function

$$\mu_{i(l)}(t) = \mu_{0(l)}(t) \exp\{X_i'\beta + z_i\}. \tag{3.1}$$

Given X_i and a subject-specific unobservable frailty u_i, $H_{i(l)}(t)$, which is the total number of observation times for type l events that have occurred up to time t, follows a nonhomogeneous Poisson process with the intensity function

$$\lambda_{i(l)}(t) = \lambda_{0(l)}(t) \exp\{X_i'\gamma + u_i\}. \tag{3.2}$$

To model the correlation between the two processes $N_{i(l)}(t)$ and $H_{i(l)}(t)$, it is assumed that (z_i, u_i) follows a bivariate normal distribution with mean (0, 0) and a variance-covariance matrix that is the same across all of the event types

$$\Sigma = \begin{pmatrix} \sigma_z^2 & \sigma_{zu} \\ \sigma_{zu} & \sigma_u^2 \end{pmatrix} \tag{3.3}$$

For the proposed model B, it is assumed that given X_i and a subject-specific and event-type-specific unobservable frailty $z_{i(l)}$, $N_{i(l)}(t), (l = 1, ..., L)$, which is the total number of type l events that have occurred up to time t, follows a nonhomogeneous Poisson process with the mean function

$$\mu_{i(l)}(t) = \mu_{0(l)}(t) \exp\{X_i'\beta + z_{i(l)}\}. \tag{3.4}$$

Given X_i and a subject-specific and event-type-specific unobservable frailty $u_{i(l)}$, $H_{i(l)}(t)$, which is the total number of observation times for type l events that have occurred up to time t, follows a nonhomogeneous Poisson process with the intensity function

$$\lambda_{i(l)}(t) = \lambda_{0(l)}(t) \exp\{X_i'\gamma + u_{i(l)}\}. \tag{3.5}$$

To model the correlation between the two processes $N_{i(l)}(t)$ and $H_{i(l)}(t)$, it is assumed that ($z_{i(l)}$, $u_{i(l)}$) follows a bivariate normal distribution with mean (0, 0) and a variance-covariance matrix that is event-type-specific

$$\Sigma_{(l)} = \begin{pmatrix} [\sigma_{z(l)}]^2 & \sigma_{zu(l)} \\ \sigma_{zu(l)} & [\sigma_{u(l)}]^2 \end{pmatrix} \tag{3.6}$$

For the proposed model C, it is assumed that given X_i and a subject-specific unobservable frailty z_i, $N_{i(l)}(t), (l = 1, ..., L)$, which is the total number of type l events that have occurred up to time t, follows a nonhomogeneous Poisson process with the mean function

$$\mu_{i(l)}(t) = \mu_{0(l)}(t) \exp\{X_i'\beta_{(l)} + z_i\}, \tag{3.7}$$

where $\beta_{(l)}$ is a vector of unknown regression parameters specifically for event type l.

For the observation process, given X_i and a subject-specific unobservable frailty $u_{i(l)}$, $H_{i(l)}(t)$, which is the total number of observation times for type l events that have occurred up to time t, follows a nonhomogeneous Poisson process with the intensity function

$$\lambda_{i(l)}(t) = \lambda_{0(l)}(t) \exp\{X_i'\gamma_{(l)} + u_i\}, \tag{3.8}$$

where $\gamma_{(l)}$ is a vector of unknown regression parameters specifically for event type l.

To model the correlation between the two processes $N_{i(l)}(t)$ and $H_{i(l)}(t)$, it is assumed that (z_i, u_i) follows a bivariate normal distribution with mean (0, 0) and a variance-covariance matrix that is the same across all of the event types

$$\Sigma = \begin{pmatrix} \sigma_z^2 & \sigma_{zu} \\ \sigma_{zu} & \sigma_u^2 \end{pmatrix} \tag{3.9}$$

3.3 Estimation Procedures

3.3.1 Multivariate Model A

For the proposed multivariate model A, we first set $T_{i(l),0} = 0$, $N_{i(l),K_i} = N_{i(l)}(T_{i(l),K_i})$ and $\Delta N_{i(l),j} = N_{i(l)}(T_{i(l),j}) - N_{i(l)}(T_{i(l),j-1})$ for $j = 1, ..., K_{i(l)}$, $i = 1, ...n$ and $l = 1, ..., L$. Then we assume that the observations from the n subjects are i.i.d. copies, given the observed data

$$O_{i(l)} = \{X_i, \boldsymbol{T_{i(l)}}, \boldsymbol{N_{i(l)}}, C_{i(l)}\},\ i = 1, ..., n,\ l = 1, ..., L$$

with $\boldsymbol{T_{i(l)}} = \{T_{i(l),0}, T_{i(l),1}, ..., T_{i(l),K_i}\}$ and $\boldsymbol{N_{i(l)}} = \{N_{i(l)}(T_{i(l),1}), ..., N_{i(l)}(T_{i(l),K_i})\}$,

The likelihood function of $\theta = (\beta, \gamma, \Sigma, \mu, \Lambda)$ is proportional to

$$\prod_{l=1}^{L}\prod_{i=1}^{n}\int\int f(\boldsymbol{T_{i(l)}}, \boldsymbol{N_{i(l)}}|X_i, C_{i(l)}, u_{i(l)}, z_{i(l)}, \beta, \gamma, \mu_{(l)}, \Lambda_{(l)})\phi(u_{i(l)}, z_{i(l)}; \Sigma)du_{i(l)}dz_{i(l)}, \quad (3.10)$$

where

$f(\boldsymbol{T_{i(l)}}, \boldsymbol{N_{i(l)}}|X_i, C_{i(l)}, u_{i(l)}, z_{i(l)}, \beta, \gamma, \mu_{(l)}, \Lambda_{(l)})$

$$\begin{aligned}
&= \frac{\exp(-\Lambda_{(l)}(C_{i(l)})\exp(X_i'\gamma + u_{i(l)}))[\Lambda_{(l)}(C_{i(l)})\exp(X_i'\gamma + u_{i(l)})]^{K_{i(l)}}}{K_{i(l)}!} \\
&\times \prod_{j=1}^{K_{i(l)}}(\frac{\Lambda_{(l)}(T_{i(l),j})}{\Lambda_{(l)}(C_{i(l)})}\{\mu_{(l)}(T_{i,j(l)}) - \mu_{(l)}(T_{i(l),j-1})\}^{\Delta N_{i(l),j}} \\
&\times \exp((X_i'\beta + z_{i(l)})\Delta N_{i(l),j}) \\
&\times \exp\{-(\mu_{(l)}(T_{i,j(l)}) - \mu_{(l)}(T_{i(l),j-1}))\exp(X_i'\beta + z_{i(l)})\}/(\Delta N_{i(l),j}!)),
\end{aligned}$$

and $\phi(.;\Sigma)$ is the density function of bivariate normal distribution with zero mean and variance-covariance matrix Σ. Noted that for the expression above, $\Lambda_{(l)}(C_{i(l)})$ can be cancelled out and the remaining parts can be simplified as follows:

$$
\begin{aligned}
&f(\boldsymbol{T_{i(l)}}, \boldsymbol{N_{i(l)}}|X_i, C_{i(l)}, u_{i(l)}, z_{i(l)}, \beta, \gamma, \mu_{(l)}, \Lambda_{(l)}) \\
&\qquad = \frac{\exp(-\Lambda_{(l)}(C_{i(l)})\exp(X_i'\gamma + u_{i(l)}))[\exp(X_i'\gamma + u_{i(l)})]^{K_{i(l)}}}{K_{i(l)}!} \\
&\qquad\quad \times \exp((X_i'\beta + z_{i(l)})N_{i(l),K_i})\exp\{-(\mu(T_{i(l),K_i})\exp(X_i'\beta + z_{i(l)})\} \\
&\qquad\quad \times \prod_{j=1}^{K_{i(l)}}\left(\Lambda_{(l)}(T_{i(l),j})\{\mu_{(l)}(T_{i,j(l)}) - \mu_{(l)}(T_{i(l),j-1})\}^{\Delta N_{i(l),j}}/(\Delta N_{i(l),j}!)\right).
\end{aligned}
$$

The likelihood function can be obtained by removing the terms that are not related to the parameters, which can be written as

$$
\begin{aligned}
L_n(\theta|O_i's) = &\left(\prod_{l=1}^{L}\prod_{i=1}^{n}\prod_{j=1}^{K_{i(l)}}\Lambda_{(l)}(T_{i(l),j})\right) \\
&\times \prod_{l=1}^{L}\prod_{i=1}^{n}\int\int \exp(-\Lambda_{(l)}(C_{i(l)})\exp(X_i'\gamma + u_{i(l)}))[\exp(X_i'\gamma + u_{i(l)})]^{K_{i(l)}} \\
&\times \exp((X_i'\beta + z_{i(l)})N_{i(l),K_i})\exp\{-\mu_{(l)}(T_{i(l),K_i})\exp(X_i'\beta + z_{i(l)})\} \\
&\times \prod_{j=1}^{K_{i(l)}}\{\mu_{(l)}(T_{i,j(l)}) - \mu_{(l)}(T_{i(l),j-1})\}^{\Delta N_{i(l),j}}\phi(u_{i(l)}, z_{i(l)};\Sigma)du_{i(l)}dz_{i(l)}.
\end{aligned}
$$

Let the true value of $\theta_{(l)}$, which is the set of true parameters of event type l, be $\theta_{0(l)} = (\beta_0, \gamma_0, \Sigma_0, \mu_{0(l)}, \Lambda_{0(l)})$ and $0 < t_1 < \ldots < t_{M_{(l)}}$ be the ordered distinction time points of $\{T_{i(l),j}\}$. Then the maximum likelihood estimation of $\theta_{0(l)}$ can be obtained by maximizing the above likelihood function subject to the monotonicity of $\Lambda_{(l)}(.)$ (e.g. non-decreasing). Following the discussion in chapter 2, the maximum of the foregoing likelihood function is infinity, because we can always choose some function $\Lambda_{(l)}(.)$ with fixed values at the $T_{i(l),j}$ while letting $\Lambda_{(l)}(t) = d\Lambda_{(l)}(t)/dt$ go to infinity. Thus we relax $\Lambda_{(l)}(.)$ to be right-continuous and allow it to have jumps at $T_{i(l),j}$. Then the likelihood function can be rewritten as

$$
\begin{aligned}
L_n(\theta|O_i's) = &\left(\prod_{l=1}^{L}\prod_{i=1}^{n}\prod_{j=1}^{K_{i(l)}} \Delta\Lambda_{(l)}(T_{i(l),j})\right) \\
&\times \prod_{l=1}^{L}\prod_{i=1}^{n}\int\int \exp(-\Lambda_{(l)}(C_{i(l)})\exp(X_i'\gamma + u_{i(l)}))[\exp(X_i'\gamma + u_{i(l)})]^{K_{i(l)}} \\
&\times \exp((X_i'\beta + z_{i(l)})N_{i(l),K_i})\exp\{-\mu(T_{i(l),K_i})\exp(X_i'\beta + z_{i(l)})\} \\
&\times \prod_{j=1}^{K_{i(l)}}\{\mu_{(l)}(T_{i,j(l)}) - \mu_{(l)}(T_{i(l),j-1})\}^{\Delta N_{i(l),j}}\phi(u_{i(l)}, z_{i(l)}; \Sigma)du_{i(l)}dz_{i(l)}.
\end{aligned}
$$

where $\Delta\Lambda_{(l)}(t) = \Lambda_{(l)}(t) - \Lambda_{(l)}(t-)$ and $\Lambda_{(l)}(t) = \sum_{m=1}^{M_{(l)}} \Delta\Lambda_{(l)}(t_{m(l)})I(t \geq t_{m(l)})$. The resulting estimators, are referred to as maximum likelihood estimators, denoted by

$\hat{\theta}_{n(l)} = (\hat{\beta}_n, \hat{\gamma}_n, \hat{\Sigma}_n, \hat{\mu}_{n(l)}, \hat{\Lambda}_{n(l)})$, $(l = 1, \ldots, L)$.

Following a similar algorithm as in Chapter 2 and by incorporating multivariate components, the complete likelihood function for multivariate model A is given by

$$
\begin{aligned}
L_n(\theta|O_i's, u_i's, z_i's) = &\prod_{l=1}^{L}\prod_{m=1}^{M_{(l)}}[\Delta\Lambda_{(l)}(t_{m(l)})]^{N_{m(l)}} \\
&\times \prod_{l=1}^{L}\prod_{i=1}^{n}\exp(-\Lambda_{(l)}(C_{i(l)})\exp(X_i'\gamma + u_{i(l)}))[\exp(X_i'\gamma + u_{i(l)})]^{K_{i(l)}} \\
&\times \exp((X_i'\beta + z_{i(l)})N_{i(l),K_i})\exp\{-\mu_{(l)}(T_{i(l),K_i})\exp(X_i'\beta + z_{i(l)})\} \\
&\times \prod_{j=1}^{K_{i(l)}}\{\mu_{(l)}(T_{i,j(l)}) - \mu_{(l)}(T_{i(l),j-1})\}^{\Delta N_{i(l),j}}\phi(u_{i(l)}, z_{i(l)}; \Sigma),
\end{aligned}
$$

where $M_{(l)}, l = 1, ..., L$ is the number of distinct observation time points for event type l. To implement the EM algorithm, we first compute the conditional expectation of the log likelihood function given the current estimate of θ and the observed data O (E-step). To this end, for multivariate model A, note that the log likelihood function can be written as

$$
\begin{aligned}
l_n(\theta|O_i's, u_i's, z_i's) &= \sum_{i=1}^{n} l_{ni}(\theta|O_i, u_i) \\
&= \sum_{l=1}^{L}\sum_{m=1}^{M_{(l)}} N_{m(l)} log\Delta\Lambda_{(l)}(t_{m(l)}) + \sum_{l=1}^{L}\sum_{i=1}^{n}\Big\{ - \Lambda_{(l)}(C_{i(l)})e^{X_i'\gamma+u_{i(l)}} \\
&\quad + K_{i(l)}(X_i'\gamma + u_{i(l)}) + N_{iK_{i(l)}}(X_i'\beta + z_{i(l)}) - \mu_{(l)}(T_{i(l),K_i})\exp(X_i'\beta + z_{i(l)}) \\
&\quad + log\phi(u_{i(l)}, z_{i(l)}; \Sigma) + \sum_{j=1}^{K_{i(l)}} \Delta N_{i(l),j} log[\mu_{(l)}(T_{i,j(l)}) - \mu_{(l)}(T_{i(l),j-1})]\Big\}.
\end{aligned}
$$

Now we consider the maximization of the above expectation, which is the M-step. For event type l, suppose that at (k)th step, the current parameter is $\theta_{(l)}^{(k)} = (\beta^{(k)}, \gamma^{(k)}, \Sigma^{(k)}, \mu_{(l)}^{(k)}, \Lambda_{(l)}^{(k)})$ for multivariate model A. Then the conditional expectation of the "complete" log-likelihood function is

$$
l_n(\theta|O_i's, \theta^{(k)}) = l_n^{(1)}(\gamma, \Lambda|O_i's, \theta^{(k)}) + l_n^{(2)}(\beta, \mu|O_i's, \theta^{(k)}) + l_n^{(3)}(\Sigma|O_i's, \theta^{(k)}), \tag{3.11}
$$

where

$$
\begin{aligned}
l_n^{(1)}(\gamma, \Lambda|O_i's, \theta^{(k)}) &= \sum_{l=1}^{L}\sum_{m=1}^{M_{(l)}} N_{m(l)} log\Delta\Lambda_{(l)}(t_{m(l)}) \\
&\quad + \sum_{l=1}^{L}\sum_{i=1}^{n}\Big[- \Lambda_{(l)}(C_{i(l)})e^{X_i'\gamma} E\{e^{u_{i(l)}}|O_i, \theta^{(k)}\} \\
&\quad + K_{i(l)}(X_i'\gamma + E\{u_{i(l)}|O_i, \theta^{(k)}\})\Big],
\end{aligned}
$$

$$l_n^{(2)}(\beta, \mu|O_i's, \theta^{(k)}) = \sum_{l=1}^{L}\sum_{i=1}^{n}\Bigg[\sum_{j=1}^{K_{i(l)}} \Delta N_{i(l),j} log[\mu_{(l)}(T_{i,j(l)}) - \mu_{(l)}(T_{i(l),j-1})]$$

$$+ N_{iK_{i(l)}}(X_i'\beta + E\{z_{i(l)}|O_i, \theta^{(k)}\})$$

$$- \mu_{(l)}(T_{i(l),K_i}) \exp(X_i'\beta) E\{e^{z_{i(l)}}|O_i, \theta^{(k)}\}\Bigg],$$

$$l_n^{(3)}(\Sigma|O_i's, \theta^{(k)}) = E\{log\phi(u_{i(l)}, z_{i(l)}; \Sigma)|O_i, \theta^{(k)}\}.$$

Therefore, the maximization of the log-likelihood function can be achieved by maximizing the above three log-likelihood functions separately, and then we can obtain the next estimate $\theta^{(k+1)}$. By maximizing the following log-likelihood function $l_n^{(1)}$

$$l_n^{(1)}(\gamma, \Lambda|O_i's, \theta^{(k)}) = \sum_{l=1}^{L}\sum_{m=1}^{M_{(l)}} N_{m(l)} log\Delta\Lambda_{(l)}(t_{m(l)})$$

$$+ \sum_{l=1}^{L}\sum_{i=1}^{n}\Bigg[- \Lambda_{(l)}(C_{i(l)}) e^{X_i'\gamma} E\{e^{u_{i(l)}}|O_i, \theta^{(k)}\}$$

$$+ K_{i(l)}(X_i'\gamma + E\{u_{i(l)}|O_i, \theta^{(k)}\})\Bigg],$$

one obtains the $(k+1)$th step estimates of Λ and γ for event type l,

$$[\Delta\Lambda_{(l)}(t_{m(l)})]^{(k+1)} = N_{m(l)}\Bigg[\sum_{i=1}^{n} I(C_{i(l)} \geq t_{m(l)}) e^{X_i'\gamma^{(k)}} E(e^{u_{i(l)}}|O_i, \theta^{(k)})\Bigg]^{-1}, \tag{3.12}$$

for $m = 1, ..., M_{(l)}$ and $\gamma^{(k+1)}$ by solving the following equation

$$\sum_{l=1}^{L}\sum_{i=1}^{n} X_i\Bigg[K_{i(l)} - [\Lambda_{(l)}^{(k)}(C_{i(l)})] \exp(X_i'\gamma) E(e^{u_{i(l)}}|O_i, \theta^{(k)})\Bigg] = 0. \tag{3.13}$$

By maximizing the following log-likelihood function

$$
\begin{aligned}
l_n^{(2)}(\beta, \mu | O_i's, \theta^{(k)}) = &\sum_{l=1}^{L} \sum_{i=1}^{n} \Bigg[\sum_{j=1}^{K_{i(l)}} \Delta N_{i(l),j} log[\mu_{(l)}(T_{i,j(l)}) - \mu_{(l)}(T_{i(l),j-1})] \\
&+ N_{iK_{i(l)}}(X_i'\beta + E\{z_{i(l)}|O_i, \theta^{(k)}\}) \\
&- \mu_{(l)}(T_{i(l),K_i}) \exp(X_i'\beta) E\{e^{z_{i(l)}}|O_i, \theta^{(k)}\} \Bigg],
\end{aligned}
$$

one obtains the $(k+1)$th step estimates of μ and β. Following the self-consistent algorithm discussed in Chapter 2, the following algorithm can be applied to estimate the $(k+1)$th step estimators for event type l:

$$
\mu_{m(l)}^{(k+1)} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{K_{i(l)}} I(t_{m(l)} \in (T_{i(l),j-1}, T_{i(l),j}]) \mu_{m(l)}^{(k)} \Delta N_{i(l),j} / \{\mu_{(l)}^{(k)}(T_{i(l),j}) - \mu_{(l)}^{(k)}(T_{i(l),j-1})\}}{\sum_{i=1}^{n} I(T_{i(l),K_i} \geq t_{m(l)}) \exp(X_i'\beta^{(k)}) E\{e^{z_{i(l)}}|O_i, \theta^{(k)}\}}, \tag{3.14}
$$

and $\beta^{(k+1)}$ by solving the following equation:

$$
\sum_{l=1}^{L} \sum_{i=1}^{n} X_i \left[N_{i(l),K_i} - \mu_{(l)}^{(k)}(T_{i(l),K_i}) \exp(X_i'\beta) E\{e^{z_{i(l)}}|O_i, \theta^{(k)}\} \right] = 0. \tag{3.15}
$$

3.3.2 Multivariate Model B

For multivariate model B, the joint modeling setup is similar to multivariate model A except for the covariance component. The likelihood function can be simplified as

$$\begin{aligned} L_n(\theta|O_i's) = &\left(\prod_{l=1}^{L}\prod_{i=1}^{n}\prod_{j=1}^{K_{i(l)}} \Delta\Lambda_{(l)}(T_{i(l),j})\right) \\ &\times \prod_{l=1}^{L}\prod_{i=1}^{n}\int\int \exp(-\Lambda_{(l)}(C_{i(l)})\exp(X_i'\gamma + u_{i(l)}))[\exp(X_i'\gamma + u_{i(l)})]^{K_{i(l)}} \\ &\times \exp((X_i'\beta + z_{i(l)})N_{i(l),K_i})\exp\{-\mu(T_{i(l),K_i})\exp(X_i'\beta + z_{i(l)})\} \\ &\times \prod_{j=1}^{K_{i(l)}}\{\mu_{(l)}(T_{i,j(l)}) - \mu_{(l)}(T_{i(l),j-1})\}^{\Delta N_{i(l),j}}\phi(u_{i(l)}, z_{i(l)}; \Sigma_{(l)})du_{i(l)}dz_{i(l)}. \end{aligned}$$

where $\Delta\Lambda_{(l)}(t) = \Lambda_{(l)}(t) - \Lambda_{(l)}(t-)$ and $\Lambda_{(l)}(t) = \sum_{m=1}^{M_{(l)}} \Delta\Lambda_{(l)}(t_{m(l)})I(t \geq t_{m(l)})$. The resulting estimators are referred to as maximum likelihood estimators, denoted by $\hat{\theta}_{n(l)} = (\hat{\beta}_n, \hat{\gamma}_n, \hat{\Sigma}_{n(l)}, \hat{\mu}_{n(l)}, \hat{\Lambda}_{n(l)})$, $(l = 1, \ldots, L)$.

Then the complete likelihood function for multivariate model B is given by

$$\begin{aligned} L_n(\theta|O_i's, u_i's, z_i's) = &\prod_{l=1}^{L}\prod_{m=1}^{M_{(l)}}[\Delta\Lambda_{(l)}(t_{m(l)})]^{N_{m(l)}} \\ &\times \prod_{l=1}^{L}\prod_{i=1}^{n}\exp(-\Lambda_{(l)}(C_{i(l)})\exp(X_i'\gamma + u_{i(l)}))[\exp(X_i'\gamma + u_{i(l)})]^{K_{i(l)}} \\ &\times \exp((X_i'\beta + z_{i(l)})N_{i(l),K_i})\exp\{-\mu_{(l)}(T_{i(l),K_i})\exp(X_i'\beta + z_{i(l)})\} \\ &\times \prod_{j=1}^{K_{i(l)}}\{\mu_{(l)}(T_{i,j(l)}) - \mu_{(l)}(T_{i(l),j-1})\}^{\Delta N_{i(l),j}}\phi(u_{i(l)}, z_{i(l)}; \Sigma_{(l)}). \end{aligned}$$

To compute the conditional expectation of the log likelihood function given the current estimate of θ and the observed data O, note that the log likelihood function can be written as

$$
\begin{aligned}
l_n(\theta|O_i's, u_i's, z_i's) &= \sum_{i=1}^{n} l_{ni}(\theta|O_i, u_i) \\
&= \sum_{l=1}^{L}\sum_{m=1}^{M_{(l)}} N_{m(l)} log\Delta\Lambda_{(l)}(t_{m(l)}) + \sum_{l=1}^{L}\sum_{i=1}^{n}\Big\{ -\Lambda_{(l)}(C_{i(l)})e^{X_i'\gamma+u_{i(l)}} \\
&\quad + K_{i(l)}(X_i'\gamma + u_{i(l)}) + N_{iK_{i(l)}}(X_i'\beta + z_{i(l)}) - \mu_{(l)}(T_{i(l),K_i})\exp(X_i'\beta + z_{i(l)}) \\
&\quad + log\phi(u_{i(l)}, z_{i(l)}; \Sigma_{(l)}) + \sum_{j=1}^{K_{i(l)}} \Delta N_{i(l),j} log[\mu_{(l)}(T_{i,j(l)}) - \mu_{(l)}(T_{i(l),j-1})]\Big\}
\end{aligned}
$$

For event type l, the conditional expectation of the “complete” log-likelihood function for the current parameter $\theta_{(l)}^{(k)} = (\beta^{(k)}, \gamma^{(k)}, \Sigma_{(l)}^{(k)}, \mu_{(l)}^{(k)}, \Lambda_{(l)}^{(k)})$ is

$$
l_n(\theta|O_i's, \theta^{(k)}) = l_n^{(1)}(\gamma, \Lambda|O_i's, \theta^{(k)}) + l_n^{(2)}(\beta, \mu|O_i's, \theta^{(k)}) + l_n^{(3)}(\Sigma|O_i's, \theta^{(k)}), \tag{3.16}
$$

where

$$
\begin{aligned}
l_n^{(1)}(\gamma, \Lambda|O_i's, \theta^{(k)}) &= \sum_{l=1}^{L}\sum_{m=1}^{M_{(l)}} N_{m(l)} log\Delta\Lambda_{(l)}(t_{m(l)}) \\
&\quad + \sum_{l=1}^{L}\sum_{i=1}^{n}\Big[-\Lambda_{(l)}(C_{i(l)})e^{X_i'\gamma} E\{e^{u_{i(l)}}|O_i, \theta^{(k)}\} \\
&\quad + K_{i(l)}(X_i'\gamma + E\{u_{i(l)}|O_i, \theta^{(k)}\})\Big],
\end{aligned}
$$

$$
\begin{aligned}
l_n^{(2)}(\beta, \mu | O_i' s, \theta^{(k)}) = \sum_{l=1}^{L} \sum_{i=1}^{n} \Bigg[& \sum_{j=1}^{K_{i(l)}} \Delta N_{i(l),j} log[\mu_{(l)}(T_{i,j(l)}) - \mu_{(l)}(T_{i(l),j-1})] \\
& + N_{iK_{i(l)}}(X_i'\beta + E\{z_{i(l)} | O_i, \theta^{(k)}\}) \\
& - \mu_{(l)}(T_{i(l),K_i}) \exp(X_i'\beta) E\{e^{z_{i(l)}} | O_i, \theta^{(k)}\} \Bigg],
\end{aligned}
$$

$$
l_n^{(3)}(\Sigma | O_i' s, \theta^{(k)}) = E\{log\phi(u_{i(l)}, z_{i(l)}; \Sigma_{(l)}) | O_i, \theta^{(k)}\}.
$$

Therefore, the maximization of the log-likelihood function can be achieved by maximizing the above three log-likelihood functions separately and then one can obtain the next estimate $\theta^{(k+1)}$. By maximizing the following log-likelihood function $l_n^{(1)}$

$$
\begin{aligned}
l_n^{(1)}(\gamma, \Lambda | O_i' s, \theta^{(k)}) = & \sum_{l=1}^{L} \sum_{m=1}^{M_{(l)}} N_{m(l)} log \Delta\Lambda_{(l)}(t_{m(l)}) \\
& + \sum_{l=1}^{L} \sum_{i=1}^{n} \Bigg[- \Lambda_{(l)}(C_{i(l)}) e^{X_i'\gamma} E\{e^{u_{i(l)}} | O_i, \theta^{(k)}\} \\
& + K_{i(l)}(X_i'\gamma + E\{u_{i(l)} | O_i, \theta^{(k)}\}) \Bigg],
\end{aligned}
$$

one obtains the $(k+1)$th step estimates of Λ and γ for event type l,

$$
[\Delta\Lambda_{(l)}(t_{m(l)})]^{(k+1)} = N_{m(l)} \Bigg[\sum_{i=1}^{n} I(C_{i(l)} \geq t_{m(l)}) e^{X_i'\gamma^{(k)}} E(e^{u_{i(l)}} | O_i, \theta^{(k)}) \Bigg]^{-1}, \qquad (3.17)
$$

for $m = 1, ..., M_{(l)}$ and $\gamma^{(k+1)}$ by solving the following equation

$$
\sum_{l=1}^{L} \sum_{i=1}^{n} X_i \Bigg[K_{i(l)} - [\Lambda_{(l)}^{(k)}(C_{i(l)})] \exp(X_i'\gamma) E(e^{u_{i(l)}} | O_i, \theta^{(k)}) \Bigg] = 0. \qquad (3.18)
$$

By maximizing the following log-likelihood function

$$
\begin{aligned}
l_n^{(2)}(\beta, \mu | O_i's, \theta^{(k)}) = & \sum_{l=1}^{L} \sum_{i=1}^{n} \Bigg[\sum_{j=1}^{K_{i(l)}} \Delta N_{i(l),j} log[\mu_{(l)}(T_{i,j(l)}) - \mu_{(l)}(T_{i(l),j-1})] \\
& + N_{iK_{i(l)}}(X_i'\beta + E\{z_{i(l)} | O_i, \theta^{(k)}\}) \\
& - \mu_{(l)}(T_{i(l),K_i}) \exp(X_i'\beta) E\{e^{z_{i(l)}} | O_i, \theta^{(k)}\} \Bigg],
\end{aligned}
$$

one obtains the $(k+1)$th step estimates of μ and β. The self-consistent algorithm discussed in Chapter 2 suggests the following $(k+1)$th step estimates of μ_m for event type l:

$$
\mu_{m(l)}^{(k+1)} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{K_{i(l)}} I(t_{m(l)} \in (T_{i(l),j-1}, T_{i(l),j}]) \mu_{m(l)}^{(k)} \Delta N_{i(l),j} / \{\mu_{(l)}^{(k)}(T_{i(l),j}) - \mu_{(l)}^{(k)}(T_{i(l),j-1})\}}{\sum_{i=1}^{n} I(T_{i(l),K_i} \geq t_{m(l)}) \exp(X_i'\beta^{(k)}) E\{e^{z_{i(l)}} | O_i, \theta^{(k)}\}}, \tag{3.19}
$$

and $\beta^{(k+1)}$ by solving the following equation:

$$
\sum_{l=1}^{L} \sum_{i=1}^{n} X_i \Bigg[N_{i(l),K_i} - \mu_{(l)}^{(k)}(T_{i(l),K_i}) \exp(X_i'\beta) E\{e^{z_{i(l)}} | O_i, \theta^{(k)}\} \Bigg] = 0. \tag{3.20}
$$

3.3.3 Multivariate Model C

For multivariate model C, the covariance component is similar to multivariate model A, and the likelihood function can be simplified as

$$
\begin{aligned}
L_n(\theta|O_i's) = & \left(\prod_{l=1}^{L}\prod_{i=1}^{n}\prod_{j=1}^{K_{i(l)}} \Delta\Lambda_{(l)}(T_{i(l),j})\right) \\
& \times \prod_{l=1}^{L}\prod_{i=1}^{n}\int\int \exp(-\Lambda_{(l)}(C_{i(l)})\exp(X_i'\gamma_{(l)} + u_{i(l)})[\exp(X_i'\gamma_{(l)} + u_{i(l)})]^{K_{i(l)}} \\
& \times \exp((X_i'\beta_{(l)} + z_{i(l)})N_{i(l),K_i})\exp\{-\mu(T_{i(l),K_i})\exp(X_i'\beta_{(l)} + z_{i(l)})\} \\
& \times \prod_{j=1}^{K_{i(l)}}\{\mu_{(l)}(T_{i,j(l)}) - \mu_{(l)}(T_{i(l),j-1})\}^{\Delta N_{i(l),j}}\phi(u_{i(l)}, z_{i(l)}; \Sigma)du_{i(l)}dz_{i(l)}.
\end{aligned}
$$

where $\Delta\Lambda_{(l)}(t) = \Lambda_{(l)}(t) - \Lambda_{(l)}(t-)$ and $\Lambda_{(l)}(t) = \sum_{m=1}^{M_{(l)}} \Delta\Lambda_{(l)}(t_{m(l)})I(t \geq t_{m(l)})$. The resulting estimators are referred as maximum likelihood estimators, denoted by $\hat{\theta}_{n(l)} = (\hat{\beta}_{n(l)}, \hat{\gamma}_{n(l)}, \hat{\Sigma}_n, \hat{\mu}_{n(l)}, \hat{\Lambda}_{n(l)})$, $(l = 1, \ldots, L)$.

For multivariate model C, the conditional expectation of the "complete" log-likelihood function is maximized based on each event type. One can obtain the $(k+1)$th step estimates of Λ for event type l as follows

$$
[\Delta\Lambda_{(l)}(t_{m(l)})]^{(k+1)} = N_{m(l)}\left[\sum_{i=1}^{n} I(C_{i(l)} \geq t_{m(l)})e^{X_i'\gamma_{(l)}^{(k)}}E(e^{u_{i(l)}}|O_i, \theta^{(k)})\right]^{-1}, \tag{3.21}
$$

for $m = 1, ..., M_{(l)}$. Then $\gamma_{(1)}^{(k+1)}, \gamma_{(2)}^{(k+1)}, \ldots, \gamma_{(L)}^{(k+1)}$ can be obtained by solving the following system of equations

$$\begin{aligned}
\sum_{i=1}^{n} X_i\Big[K_{i(1)} - [\Lambda^{(k)}_{(1)}(C_{i(1)})]\exp(X_i'\gamma_{(1)})E(e^{u_{i(1)}}|O_i,\theta^{(k)})\Big] &= 0 \\
\sum_{i=1}^{n} X_i\Big[K_{i(2)} - [\Lambda^{(k)}_{(2)}(C_{i(2)})]\exp(X_i'\gamma_{(2)})E(e^{u_{i(2)}}|O_i,\theta^{(k)})\Big] &= 0 \\
&\vdots \\
\sum_{i=1}^{n} X_i\Big[K_{i(L)} - [\Lambda^{(k)}_{(L)}(C_{i(L)})]\exp(X_i'\gamma_{(L)})E(e^{u_{i(L)}}|O_i,\theta^{(k)})\Big] &= 0
\end{aligned} \tag{3.22}$$

The $(k+1)$th step estimates of μ can also be obtained following the self-consistent algorithm discussed in Chapter 2, which yields the following $(k+1)$th step estimate of μ_m for event type l:

$$\mu^{(k+1)}_{m(l)} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{K_{i(l)}} I(t_{m(l)} \in (T_{i(l),j-1}, T_{i(l),j}])\mu^{(k)}_{m(l)}\Delta N_{i(l),j}/\{\mu^{(k)}_{(l)}(T_{i(l),j}) - \mu^{(k)}_{(l)}(T_{i(l),j-1})\}}{\sum_{i=1}^{n} I(T_{i(l),K_i} \geq t_{m(l)})\exp(X_i'\beta^{(k)}_{(l)})E\{e^{z_{i(l)}}|O_i,\theta^{(k)}\}}, \tag{3.23}$$

and $\beta^{(k+1)}_{(1)}, \beta^{(k+1)}_{(2)}, \ldots, \beta^{(k+1)}_{(L)}$ can be obtained by solving the following system of equations

$$\begin{aligned}
\sum_{i=1}^{n} X_i\Big[N_{i(1),K_i} - \mu^{(k)}_{(1)}(T_{i(1),K_i})\exp(X_i'\beta_{(1)})E\{e^{z_{i(1)}}|O_i,\theta^{(k)}\}\Big] &= 0 \\
\sum_{i=1}^{n} X_i\Big[N_{i(2),K_i} - \mu^{(k)}_{(2)}(T_{i(2),K_i})\exp(X_i'\beta_{(2)})E\{e^{z_{i(2)}}|O_i,\theta^{(k)}\}\Big] &= 0 \\
&\vdots \\
\sum_{i=1}^{n} X_i\Big[N_{i(L),K_i} - \mu^{(k)}_{(L)}(T_{i(L),K_i})\exp(X_i'\beta_{(L)})E\{e^{z_{i(L)}}|O_i,\theta^{(k)}\}\Big] &= 0
\end{aligned} \tag{3.24}$$

3.3.4 Estimation of Σ

For multivariate model A, the following log-likelihood function

$$l_n^{(3)}(\Sigma|O_i's, \theta^{(k)}) = E\{log\phi(u_{i(l)}, z_{i(l)}; \Sigma)|O_i, \theta^{(k)}\}.$$

can be maximized to obtain the $(k+1)$th step estimates of Σ as

$$\Sigma^{(k+1)} = \frac{1}{nL} \sum_{l=1}^{L} \sum_{i=1}^{n} E\Big[\begin{pmatrix} [z_{i(l)}]^2 & z_{i(l)}u_{i(l)} \\ z_{i(l)}u_{i(l)} & [u_{i(l)}]^2 \end{pmatrix} |O_i, \theta^{(k)}\Big]. \tag{3.25}$$

Then the maximum likelihood estimate of θ can be obtained by the following step:

Step 1. Start with the initial value $\theta^{(0)}$;

Step 2. At the kth step, we calculate the $k+1$th step $\theta^{(k+1)}$; and then iterate this step until the desired convergence achieves. To calculate $E[[z_{i(l)}]^2|O_i, \theta^{(k)}]$, $E[z_{i(l)}u_{i(l)}|O_i, \theta^{(k)}]$ and $E[[u_{i(l)}]^2|O_i, \theta^{(k)}]$,

we let

$$\begin{aligned} f_i^*(\boldsymbol{u_i^*}, \boldsymbol{z_i^*}; \theta) = \exp\Big\{ \sum_{l=1}^{L} \big[-\Lambda_{(l)}(C_{i(l)})\big] \exp(X_i'\gamma + u_{i(l)}) \\ + K_{i(l)}u_{i(l)} + z_{i(l)}N_{i(l),K_i} \\ - \mu_{(l)}(T_{i(l),K_i}) \exp(X_i'\beta + z_{i(l)})\Big\}, \end{aligned}$$

for $i = 1, ..., n$, $\boldsymbol{u_i^*} = (u_{i(1)}, ..., u_{i(L)})$ and $\boldsymbol{z_i^*} = (z_{i(1)}, ..., z_{i(L)})$. Generally, for some function $h(u_{i(l)}, z_{i(l)})$,

$$E\{h(u_{i(l)}, z_{i(l)})|O_i, \theta^{(k)}\} = \int\int h(u_{i(l)}, z_{i(l)}) f(\boldsymbol{u_i^*}, \boldsymbol{z_i^*}|O_i, \theta^{(k)}) du_i dz_i,$$

where

$$f(\boldsymbol{u_i^*}, \boldsymbol{z_i^*}|O_i, \theta^{(k)}) = \frac{f_i^*(\boldsymbol{u_i^*}, \boldsymbol{z_i^*}; \theta^{(k)})\phi(\boldsymbol{u_i^*}, \boldsymbol{z_i^*}; \Sigma)}{\int\int f_i^*(\boldsymbol{u_i^*}, \boldsymbol{z_i^*}; \theta^{(k)})\phi(\boldsymbol{u_i^*}, \boldsymbol{z_i^*}; \Sigma) d\boldsymbol{u_i^*}\boldsymbol{z_i^*}}, \tag{3.26}$$

is the conditional density of U_i and Z_i given O_i and θ. It is apparent that this integration has no closed form. For this situation, with $\theta = \theta^{(k)}$, let $\{u_{i(l),q}, z_{i(l),q}; i = 1, ..., n, q = 1, ..., Q, l = 1, ..., L\}$ be L sets of Q i.i.d. samples from $N(0, \Sigma^{(k)})$, where Q is sufficiently large. Then one can approximate $E\{h(u_{i(l)}, z_{i(l)})|O_i, \theta^{(k)}\}$ by

$$\hat{E}\{h(u_{i(l)}, z_{i(l)})|O_i, \theta^{(k)}\} = \frac{\sum_{q=1}^{Q} h(u_{i(l),q}, z_{i(l),q}) f_i^*(\boldsymbol{u_{iq}^*}, \boldsymbol{z_{iq}^*}; \theta^{(k)})}{\sum_{q=1}^{Q} f_i^*(\boldsymbol{u_{iq}^*}, \boldsymbol{z_{iq}^*}; \theta^{(k)})}. \tag{3.27}$$

Following equation 3.27 above, $E[[z_{i(l)}]^2|O_i, \theta^{(k)}]$ can be approximated by

$$\hat{E}\{[z_{i(l)}]^2|O_i, \theta^{(k)}\} = \frac{\sum_{q=1}^{Q} [z_{i(l),q}]^2 f_i^*(\boldsymbol{u_{iq}^*}, \boldsymbol{z_{iq}^*}; \theta^{(k)})}{\sum_{q=1}^{Q} f_i^*(\boldsymbol{u_{iq}^*}, \boldsymbol{z_{iq}^*}; \theta^{(k)})}. \tag{3.28}$$

$E[z_{i(l)} u_{i(l)}|O_i, \theta^{(k)}]$ can be approximated by

$$\hat{E}\{z_{i(l)} u_{i(l)}|O_i, \theta^{(k)}\} = \frac{\sum_{q=1}^{Q} z_{i(l),q} u_{i(l),q} f_i^*(\boldsymbol{u_{iq}^*}, \boldsymbol{z_{iq}^*}; \theta^{(k)})}{\sum_{q=1}^{Q} f_i^*(\boldsymbol{u_{iq}^*}, \boldsymbol{z_{iq}^*}; \theta^{(k)})}. \tag{3.29}$$

and $E[[u_{i(l)}]^2|O_i, \theta^{(k)}]$ can be approximated by

$$\hat{E}\{[u_{i(l)}]^2|O_i, \theta^{(k)}\} = \frac{\sum_{q=1}^{Q} [u_{i(l),q}]^2 f_i^*(\boldsymbol{u_{iq}^*}, \boldsymbol{z_{iq}^*}; \theta^{(k)})}{\sum_{q=1}^{Q} f_i^*(\boldsymbol{u_{iq}^*}, \boldsymbol{z_{iq}^*}; \theta^{(k)})}. \tag{3.30}$$

For multivariate model B, the following log-likelihood function

$$l_n^{(3)}(\Sigma|O_i's, \theta^{(k)}) = E\{log\phi(u_{i(l)}, z_{i(l)}; \Sigma_{(l)})|O_i, \theta^{(k)}\}.$$

can be maximized to obtain the $(k+1)$th step estimates of $\Sigma_{(l)}$ as

$$\Sigma_{(l)}^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n} E\Big[\begin{pmatrix} [z_{i(l)}]^2 & z_{i(l)}u_{i(l)} \\ z_{i(l)}u_{i(l)} & [u_{i(l)}]^2 \end{pmatrix} |O_{i(l)}, \theta^{(k)}\Big]. \quad (3.31)$$

Then the maximum likelihood estimate of θ for event type l can be obtained using the following step:

Step 1. Start with the initial value $\theta_{(l)}^{(0)}$;

Step 2. At the kth step, we calculate the $k+1$th step $\theta_{(l)}^{(k+1)}$; and then iterate this step until the desired convergence achieves. To calculate $E[[z_{i(l)}]^2|O_{i(l)}, \theta_{(l)}^{(k)}]$, $E[z_{i(l)}u_{i(l)}|O_{i(l)}, \theta_{(l)}^{(k)}]$ and $E[[u_{i(l)}]^2|O_{i(l)}, \theta_{(l)}^{(k)}]$,

we let

$$\begin{aligned} f_{i(l)}^*(u_{i(l)}, z_{i(l)}; \theta_{(l)}) = & \exp(-\Lambda_{(l)}(C_{i(l)})\exp(X_i'\gamma + u_{i(l)}) \\ & + K_{i(l)}u_{i(l)} + z_{i(l)}N_{i(l),K_i} - \mu_{(l)}(T_{i(l),K_i})\exp(X_i'\beta + z_{i(l)})), \end{aligned}$$

for $i = 1, ..., n$. Generally, for some function $h(u_i, z_i)$ for event type l,

$$E\{h(u_{i(l)}, z_{i(l)})|O_{i(l)}, \theta_{(l)}^{(k)}\} = \int\int h(u_{i(l)}, z_{i(l)}) f(u_{i(l)}, z_{i(l)}|O_{i(l)}, \theta_{(l)}^{(k)}) du_{i(l)} dz_{i(l)},$$

where

$$f(u_{i(l)}, z_{i(l)}|O_{i(l)}, \theta_{(l)}) = \frac{f_{i(l)}^*(u_{i(l)}, z_{i(l)}; \theta_{(l)})\phi(u_{i(l)}, z_{i(l)}; \Sigma_{(l)})}{\int\int f_{i(l)}^*(u_{i(l)}, z_{i(l)}; \theta_{(l)})\phi(u_{i(l)}, z_{i(l)}; \Sigma_{(l)}) du_{i(l)} z_{i(l)}}, \tag{3.32}$$

is the conditional density of U_i and Z_i given O_i and θ. It is apparent that this integration has no closed form. For this situation, with $\theta_{(l)} = \theta_{(l)}^{(k)}$, let $\{u_{i(l),q}, z_{i(l),q}; i = 1, ..., n, q = 1, ..., Q, l = 1, ..., L\}$ be Q i.i.d. samples from $N(0, \Sigma_{(l)}^{(k)})$, where Q is sufficiently large. Then one can approximate $E\{h(u_{i(l)}, z_{i(l)})|O_{i(l)}, \theta_{(l)}^{(k)}\}$ by

$$\hat{E}\{h(u_{i(l)}, z_{i(l)})|O_{i(l)}, \theta_{(l)}^{(k)}\} = \frac{\sum_{q=1}^{Q} h(u_{i(l),q}, z_{i(l),q}) f_{i(l)}^*(u_{i(l),q}, z_{i(l),q}; \theta_{(l)}^{(k)})}{\sum_{q=1}^{Q} f_{i(l)}^*(u_{i(l),q}, z_{i(l),q}; \theta_{(l)}^{(k)})}. \tag{3.33}$$

Therefore, following the above equation, $E[[z_{i(l)}]^2|O_{i(l)}, \theta_{(l)}^{(k)}]$ can be approximated by

$$\hat{E}\{[z_{i(l)}]^2|O_{i(l)}, \theta_{(l)}^{(k)}\} = \frac{\sum_{q=1}^{Q} [z_{i(l),q}]^2 f_{i(l)}^*(u_{i(l),q}, z_{i(l),q}; \theta_{(l)}^{(k)})}{\sum_{q=1}^{Q} f_{i(l)}^*(u_{i(l),q}, z_{i(l),q}; \theta_{(l)}^{(k)})}. \tag{3.34}$$

$E[z_{i(l)}u_{i(l)}|O_{i(l)}, \theta_{(l)}^{(k)}]$ can be approximated by

$$\hat{E}\{z_{i(l)}u_{i(l)}|O_{i(l)}, \theta_{(l)}^{(k)}\} = \frac{\sum_{q=1}^{Q} z_{i(l),q} u_{i(l),q} f_{i(l)}^*(u_{i(l),q}, z_{i(l),q}; \theta_{(l)}^{(k)})}{\sum_{q=1}^{Q} f_{i(l)}^*(u_{i(l),q}, z_{i(l),q}; \theta_{(l)}^{(k)})}. \tag{3.35}$$

$E[[u_{i(l)}]^2|O_{i(l)}, \theta_{(l)}^{(k)}]$ can be approximated by

$$\hat{E}\{[u_{i(l)}]^2|O_{i(l)}, \theta_{(l)}^{(k)}\} = \frac{\sum_{q=1}^{Q}[u_{i(l),q}]^2 f_{i(l)}^*(u_{i(l),q}, z_{i(l),q}; \theta_{(l)}^{(k)})}{\sum_{q=1}^{Q} f_{i(l)}^*(u_{i(l),q}, z_{i(l),q}; \theta_{(l)}^{(k)})}. \tag{3.36}$$

For multivariate model C, the estimation procedure of Σ is very similar to that in multivariate model A, based on the fact that $\beta's$ and $\gamma's$ are different across event types in multivariate model C, the definition of $f_i^*(\boldsymbol{u_i^*}, \boldsymbol{z_i^*}; \theta)$ is slightly different, which is given by

$$\begin{aligned} f_i^*(\boldsymbol{u_i^*}, \boldsymbol{z_i^*}; \theta) = \exp\Bigg\{ \sum_{l=1}^{L} \big[-\Lambda_{(l)}(C_{i(l)})\big] \exp(X_i'\gamma_{(l)} + u_{i(l)}) \\ + K_{i(l)}u_{i(l)} + z_{i(l)}N_{i(l),K_i} \\ - \mu_{(l)}(T_{i(l),K_i}) \exp(X_i'\beta_{(l)} + z_{i(l)})\Bigg\}, \end{aligned}$$

for $i = 1, ..., n$, $\boldsymbol{u_i^*} = (u_{i(1)}, ..., u_{i(L)})$ and $\boldsymbol{z_i^*} = (z_{i(1)}, ..., z_{i(L)})$. Then for some function $h(u_{i(l)}, z_{i(l)})$,

$$E\{h(u_{i(l)}, z_{i(l)})|O_i, \theta^{(k)}\} = \int\int h(u_{i(l)}, z_{i(l)}) f(\boldsymbol{u_i^*}, \boldsymbol{z_i^*}|O_i, \theta^{(k)}) du_i dz_i,$$

where

$$f(\boldsymbol{u_i^*}, \boldsymbol{z_i^*}|O_i, \theta^{(k)}) = \frac{f_i^*(\boldsymbol{u_i^*}, \boldsymbol{z_i^*}; \theta^{(k)})\phi(\boldsymbol{u_i^*}, \boldsymbol{z_i^*}; \Sigma)}{\int\int f_i^*(\boldsymbol{u_i^*}, \boldsymbol{z_i^*}; \theta^{(k)})\phi(\boldsymbol{u_i^*}, \boldsymbol{z_i^*}; \Sigma) d\boldsymbol{u_i^*}\boldsymbol{z_i^*}}, \tag{3.37}$$

is the conditional density of U_i and Z_i given O_i and θ. It is apparent that this integration has no closed form. For this situation, with $\theta = \theta^{(k)}$, let $\{u_{i(l),q}, z_{i(l),q}; i = 1, ..., n, q = 1, ..., Q, l =$

$1, ..., L\}$ be L sets of Q i.i.d. samples from $N(0, \Sigma^{(k)})$, where Q is sufficiently large. Then one can approximate $E\{h(u_{i(l)}, z_{i(l)})|O_i, \theta^{(k)}\}$ by

$$\hat{E}\{h(u_{i(l)}, z_{i(l)})|O_i, \theta^{(k)}\} = \frac{\sum_{q=1}^{Q} h(u_{i(l),q}, z_{i(l),q}) f_i^*(\boldsymbol{u}_{\boldsymbol{iq}}^*, \boldsymbol{z}_{\boldsymbol{iq}}^*; \theta^{(k)})}{\sum_{q=1}^{Q} f_i^*(\boldsymbol{u}_{\boldsymbol{iq}}^*, \boldsymbol{z}_{\boldsymbol{iq}}^*; \theta^{(k)})}. \tag{3.38}$$

Following the equation above, $E[[z_{i(l)}]^2|O_i, \theta^{(k)}]$ can be approximated by

$$\hat{E}\{[z_{i(l)}]^2|O_i, \theta^{(k)}\} = \frac{\sum_{q=1}^{Q} [z_{i(l),q}]^2 f_i^*(\boldsymbol{u}_{\boldsymbol{iq}}^*, \boldsymbol{z}_{\boldsymbol{iq}}^*; \theta^{(k)})}{\sum_{q=1}^{Q} f_i^*(\boldsymbol{u}_{\boldsymbol{iq}}^*, \boldsymbol{z}_{\boldsymbol{iq}}^*; \theta^{(k)})}. \tag{3.39}$$

$E[z_{i(l)} u_{i(l)}|O_i, \theta^{(k)}]$ can be approximated by

$$\hat{E}\{z_{i(l)} u_{i(l)}|O_i, \theta^{(k)}\} = \frac{\sum_{q=1}^{Q} z_{i(l),q} u_{i(l),q} f_i^*(\boldsymbol{u}_{\boldsymbol{iq}}^*, \boldsymbol{z}_{\boldsymbol{iq}}^*; \theta^{(k)})}{\sum_{q=1}^{Q} f_i^*(\boldsymbol{u}_{\boldsymbol{iq}}^*, \boldsymbol{z}_{\boldsymbol{iq}}^*; \theta^{(k)})}. \tag{3.40}$$

$E[[u_{i(l)}]^2|O_i, \theta^{(k)}]$ can be approximated by

$$\hat{E}\{[u_{i(l)}]^2|O_i, \theta^{(k)}\} = \frac{\sum_{q=1}^{Q} [u_{i(l),q}]^2 f_i^*(\boldsymbol{u}_{\boldsymbol{iq}}^*, \boldsymbol{z}_{\boldsymbol{iq}}^*; \theta^{(k)})}{\sum_{q=1}^{Q} f_i^*(\boldsymbol{u}_{\boldsymbol{iq}}^*, \boldsymbol{z}_{\boldsymbol{iq}}^*; \theta^{(k)})}. \tag{3.41}$$

3.4 Discussion

In this chapter, we discussed three different model assumptions and the corresponding estimation procedures for the situation in which multiple related types of events are presented. The performance of each of the three proposed models will be demonstrated by simulation studies in Chapter 5. Application to real-world data will be implemented and discussed in Chapter 6. Our proposed approaches are capable of incorporating the same or event-type-specific covariate effects and flexible in modeling different types of correlation between the observation process and the point process that characterizes the occurrence of the event depending on the major research purposes. Furthermore, the algorithms for all of the three models are relatively easy to implement although the bootstrap methods for the variance estimation may take a longer computation time in real-world data analysis. One limitation of the proposed approach is the lack of consideration of an informative follow-up time process. A possible direction for future research is to develop an approach that allows the follow-up time to be related with the observation process as well as the point process of interest. This situation can occur if the disease of interest is not indolent and a relatively large proportion of severe patients may tend to drop out of the study earlier. This problem has been discussed by some previous literature with multi-step inference procedures (He et al. 2009), but has not been tackled in the multivariate panel count data setting.

Chapter 4: Model Checking

4.1 Overview

This chapter presents model checking procedures for the proposed univariate models as well as all types of multivariate models. The procedures are developed based on the model checking procedures discussed in Lin et al. (1993), which were derived from cumulative sums of martingale-based residuals over covariate values. The underlying distributions of such stochastic processes can be approximated by zero-mean Gaussian processes. A number of simulated realizations from the approximate null distribution can be assessed to compare with each observed process. The functional form of a covariate can then be checked to assess the objectivity of the proposed models by visually assessing how unusual the observed pattern is and by a supremum tests, which will be discussed later.

4.2 Martingale-Based Residuals

Martingale residuals can provide useful clues on appropriate functional forms of covariates by generating plots of the martingale residuals versus covariates, which are helpful to detect potential model departures. To describe the martingale residuals, the counting process is defined as

$$N_i(t) = \sum_{k=1}^{K_i} \Delta N_{ik} I\{T_{ik} \leq t\} \quad (i = 1, ..., n), \tag{4.1}$$

where $\Delta N_{ik} = N_i(T_{ik}) - N_i(T_{ik-1})$. These processes have the intensity functions

$$\mu_i(t) = \mu_0(t) \exp\{X_i'\beta + Z_i\} \quad (i = 1, ..., n). \tag{4.2}$$

The differences between the counting processes and their respective integrated intensity functions can be defined as

$$M_i(t) = N_i(t) - \int_0^t \mu_0(u) \exp\{X_i'\beta + Z_i\} du \quad (i = 1, ..., n), \tag{4.3}$$

and they are martingales. The martingale residuals are defined as

$$\hat{M}_i(t) = N_i(t) - \int_0^t \exp\{X_i'\hat{\beta}\} E\{e^{z_i} | O_i, \hat{\theta}\} d\hat{\mu}(u) \quad (i = 1, ..., n), \tag{4.4}$$

where

$$\hat{\mu}(t) = \sum_{m=1}^{M} \mu_m I(t \geq t_m),$$

M is the number of the distinct time points $T_{i,j}'s$, and

$$\mu_m = \mu(t_m) - \mu(t_{m-1}).$$

The quantities μ_m can be estimated by using the self-consistent algorithm discussed in Chapter 2. The martingale-based residuals can then be grouped cumulatively with respect to covariate values. We then define the following stochastic process:

$$W_x(t,x) = \sum_{i=1}^{n} f(X_i) I(X_i \leq x) \hat{M}_i(t), \tag{4.5}$$

where $f(.)$ is a known smooth function, $x = (x_1, ..., x_p)' \in R^p$, and the event $\{X_i \leq x\}$ means that all the p components of X_i are no larger than the respective components of x. This process will fluctuate randomly around zero if model (4.2) holds. The approximation of the null distribution is shown in the next section.

The model checking procedures can be extended to the multivariate analysis. In the multivariate modeling, the counting process for event type l is defined as

$$N_{i(l)}(t) = \sum_{k=1}^{K_{i(l)}} \Delta N_{i(l),k} I\{T_{i(l),k} \leq t\} \quad (i = 1, ..., n), \tag{4.6}$$

where $\Delta N_{i(l),k} = N_{i(l)}(T_{i(l),k}) - N_{i(l)}(T_{i(l),k-1})$. For each multivariate model, these processes have different mean functions.

For the multivariate model A, the processes have the following conditional mean functions

$$\mu_{i(l)}(t) = \mu_{0(l)}(t) \exp\{X_i'\beta + Z_i\} \quad (i = 1, ..., n). \tag{4.7}$$

For the multivariate model B, the processes have the following conditional mean functions

$$\mu_{i(l)}(t) = \mu_{0(l)}(t) \exp\{X_i'\beta + Z_{i(l)}\} \quad (i = 1, ..., n). \tag{4.8}$$

For the multivariate model C, the processes have the following conditional mean functions

$$\mu_{i(l)}(t) = \mu_{0(l)}(t) \exp\{X_i'\beta_{(l)} + Z_i\} \quad (i = 1, ..., n). \tag{4.9}$$

Then for the multivariate model A, the differences between the counting processes and their respective integrated intensity functions for event type l are defined as

$$M_{i(l)}(t) = N_{i(l)}(t) - \int_0^t \mu_{0(l)}(u) \exp\{X_i'\beta + Z_i\} du \quad (i = 1, ..., n), \tag{4.10}$$

which are martingales.

Similarly, for the multivariate model B, the differences between the counting processes and their respective integrated intensity functions for event type l are martingales and are defined as

$$M_{i(l)}(t) = N_{i(l)}(t) - \int_0^t \mu_{0(l)}(u) \exp\{X_i'\beta + Z_{i(l)}\} du \quad (i = 1, ..., n), \tag{4.11}$$

The differences between the counting processes and their respective integrated intensity functions for event type l for the multivariate model C are also martingales

$$M_{i(l)}(t) = N_{i(l)}(t) - \int_0^t \mu_{0(l)}(u) \exp\{X_i'\beta_{(l)} + Z_i\} du \quad (i = 1, ..., n). \tag{4.12}$$

The martingale residuals for event type l can then be obtained from the above expression. For the multivariate model A, the corresponding martingale residuals are defined as

$$\hat{M}_{i(l)}(t) = N_{i(l)}(t) - \int_0^t \exp\{X_i'\hat{\beta}\} E\{e^{z_i} | O_i, \hat{\theta}_{(l)}\} d\hat{\mu}_{(l)}(u) \quad (i = 1, ..., n), \tag{4.13}$$

where

$$\hat{\mu}_{(l)}(t) = \sum_{m=1}^{M_{(l)}} \mu_{m(l)} I(t \geq t_m),$$

where $M_{(l)}$ is the number of the distinct time points $T'_{ij}s$ for event type l and

$$\mu_{m(l)} = \mu_{(l)}(t_m) - \mu_{(l)}(t_{m-1}).$$

Similarly, the martingale residuals for event type l in the multivariate model B are defined as

$$\hat{M}_{i(l)}(t) = N_{i(l)}(t) - \int_0^t \exp\{X_i'\hat{\beta}\} E\{e^{z_{i(l)}}|O_i, \hat{\theta}_{(l)}\} d\hat{\mu}_{(l)}(u) \quad (i = 1, ..., n). \tag{4.14}$$

Similarly the martingale residuals for event type l in the multivariate model C are defined as

$$\hat{M}_{i(l)}(t) = N_{i(l)}(t) - \int_0^t \exp\{X_i'\hat{\beta}_{(l)}\} E\{e^{z_i}|O_i, \hat{\theta}_{(l)}\} d\hat{\mu}_{(l)}(u) \quad (i = 1, ..., n). \tag{4.15}$$

For all of the multivariate models, μ_m can be estimated by using the self-consistent algorithm discussed in Chapter 2. The martingale-based residuals can then be grouped cumulatively with respect to covariate values. Then for each event type l, We can define the following stochastic process:

$$W_{x(l)}(t, x) = \sum_{i=1}^{n} f(X_i) I(X_i \leq x) \hat{M}_{i(l)}(t), \tag{4.16}$$

where $f(.)$ is a known smooth function, $x = (x_1, ..., x_p)' \in R^p$, and the event $\{X_i \leq x\}$ means that all the p components of X_i are no larger than the respective components of x. Again, this process will fluctuate randomly around zero if model 4.2 holds.

4.3 Null Distributions

To check the functional form of a covariate, Lin et al. (1993) proposed an objective graphical solution plots the partial-sum processes of the $\hat{M}_i(.)$ instead of plotting the raw martingale residuals,

$$W_0(x) = \sum_{i=1}^{n} I(X_i \leq x)\hat{M}_i(\infty), \tag{4.17}$$

The partial-sum processes $W_0(x)$ is a special case of $W_x(t, x)$ with $f(X_i) = 1$ and $t = \infty$. Then the null distribution of $W_0(x)$ can be approximated through Monte Carlo simulations. Here, we derive the following approximation of $W_0(x)$ for the proposed univariate model

$$\begin{aligned}\hat{W}_0(x) =& \sum_{i=1}^{n} N_{iK_i}\Big\{I(X_i \leq x) - g(\hat{\beta}, x)\Big\}G_i \\ &- \sum_{i=1}^{n} \mu(T_{i,K_i})\exp(X_i'\hat{\beta})E(e^{z_j}|O_i, \hat{\theta})I(X_i \leq x)\Big\{X_i - \bar{X}(\hat{\beta})\Big\} \\ &\times \mathscr{F}^{-1}(\hat{\beta})\sum_{i=1}^{n} N_{iK_i}\Big\{X_i - \bar{X}(\hat{\beta})\Big\}G_i,\end{aligned} \tag{4.18}$$

where

$$g(\hat{\beta}, x) = \frac{\sum_{j=1}^{n}\exp(X_j'\hat{\beta})E(e^{z_j}|O_i, \hat{\theta})I(X_j \leq x)}{\sum_{j=1}^{n}\exp(X_j'\hat{\beta})E(e^{z_j}|O_i, \hat{\theta})},$$

and

$$\bar{X}(\hat{\beta}) = \frac{\sum_{j=1}^{n} \exp(X_j'\hat{\beta}) E(e^{z_j}|O_i, \hat{\theta}) X_j}{\sum_{j=1}^{n} \exp(X_j'\hat{\beta}) E(e^{z_j}|O_i, \hat{\theta})}.$$

Here $\{G_i; i = 1, ..., n\}$ denotes a random sample of standard normal variables, and $\mathscr{F}^{-1}(\hat{\beta})$ is the inverse of the Fisher information derived from the proposed likelihood function. Similar techniques can also be extended to multivariate cases. Under the multivariate model A, the null distribution of $W_x(t, x)$ for event type l can also be approximated through Monte Carlo simulations. We then obtain

$$\begin{aligned}
\hat{W}_{0(l)}(x) = & \sum_{i=1}^{n} N_{iK_i(l)} \Big\{ I(X_i \le x) - g_{(l)}(\hat{\beta}, x) \Big\} G_i \\
& - \sum_{i=1}^{n} \mu_{(l)}(T_{i,K_i}) \exp(X_i'\hat{\beta}) E(e^{z_j}|O_i, \hat{\theta}_{(l)}) I(X_i \le x) \Big\{ X_i - \bar{X}(\hat{\beta}) \Big\} \\
& \times \mathscr{F}_{(l)}^{-1}(\hat{\beta}) \sum_{i=1}^{n} N_{iK_i(l)} \Big\{ X_i - \bar{X}(\hat{\beta}) \Big\} G_i,
\end{aligned} \tag{4.19}$$

where

$$g_{(l)}(\hat{\beta}, x) = \frac{\sum_{j=1}^{n} \exp(X_j'\hat{\beta}) E(e^{z_j}|O_i, \hat{\theta}_{(l)}) I(X_j \le x)}{\sum_{j=1}^{n} \exp(X_j'\hat{\beta}) E(e^{z_j}|O_i, \hat{\theta}_{(l)})},$$

and

$$\bar{X}(\hat{\beta}) = \frac{\sum_{j=1}^{n} \exp(X_j'\hat{\beta}) E(e^{z_j}|O_i, \hat{\theta}_{(l)}) X_j}{\sum_{j=1}^{n} \exp(X_j'\hat{\beta}) E(e^{z_j}|O_i, \hat{\theta}_{(l)})}.$$

Under the multivariate model B, the null distribution of $W_x(t, x)$ for event type l can also be approximated through Monte Carlo simulations as

$$
\begin{aligned}
\hat{W}_{0(l)}(x) =& \sum_{i=1}^{n} N_{iK_i(l)} \Big\{ I(X_i \leq x) - g_{(l)}(\hat{\beta}, x) \Big\} G_i \\
& - \sum_{i=1}^{n} \mu_{(l)}(T_{i,K_i}) \exp(X_i'\hat{\beta}) E(e^{z_{j(l)}} | O_i, \hat{\theta}_{(l)}) I(X_i \leq x) \Big\{ X_i - \bar{X}(\hat{\beta}) \Big\} \\
& \times \mathscr{F}_{(l)}^{-1}(\hat{\beta}) \sum_{i=1}^{n} N_{iK_i(l)} \Big\{ X_i - \bar{X}(\hat{\beta}) \Big\} G_i,
\end{aligned} \tag{4.20}
$$

where

$$
g_{(l)}(\hat{\beta}, x) = \frac{\sum_{j=1}^{n} \exp(X_j'\hat{\beta}) E(e^{z_{j(l)}} | O_i, \hat{\theta}_{(l)}) I(X_j \leq x)}{\sum_{j=1}^{n} \exp(X_j'\hat{\beta}) E(e^{z_{j(l)}} | O_i, \hat{\theta}_{(l)})},
$$

and

$$
\bar{X}(\hat{\beta}) = \frac{\sum_{j=1}^{n} \exp(X_j'\hat{\beta}) E(e^{z_{j(l)}} | O_i, \hat{\theta}_{(l)}) X_j}{\sum_{j=1}^{n} \exp(X_j'\hat{\beta}) E(e^{z_{j(l)}} | O_i, \hat{\theta}_{(l)})}.
$$

For the multivariate model C, the approximated null distribution of $W_x(t, x)$ for event type l can be obtained by

$$
\begin{aligned}
\hat{W}_{0(l)}(x) =& \sum_{i=1}^{n} N_{iK_i(l)} \Big\{ I(X_i \leq x) - g_{(l)}(\hat{\beta}_{(l)}, x) \Big\} G_i \\
& - \sum_{i=1}^{n} \mu_{(l)}(T_{i,K_i}) \exp(X_i'\hat{\beta}_{(l)}) E(e^{z_j} | O_i, \hat{\theta}_{(l)}) I(X_i \leq x) \Big\{ X_i - \bar{X}(\hat{\beta}_{(l)}) \Big\} \\
& \times \mathscr{F}_{(l)}^{-1}(\hat{\beta}_{(l)}) \sum_{i=1}^{n} N_{iK_i(l)} \Big\{ X_i - \bar{X}(\hat{\beta}_{(l)}) \Big\} G_i,
\end{aligned} \tag{4.21}
$$

where

$$
g_{(l)}(\hat{\beta}_{(l)}, x) = \frac{\sum_{j=1}^{n} \exp(X_j'\hat{\beta}_{(l)}) E(e^{z_j} | O_i, \hat{\theta}_{(l)}) I(X_j \leq x)}{\sum_{j=1}^{n} \exp(X_j'\hat{\beta}_{(l)}) E(e^{z_j} | O_i, \hat{\theta}_{(l)})},
$$

and

$$\bar{X}(\hat{\beta}_{(l)}) = \frac{\sum_{j=1}^{n} \exp(X_j'\hat{\beta}_{(l)}) E(e^{z_j}|O_i, \hat{\theta}_{(l)}) X_j}{\sum_{j=1}^{n} \exp(X_j'\hat{\beta}_{(l)}) E(e^{z_j}|O_i, \hat{\theta}_{(l)})}.$$

4.4 Supremum Test

To further evaluate the objectivity of the residual plot, the extremity of $w_0(.)$ can be identified. Since $W_0(.)$ should fluctuate randomly around zero under the null hypothesis, a natural numerical measure is given by

$$s_0 = sup_x|w_0(x)|.$$

From the definition above, an unusually large value of s_0 suggests the inappropriateness of the functional form for the covariates. The p-value of the test can be approximated by pr$(\hat{S}_0 \geq s_0)$, where

$$\hat{S}_0 = sup_x|\hat{W}_0(x)|.$$

In addition, pr$(\hat{S}_0 \geq s_0)$ can be estimated by generating a large number of normal samples $\{G_i\}$ and the calculation of pr$(\hat{S}_0 \geq s_0)$ is conditional on $\{N_{iK_i}, X_i\}$.

Under the multivariate setup, a similar supremum test can be conducted to evaluate the objectivity of the residual plot for each event type l, and the extremity of $w_{0(l)}(.)$ can then be identified. The p-value of the test can be approximated following similar set-up in the univariate analysis. In Chapter 6, we will apply the model checking techniques discussed above to both univariate and multivariate analyses of bivariate panel count data from the skin cancer chemoprevention trial. Plots of cumulative martingale residuals will be used to evaluate the appropriateness of the functional forms of covariates, and the supremum tests will be conducted to further assess the objectivity of the residual plots under each proposed model.

Chapter 5: Simulation Studies

5.1 Univariate Panel Count Data

5.1.1 Description of Simulation

Simulation studies were conducted to evaluate the performance of the proposed methodology under various situations. In the simulation study, the distribution of C_i, the follow-up time for patient i, does not depend on the covariates, and the covariates $x_i^{\prime}s$ were assumed to follow two types of distributions, a normal distribution with mean 0 and variance 0.25, and a Bernoulli distribution with success probability 0.5. The follow-up times, $C_i, i = 1, ...n$, were generated from a uniform distribution over $(\tau/2, \tau)$, where τ is set to 10. The subject-specific frailties z_i and u_i were generated from a bivariate normal distribution with mean $(0, 0)$ and variance-covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_z^2 & \sigma_{zu} \\ \sigma_{zu} & \sigma_u^2 \end{pmatrix},$$

where the initial values for the variance-covariance matrix were set to $\sigma_{z_0}^2 = 0.25$, $\sigma_{zu_0} = 0.1$, and $\sigma_{u_0}^2 = 0.2$. For the event counting process, we assumed that N_i is a nonhomogeneous Poisson process with $\mu_0(t) = t/(5\tau)$. For the observation process, we assumed that H_i is a nonhomogeneous Poisson process with $\lambda_0(t) = t/(20\tau)$. Then given x_i and u_i, K_i^*, which is the number of observation times for subject i, followed a Poisson distribution with the mean function $\Lambda(C_i|x_i, u_i) = \Lambda_0(C_i)e^{\gamma x_i + u_i}$, and the observation times $(t_{i1}, t_{i2}, ..., t_{iK_i^*})$ are the order statistics of the random sample of size K_i^* from the uniform distribution over $(0, C_i)$.

The second step was to generate panel count data $N_i(t_{ij})'s$ by assuming that

$$N_i(t_{ij}) = N_i^*(t_{i1}) + N_i^*(t_{i2} - t_{i1}) + ... + N_i^*(t_{ij} - t_{ij-1})$$

where $j = 1, ..., K_i^*$ and $i = 1, ..., n$. Given x_i and z_i, $N_i^*(t)$ followed a Poisson distribution with the mean function $\mu_0(t)e^{\beta x_i + z_i}$. The results presented below were based on two sample sizes, n = 100 and n = 200. The replications were set to be 1,000 for obtaining the coverage probabilities for β. For the size of the bootstrap samples, we considered multiple values in the simulation studies and obtained similar results. Therefore, the results with the size 20 of bootstrap samples are presented.

5.1.2 Simulation Results

Table 5.1 shows the simulation results when the $x_i's$ follow a normal distribution and $\gamma = 1$ for the observation process. The true value of β was set to be -1, 0, or 1. In the table, the average of estimates of β, the sample standard deviations of the estimates (SSD), the means of bootstrap standard deviation estimates (BSD), and the 95% empirical coverage probabilities for β (CP) are given.

The results for the case when the $x_i's$ follow a Bernoulli distribution are presented in Table 5.2. In the table, all of other setups are the same as in Table 5.1.

Table 5.1: Univariate Model: Estimation of β, $x_i \sim N(0, 0.25)$

	True β					
	-1		**0**		**1**	
Sample Size	**N=100**	**N=200**	**N=100**	**N=200**	**N=100**	**N=200**
β	-0.9910	-0.9919	0.0099	-0.0060	0.9792	1.0158
SSD	0.1612	0.1179	0.1453	0.1005	0.1554	0.1190
BSD	0.1487	0.1119	0.1333	0.0989	0.1528	0.1128
CP	0.932	0.961	0.941	0.963	0.942	0.961

Table 5.2: Univariate Model: Estimation of β, $x_i \sim Bernoulli(0.5)$

	True β					
	-1		**0**		**1**	
Sample Size	**N=100**	**N=200**	**N=100**	**N=200**	**N=100**	**N=200**
β	-1.0108	-0.996	-0.0075	-0.0004	1.0226	1.0043
SSD	0.1427	0.1042	0.1405	0.0969	0.1308	0.1035
BSD	0.1450	0.1013	0.1352	0.1015	0.1431	0.1020
CP	0.963	0.955	0.942	0.956	0.965	0.959

5.2 Multivariate Panel Count Data

5.2.1 Description of Simulation

Following similar setup in the simulation studies for univariate panel count data, the covariates $x_i's$ were assumed to follow a normal distribution with mean 0 and variance 0.25. The follow-up times C_i was generated from a uniform distribution over $(\tau/2, \tau)$ for each event of interest. The subject-specific frailty z_i and u_i were assumed to be event-type specific, and were generated from bivariate normal distributions with mean $(0, 0)$ and variance-covariance matrix

$$\Sigma_{(l)} = \begin{pmatrix} \sigma^2_{z(l)} & \sigma_{zu(l)} \\ \sigma_{zu(l)} & \sigma^2_{u(l)} \end{pmatrix}, \tag{5.1}$$

where $l = 1, ..., L$, and L was set to 2 in the simulation study. The initial values for the first variance-covariance matrix was set to $[\sigma_{z_0(1)}]^2 = 0.4$, $\sigma_{zu_0(1)} = 0.1$, and $[\sigma_{u_0(1)}]^2 = 0.2$; and the second variance-covariance matrix was set to $[\sigma_{z_0(2)}]^2 = 0.2$, $\sigma_{zu_0(2)} = 0.05$, and $[\sigma_{z_0(2)}]^2 = 0.1$. For both types of events, the event count process, $N_{i(l)}$ was assumed to follow a nonhomogeneous Poisson process with $\mu_{0(l)}(t) = t/(5\tau)$, and the observation process $H_{i(l)}$ was assumed to follow a nonhomogeneous Poisson process with $\lambda_{0(l)}(t) = t/(20\tau)$. Then given x_i and $u_{i(l)}$, $K^*_{i(l)}$, which is the number of observation times for event type l for subject i, followed a Poisson distribution with the mean functions $\Lambda_{(l)}(C_{i(l)}|x_i, u_{i(l)}) = \Lambda_{0(l)}(C_{i(l)})e^{\gamma x_i + u_{i(l)}}$, and the observation times $(t_{i(l),1}, t_{i(l),2}, ..., t_{i(l),K^*_i})$ are the order statistics of the random sample of size $K^*_{i(l)}$ from the uniform distribution over $(0, C_{i(l)})$.

Then the panel count data $N_{i(l)}(t_{ij})'s$ were generated by assuming that

$$N_{i(l)}(t_{ij}) = N^*_{i(l)}(t_{i(l),1}) + N^*_{i(l)}(t_{i(l),2} - t_{i(l),1}) + ... + N^*_{i(l)}(t_{i(l),j} - t_{i(l),j-1})$$

where $j = 1, ..., K^*_{i(l)}$ and $i = 1, ..., n$. Given x_i and $z_{i(l)}$, $N^*_{i(l)}(t)$ followed a Poisson distribution with the mean function $\mu_{0(l)}(t)e^{\beta x_i + z_{i(l)}}$. The results presented below are based on n = 100 and n = 200. The replications were set to be 1,000 for obtaining the coverage probabilities for β. The size of bootstrap samples was set to be 20, consistent with the setup in the simulation studies for univariate panel count data.

5.2.2 Simulation Results

5.2.2.1 Multivariate Model A

Table 5.3 shows the simulation results for model A when β, γ, and Σ are all constant. In this simulation setup, $x_i's$ followed a normal distribution, and the true value of γ was set to 1 for the observation process. The true value of β was set to be -1, 0, or 1. In the table, the average of estimates of β, the sample standard deviations of the estimates (SSD), the means of bootstrap standard deviation estimates (BSD), and the 95% empirical coverage probabilities for β (CP) are presented.

Table 5.3: Multivariate Model A: Estimation of β, $x_i \sim N(0, 0.25)$

	True β					
	-1		**0**		**1**	
Sample Size	**N=100**	**N=200**	**N=100**	**N=200**	**N=100**	**N=200**
β	-1.0137	-1.0118	0.0102	0.0023	1.0203	1.0012
SSD	0.1135	0.0818	0.1115	0.0776	0.1190	0.0854
BSD	0.1032	0.0760	0.1022	0.0762	0.1081	0.0809
CP	0.932	0.940	0.934	0.956	0.942	0.948

5.2.2.2 Multivariate Model B

Table 5.4 shows the simulation results for model B when β and γ are constant but Σ is event-type-specific. In this simulation setup, $x_i's$ followed a normal distribution, and the true value of γ was set to 1 for the observation process. The true value of β was set to be -1, 0, or 1. In the table, the average of estimates of β, the sample standard deviations of the estimates (SSD), the means of bootstrap standard deviation estimates (BSD), and the 95% empirical coverage probabilities for β (CP) are reported.

Table 5.4: Multivariate Model B: Estimation of β, $x_i \sim N(0, 0.25)$

	True β					
	-1		**0**		**1**	
Sample Size	**N=100**	**N=200**	**N=100**	**N=200**	**N=100**	**N=200**
β	-1.0111	-1.0025	0.0096	0.0026	0.9885	1.0041
SSD	0.1140	0.0833	0.0993	0.0777	0.1038	0.0767
BSD	0.1013	0.0718	0.1028	0.0699	0.0987	0.0730
CP	0.928	0.946	0.962	0.942	0.958	0.948

5.2.2.3 Multivariate Model C

Table 5.5 shows the simulation results for model C when β and γ are event-type-specific but Σ is constant. In this simulation setup, $x_i's$ followed a normal distribution, and the true value of γ was set to 1 for the observation process. The true value of $\beta's$ was set to have three combinations as listed on the second row of the table. In the table, the average of estimates for each β, the sample standard deviations of the estimates (SSD), the means of bootstrap standard deviation estimates (BSD), and the 95% empirical coverage probabilities for each β (CP) are presented.

Table 5.5: Multivariate Model C: Estimation of $\beta's$, $x_i \sim N(0, 0.25)$

Sample Size	**N=100**	**N=200**	**N=100**	**N=200**	**N=100**	**N=200**
True (β_1, β_2)	(-1, 0)	(-1, 0)	(1, 0)	(1, 0)	(1, -1)	(1, -1)
$\hat{\beta}_1$	-1.0285	-1.0073	0.9892	1.0123	1.0436	1.0226
$\hat{\beta}_2$	0.0147	-0.0064	0.0175	0.0043	-1.0100	-1.0398
SSD $(\hat{\beta}_1)$	0.1930	0.1354	0.1960	0.1316	0.1967	0.1390
SSD $(\hat{\beta}_2)$	0.1708	0.1280	0.1875	0.1308	0.1740	0.1471
BSD $(\hat{\beta}_1)$	0.1666	0.1193	0.1630	0.1155	0.1689	0.1185
BSD $(\hat{\beta}_2)$	0.1590	0.1167	0.1601	0.1129	0.1622	0.1217
CP $(\hat{\beta}_1)$	0.916	0.924	0.920	0.918	0.918	0.928
CP $(\hat{\beta}_2)$	0.942	0.946	0.942	0.926	0.948	0.902

5.3 Discussion

The results from Table 5.1 and Table 5.2 suggest that the proposed estimation procedure yields a consistent performance for different distributions of $x_i's$. The point estimates seem to be nearly unbiased. In addition, the bootstrap sample standard deviation estimates are quite close to the sample standard deviation of the estimates, which indicates that the proposed variance estimation procedure works well. The estimates of variances become smaller as the sample size increases. It is also shown that the coverage probabilities are consistent with the nominal levels.

In addition to the coverage probabilities, we also constructed the quantile plots of the standardized estimates of β to assess the performance of the normal approximation to the finite distribution of the estimate of β. The QQ plots for the univariate cases are presented in Figure 5.1 to 5.3, which indicate a good match between the quantiles of our estimators and the quantiles for a standard normal distribution. The QQ plots for all of the three types of multivariate models are also presented (Figure 5.4 to Figure 5.6 for the multivariate model A, Figure 5.7 to Figure 5.9 for the multivariate model B, and Figure A.1 to Figure A.6 in Appendix A for the multivariate model C) to demonstrate appropriate normal approximations.

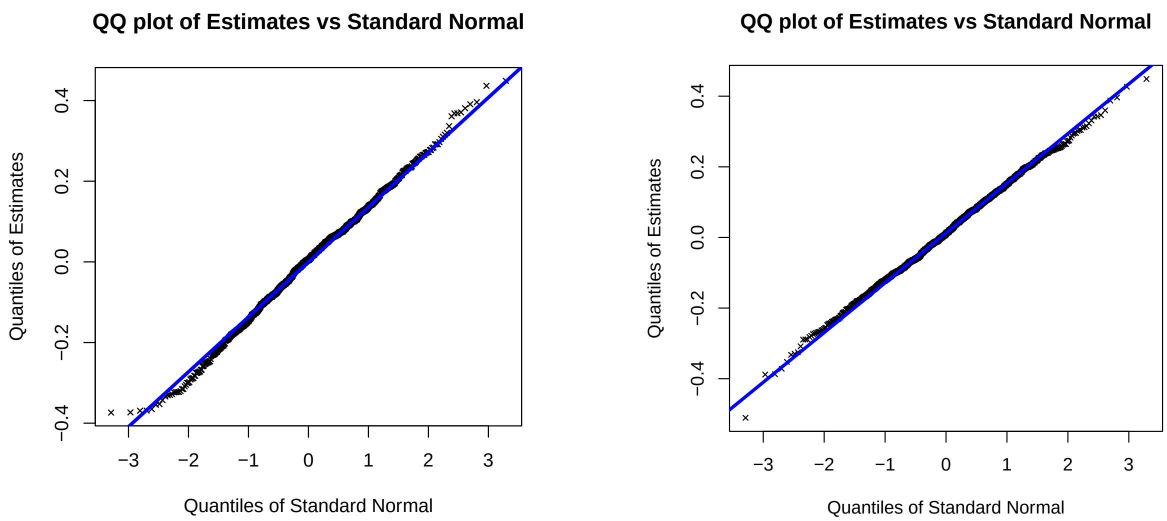


Figure 5.1: Univariate Case: QQ plot for $\beta = 0, N = 100$ (Left) and $N = 200$ (Right)

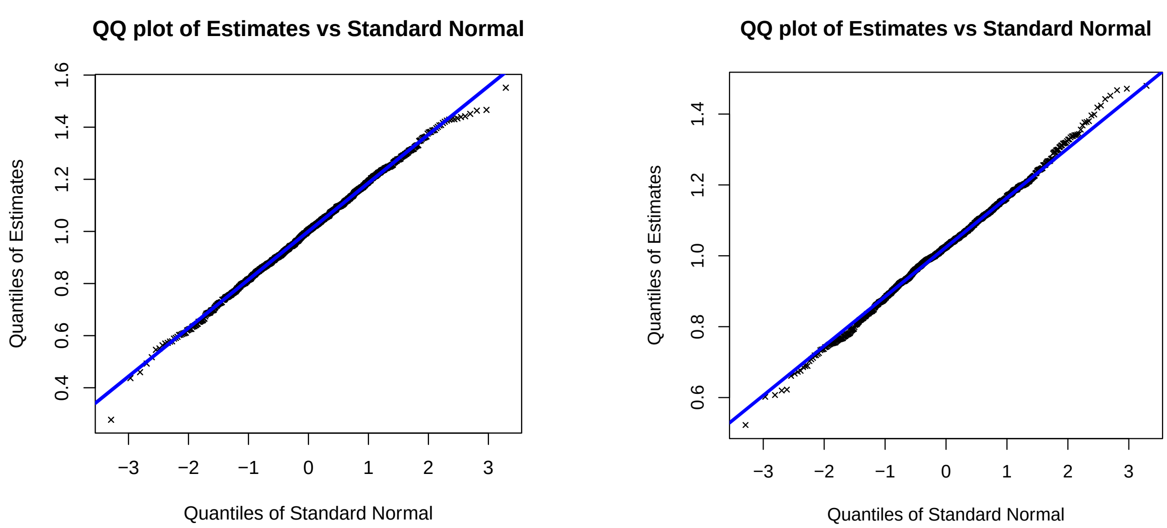


Figure 5.2: Univariate Case: QQ plot for $\beta = 1, N = 100$ (Left) and $N = 200$ (Right)

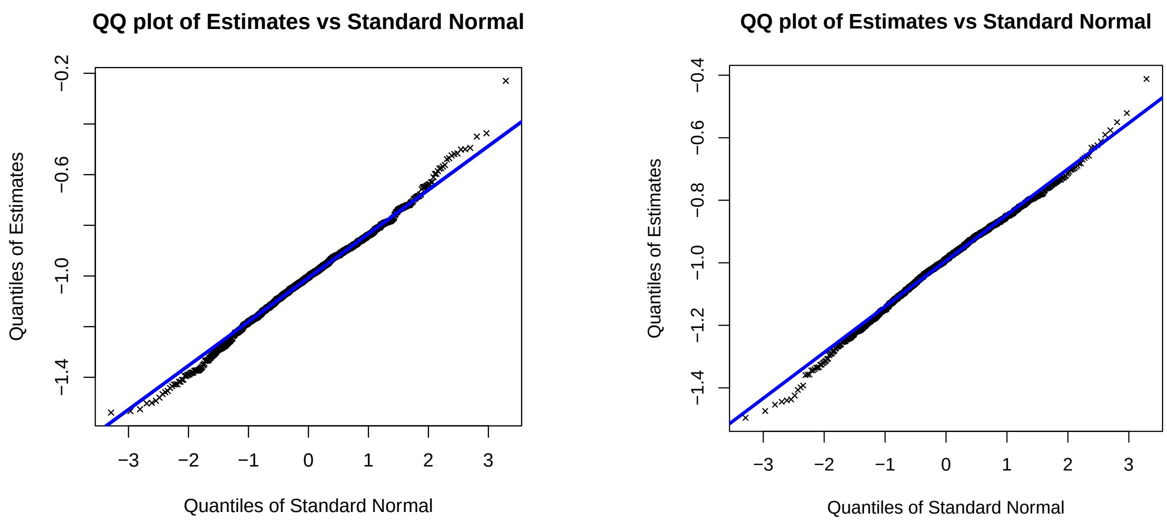


Figure 5.3: Univariate Case: QQ plot for $\beta = -1, N = 100$ (Left) and $N = 200$ (Right)

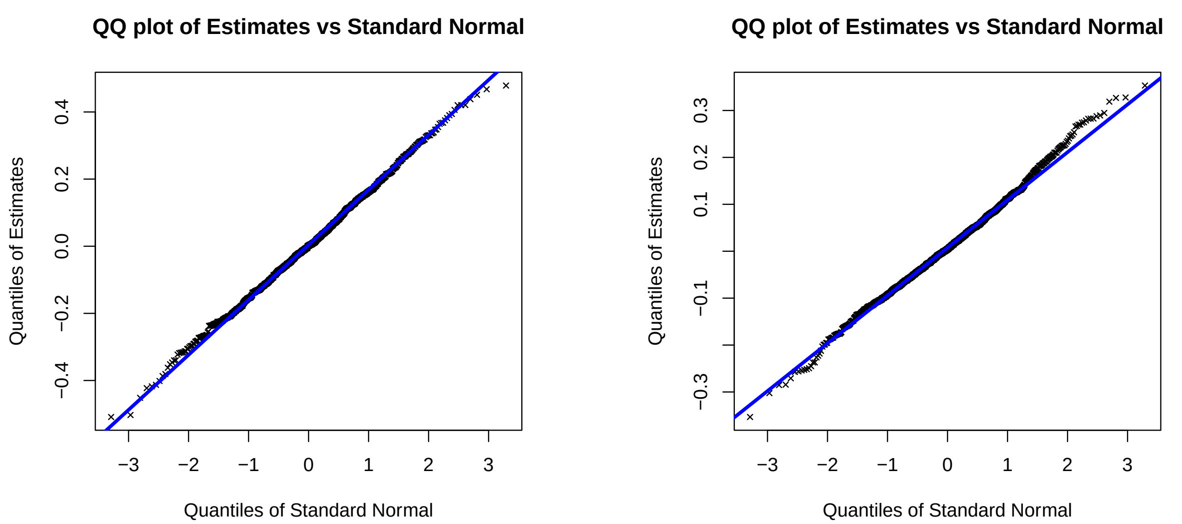


Figure 5.4: Multivariate Model A: QQ plot for $\beta = 0, N = 100$ (Left) and $N = 200$ (Right)

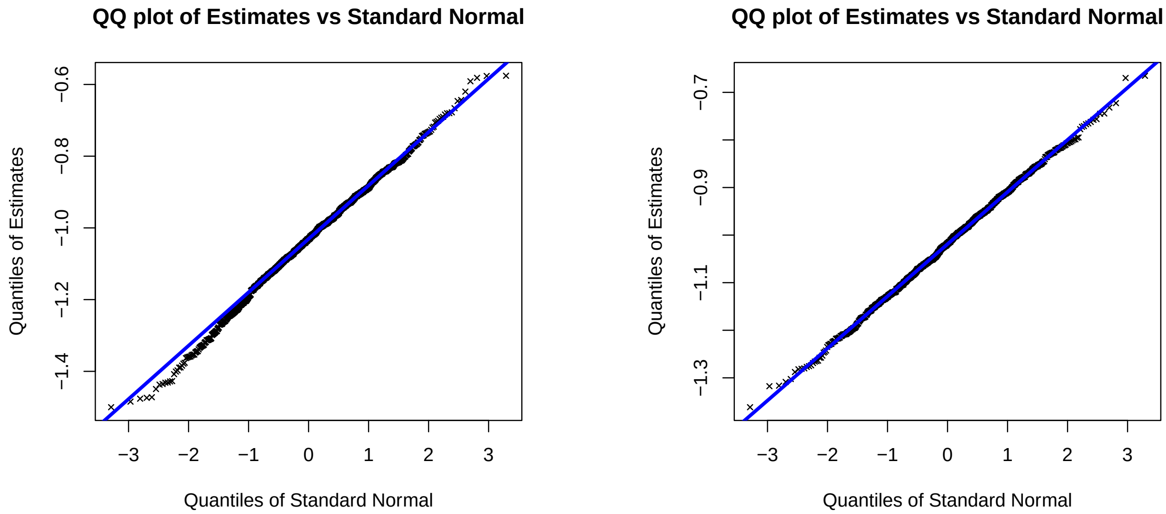


Figure 5.5: Multivariate Model A: QQ plot for $\beta = -1, N = 100$ (Left) and $N = 200$ (Right)

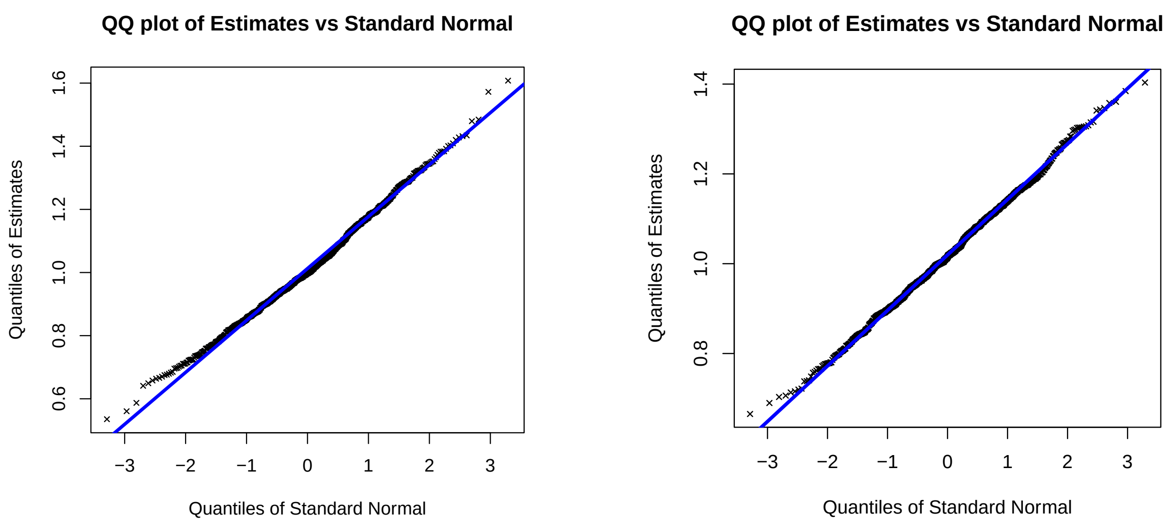


Figure 5.6: Multivariate Model A: QQ plot for $\beta = 1, N = 100$ (Left) and $N = 200$ (Right)

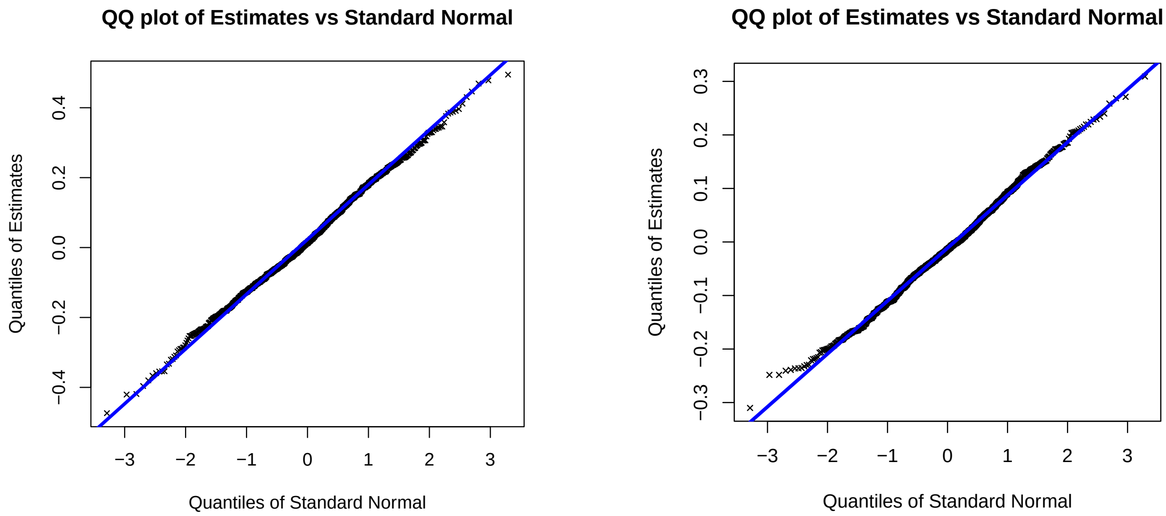


Figure 5.7: Multivariate Model B: QQ plot for $\beta = 0$, $N = 100$ (Left) and $N = 200$ (Right)

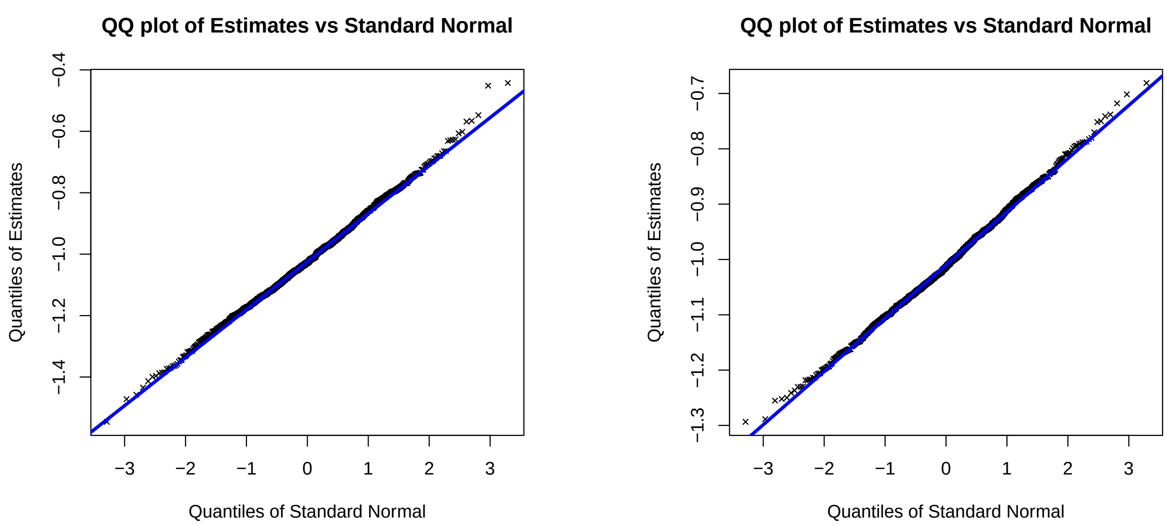


Figure 5.8: Multivariate Model B: QQ plot for $\beta = -1$, $N = 100$ (Left) and $N = 200$ (Right)

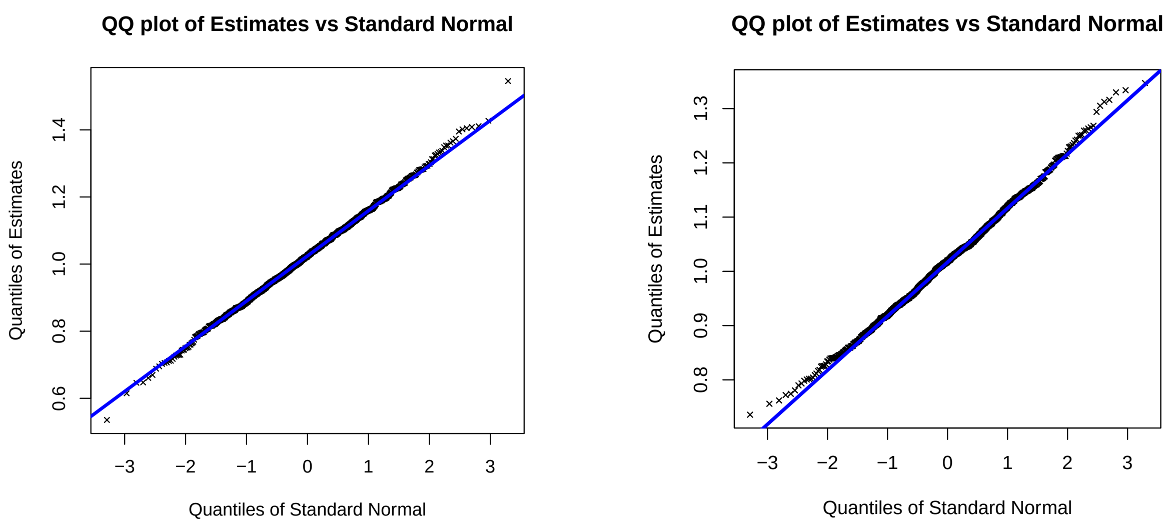


Figure 5.9: Multivariate Model B: QQ plot for $\beta = 1, N = 100$ (Left) and $N = 200$ (Right)

Chapter 6: Application: The Skin Cancer Chemoprevention Trial

6.1 Study Background

In this chapter, our proposed methodology is applied to the bivariate panel count data on the occurrences of the two types of related non-melanoma skin cancers, basal cell carcinoma and squamous cell carcinoma, from the skin cancer chemoprevention trial, which was conducted by the University of Wisconsin Comprehensive Cancer Center. It was a double-blinded and placebo-controlled randomized phase III clinical trial. The primary objective of this trial was to evaluate the effectiveness of 0.5 $g/m^2/day$ PO difluoromethylornithine (DFMO) in reducing new skin cancers in a population of the patients with history of non-melanoma skin cancers (i.e., basal cell carcinoma and squamous cell carcinoma).

For each patient, the numbers of occurrences of both basal cell carcinoma and squamous cell carcinoma and their correpsonding observation times were recorded. The observation times and follow-up times varied from patient to patient. In this dataset, there exist two types of recurrent events defined by the two types of skin cancers. The bivariate panel count data for the occurrences of these two types of skin cancers were possibly correlated.

This study involved 291 skin cancer patients in total with one removal, as one patient did not give any observation. The remaining 290 patients were randomized into a treatment group with 143 patients and a placebo group with 147 patients. Among these patients, the number of observations ranged from 1 to 17. With respect to the number of the recurrent events, the number of new basal cell carcinomas ranged from 0 to 16, while the number of new squamous cell carcinomas ranged from 0 to 23. For each patient, information was also available on three baseline covariates, gender, age at diagnosis, and the number of prior skin cancers from the first diagnosis to randomization. For the analysis, basal cell carcinoma and squamous cell carcinoma, were treated as two types of recurrent events of interest. In addition to the treatment indicator, the number of initial tumors was included in the analysis. Table 6.6 and Table 6.7 also presents the results for multivariate model A and model B when including the number of initial tumors, gender and age in the analysis.

6.2 Statistical Analysis

Following the joint modeling procedure discussed in Chapter 2 and Chapter 3, we first set up the EM-algorithm by initializing the $\beta's$ and $\gamma's$ and the bivariate normal distribution of the subject-specific frailty z_i and u_i with mean $(0,0)$ and a pre-specified variance-covariance matrix. The selection of the initial values of $\beta's$ and $\gamma's$ and the initial values of the parameters of the variance-covariance matrix was shown to have no impact on the convergence by simulation studies. The iterative process of the EM-algorithm was terminated if the following criteria were met:

$$|\hat{\beta}_{(n)} - \hat{\beta}_{(n-1)}| < \epsilon$$

and

$$|\hat{\gamma}_{(n)} - \hat{\gamma}_{(n-1)}| < \epsilon,$$

where ϵ is set to 0.001 in this case for optimal convergence. After the point estimates of the covariate effects have been obtained, the second step is to apply the bootstrap approach with a size of 1000 bootstrap samples for the estimation of the variance of the point estimates. A 95% confidence interval of each regression parameter can then be calculated.

6.3 Model Checking

All of the plots for model checking are presented in Appendix B. Figure B.1 and Figure B.2 plot the cumulative martingale residuals against the initial number of tumors with both the univariate model of basal cell carcinoma and the univariate model of squamous cell carcinoma, respectively. The pattern of both plots suggests that the functional forms of the covariate are satisfactory, and the p-value of the supremum test are both greater than 0.2.

Figure B.3 through Figure B.8 plot the cumulative martingale residuals for each skin cancer cell against the initial number of tumors for all of the three types of multivariate models: Model A, Model B, and Model C. All the plots indicate that the functional forms of the covariate are very satisfactory in all the multivariate models, and the p-values of the supremum tests are all close to 1.

6.4 Results and Discussion

The results for the univariate model are given in Table 6.1 for basal cell carcinoma and Table 6.2 for squamous cell carcinoma. The results for each of the three multivariate models are given in Table 6.3 through Table 6.5. Those tables include the point estimates of regression parameters, the estimated standard errors of the estimates, and the estimated 95% confidence intervals (CI). The estimated standard errors of $\hat{\beta}'s$ were obtained by the means of estimates for the bootstrap samples of size 1000. Tables 6.1 and 6.2 give essentially the same conclusion that the DFMO treatment reduced the recurrences of the skin cancer, but the treatment effect was not significant. However, the initial number of prior skin cancers was significantly positively associated with the recurrences of both basal and squamous cell carcinoma. It is worth noting that the standard error of $\hat{\beta}$ from Model B (Table 6.4) is slightly smaller than that from Model A (Table 6.3), which suggests an small increase in efficiency when modeling the correlation separately by event type. The results for Model C (Table 6.5) indicates that the DFMO treatment was more effective in reducing the recurrences of new basal cell carcinoma in this study. In addition, the number of initial tumors had a significantly positive effect on the recurrences of both types of skin cancers. This results discussed above are generally consistent with those in Li et al. (2011).

Table 6.1: Estimation of regression parameters for the skin cancer chemoprevention trial under univariate model for basal cell carcinoma

	DFMO treatment	**Prior skin cancer number**
$\hat{\beta}$	-1.1772	0.0287
SE($\hat{\beta}$)	0.2581	0.0053
95% CI	(-1.6830, -0.6714)	(0.0183, 0.0391)

Table 6.2: Estimation of regression parameters for the skin cancer chemoprevention trial under univariate model for squamous cell carcinoma

	DFMO treatment	**Prior skin cancer number**
$\hat{\beta}$	-0.8671	0.0195
SE($\hat{\beta}$)	0.5490	0.0315
95% CI	(-1.9431, 0.2090)	(-0.0423, 0.0814)

Table 6.3: Estimation of regression parameters for the skin cancer chemoprevention trial under Model A (Σ is constant)

	DFMO treatment	**Prior skin cancer number**
$\hat{\beta}$	-0.1196	0.0911
SE($\hat{\beta}$)	0.1701	0.0148
95% CI	(-0.4530, 0.2138)	(0.0621, 0.1201)

Table 6.4: Estimation of regression parameters for the skin cancer chemoprevention trial under Model B (Σ is event-type-specific)

	DFMO treatment	**Prior skin cancer number**
$\hat{\beta}$	-0.1685	0.0807
SE($\hat{\beta}$)	0.1497	0.0118
95% CI	(-0.4619, 0.1249)	(0.0576, 0.1038)

Table 6.5: Estimation of regression parameters for the skin cancer chemoprevention trial under Model C (event-type-specific $\beta's$ with a constant Σ)

	DFMO treatment		**Prior skin cancer number**	
	Basal Cell	Squamous Cell	Basal Cell	Squamous Cell
$\hat{\beta}$	-0.4082	0.1221	0.0744	0.1018
SE($\hat{\beta}$)	0.1456	0.1522	0.0156	0.0110
95% CI	(-0.6936, -0.1228)	(-0.1762, 0.4205)	(0.0438, 0.1051)	(0.0801, 0.1234)

Table 6.6: Estimation of regression parameters for the skin cancer chemoprevention trial under Model A with three covariates

	DFMO treatment	**Prior skin cancer number**	**Gender**	**Age**
$\hat{\beta}$	-0.0454	0.1037	0.1767	-0.0160
SE($\hat{\beta}$)	0.1832	0.0150	0.1581	0.0069
95% CI	(-0.4044, 0.3136)	(0.0743, 0.1331)	(-0.1330, 0.4865)	(-0.0295, -0.0025)

Table 6.7: Estimation of regression parameters for the skin cancer chemoprevention trial under Model B with three covariates

	DFMO treatment	**Prior skin cancer number**	**Gender**	**Age**
$\hat{\beta}$	-0.0709	0.1240	0.1083	-0.0128
SE($\hat{\beta}$)	0.1446	0.0159	0.1485	0.0056
95% CI	(-0.3544, 0.2126)	(0.0929, 0.1551)	(-0.1827, 0.3993)	(-0.0237, -0.0019)

Chapter 7: Conclusions and Future Research

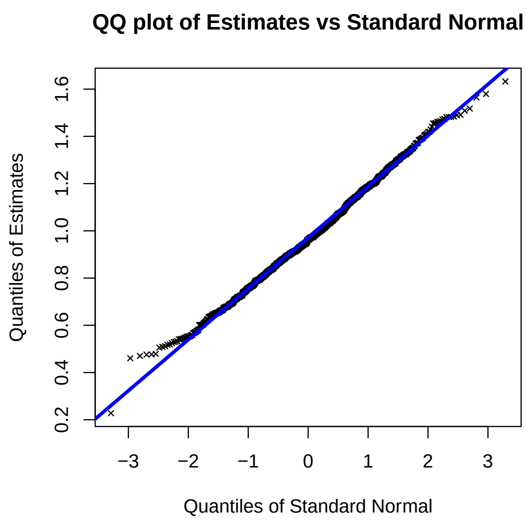
In the preceding chapters, we have discussed semiparametric regression analysis of univariate and multivariate panel count data for the underlying recurrent event processes of interest by incorporating an informative observation time process on study subjects. Such analyses can be applied to a wide variety of biomedical studies when data are collected with several related types of recurrent events and the observation processes vary from subject to subject. To deal with this problem, we presented multiple types of multivariate models for the variance-covariance matrix between two processes and a joint modeling inference procedure for the estimation of parameters. The finite sample and asymptotic properties of the proposed estimates have been validated through simulation studies. The proposed methodology was applied to the multivariate analysis of the skin cancer chemoprevention trial, and the results were consistent with the findings from previous literature. Martingale-residual based model checking techniques were developed, and the appropriateness of the functional forms of covariates and the objectivity of the residual plots were evaluated and assessed for each proposed model.

An advantage of our approach is that the variance-covariance matrix between the observation time process and the underlying recurrent event process of interest can be modeled as constant across different event types or event-type specific, depending on the research interest or prior knowledge of those related types of recurrent events. In addition, multivariate model C allows event-type-specific covariate effects with the assumption of a constant variance-covariance matrix between the two processes.

There exists several directions for future research. The first is that both processes in the proposed methodology rely on the assumption of nonhomogeneous Poisson processes. In practice, this may not be always true as it may be hard to justify that the variance of $N_i(t)$ appears to be the same as the expectation of $N_i(t)$, which are assumed to be identical under a Poisson model. Though the incorporation of unobservable random effects in the proposed model helps accounting for those interindividual variation in the event occurrence, a generalization of the proposed methodology to the situations where the Poisson assumption does not hold would be useful.

The second direction is that the follow-up time process has been assumed to be non-informative in the proposed approach. A possible direction is to develop a new model that allows the follow-up time to be related with the observation process as well as the underlying recurrent event process of interest. This situation can occur if the disease is severe with possibly high dropout rate during the follow-up period. However, the joint modeling of three processes may result in multi-step inference procedures, which are usually quite time-consuming and may reduce the overall efficiency of the estimators.

Appendix A: QQ Plot for Multivariate Case C



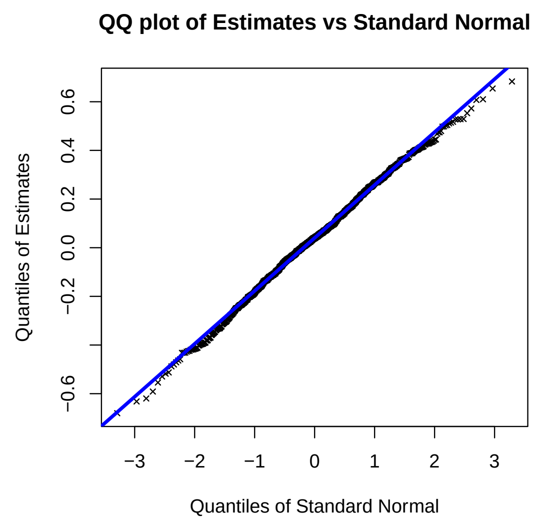


Figure A.1: QQ Plot for $\beta_1 = 0, \beta_2 = 1, N = 100$: plot for β_1 (Left) and β_2 (Right)

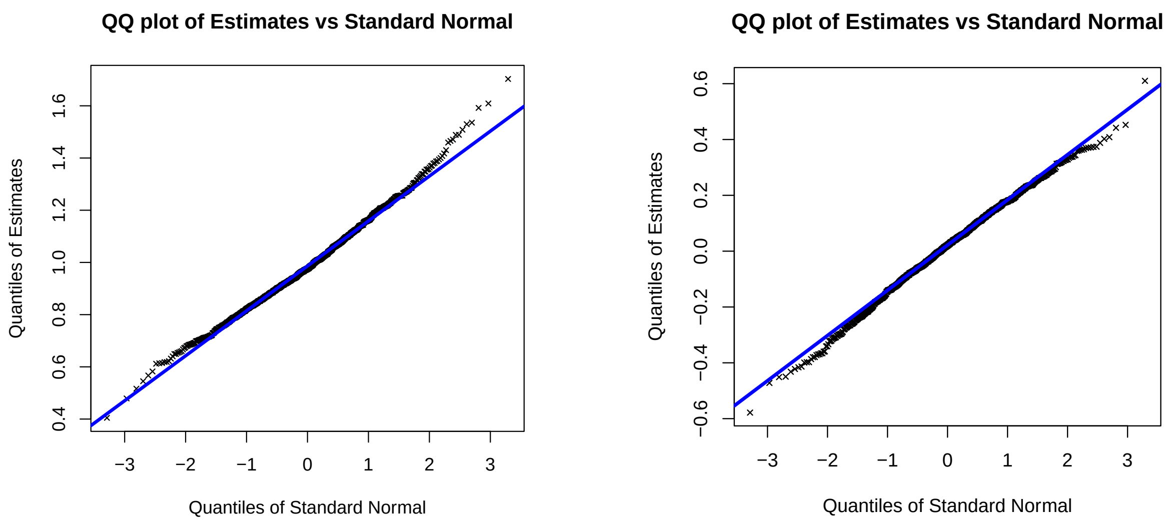


Figure A.2: QQ Plot for $\beta_1 = 0, \beta_2 = 1, N = 200$: plot for β_1 (Left) and β_2 (Right)

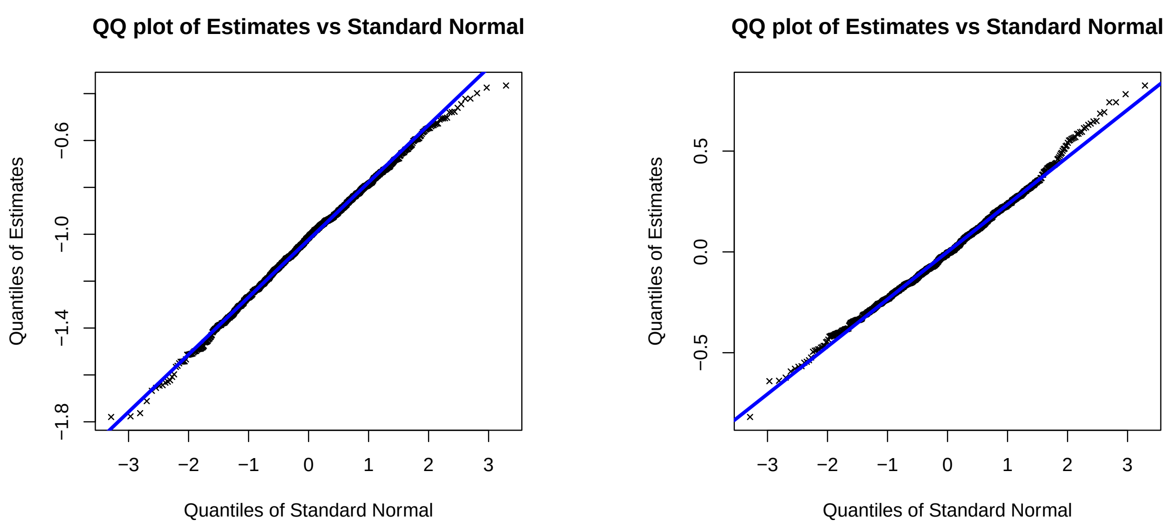


Figure A.3: QQ Plot for $\beta_1 = 0, \beta_2 = -1, N = 100$: plot for β_1 (Left) and β_2 (Right)

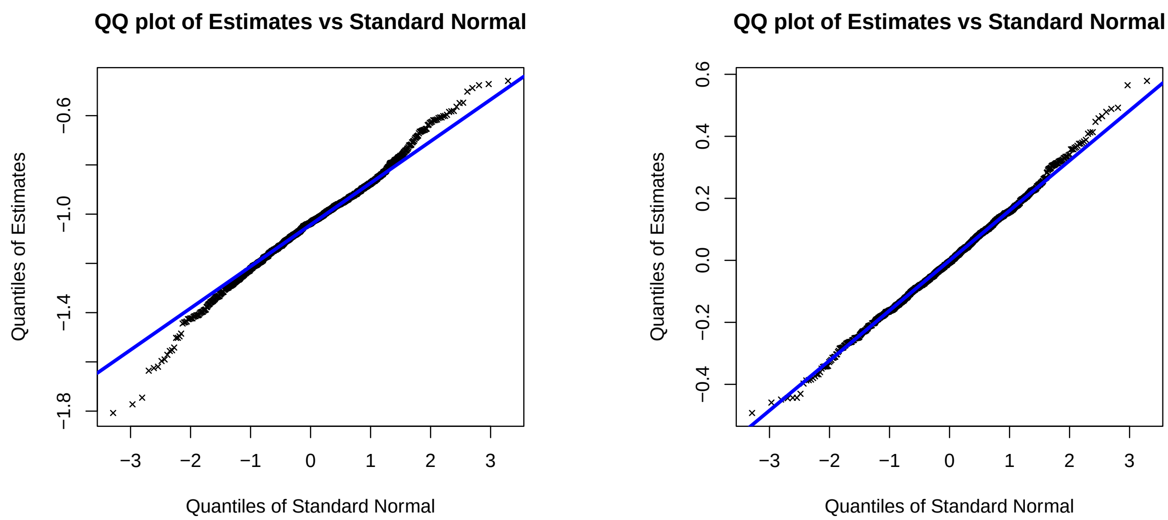


Figure A.4: QQ Plot for $\beta_1 = 0, \beta_2 = -1, N = 200$: plot for β_1 (Left) and β_2 (Right)

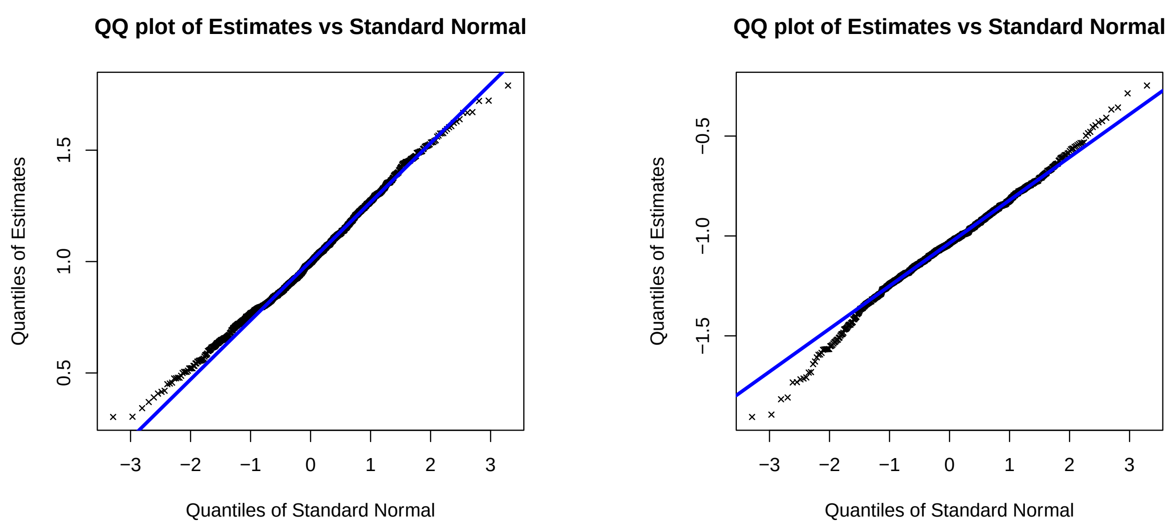
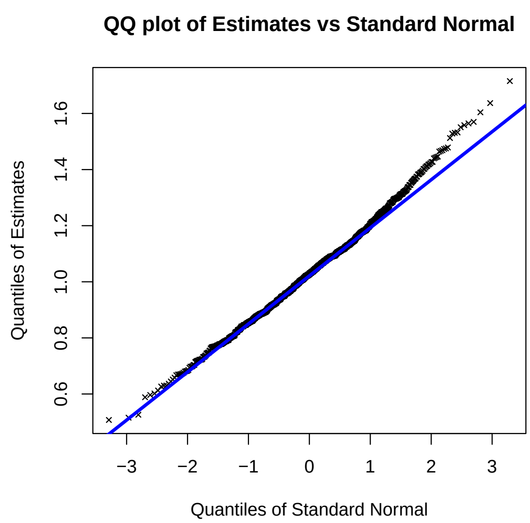


Figure A.5: QQ Plot for $\beta_1 = 1, \beta_2 = -1, N = 100$: plot for β_1 (Left) and β_2 (Right)



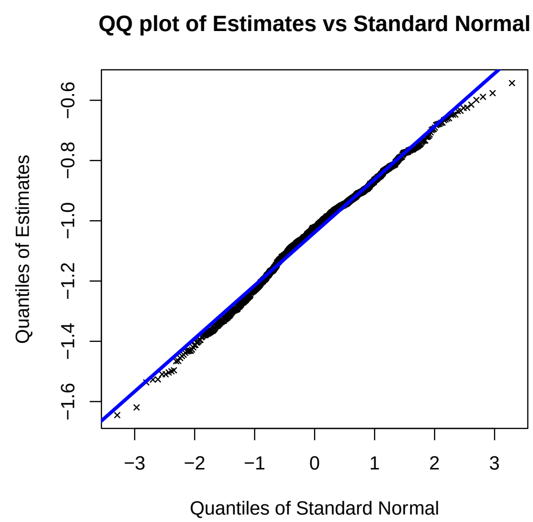


Figure A.6: QQ Plot for $\beta_1 = 1, \beta_2 = -1, N = 200$: plot for β_1 (Left) and β_2 (Right)

Appendix B: Figures of Model Checking

B.1 Univariate Models

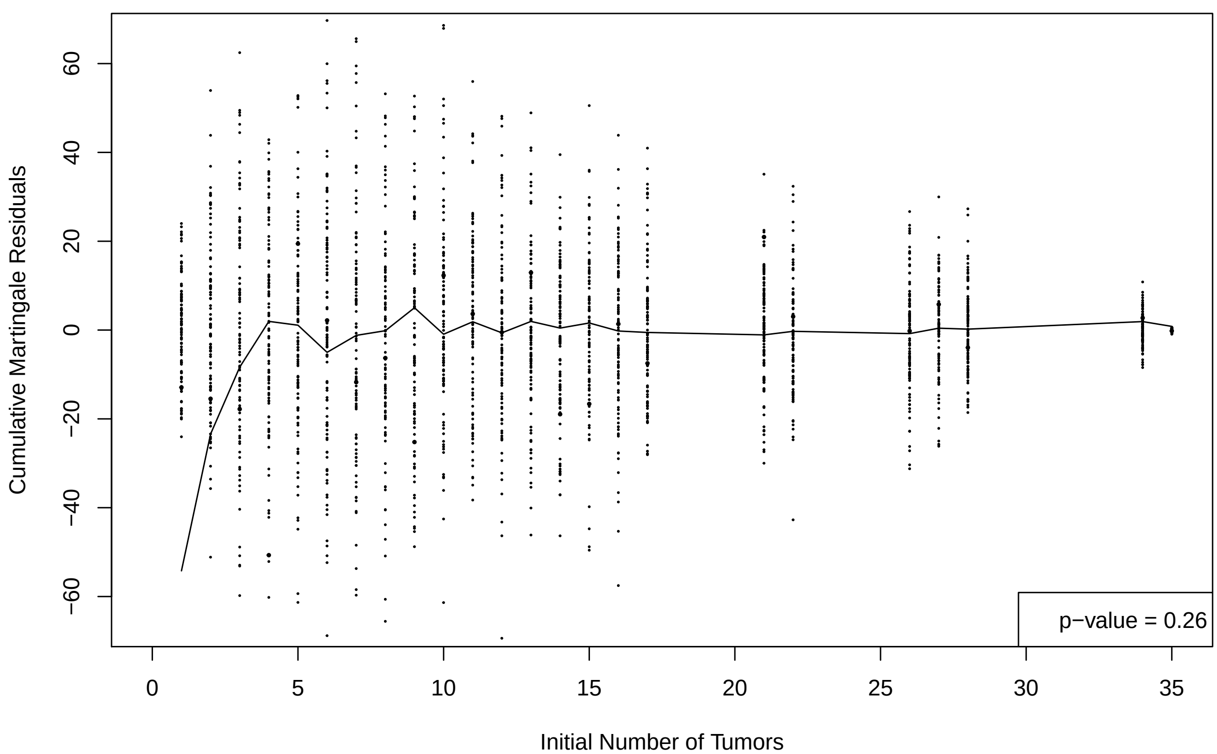


Figure B.1: Plot of cumulative martingale residuals in the Univariate Model for basal cell carcinoma.

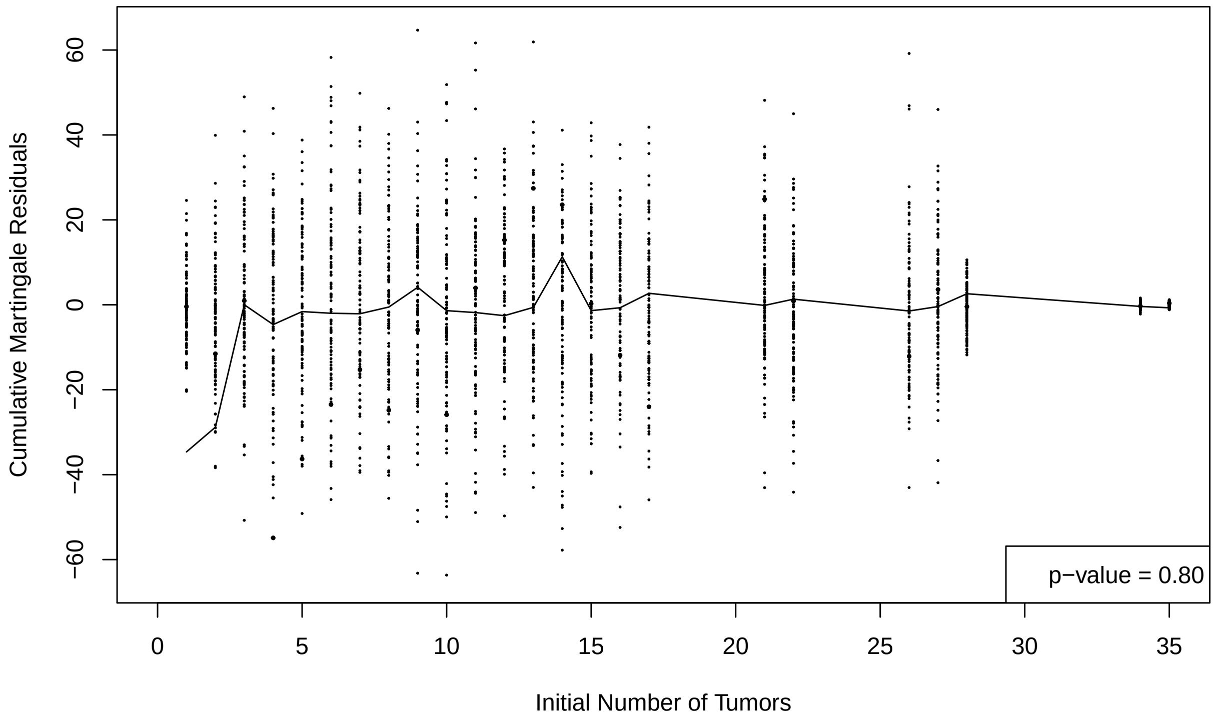


Figure B.2: Plot of cumulative martingale residuals in the Univariate Model for squamous cell carcinoma.

B.2 Multivariate Models

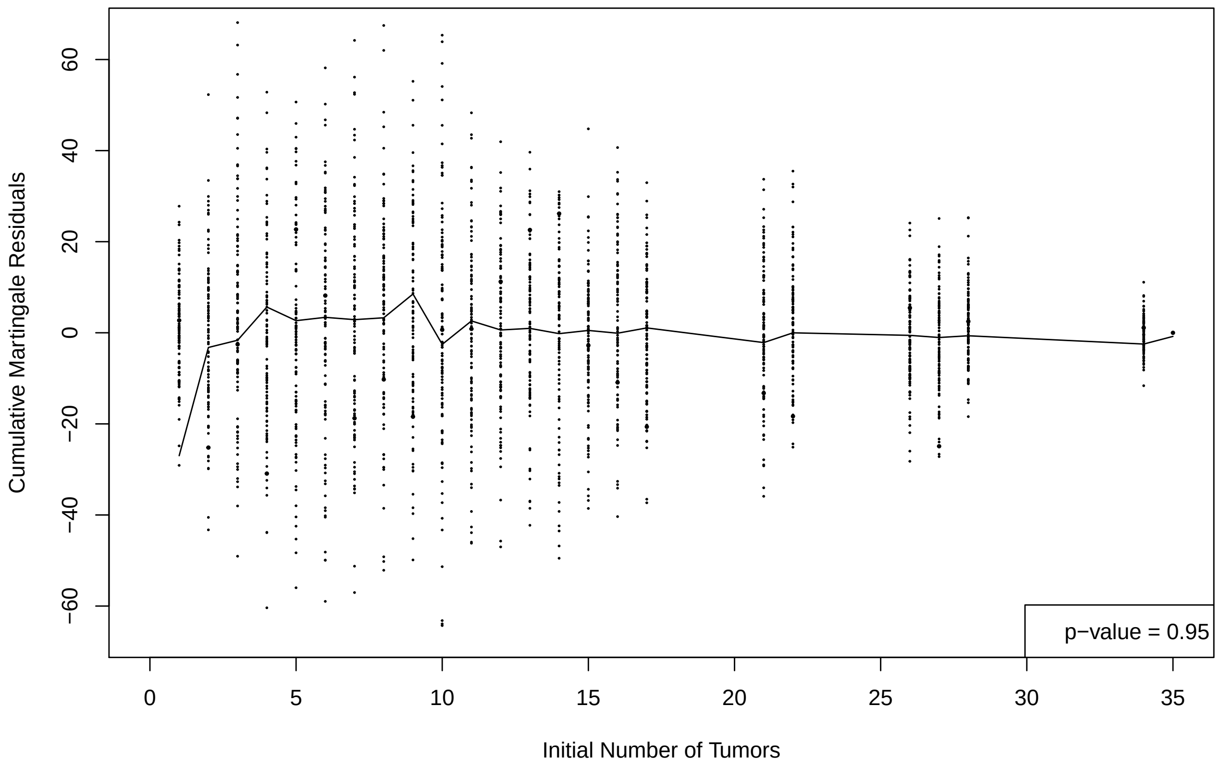


Figure B.3: Plot of cumulative martingale residuals in Multivariate Model A for basal cell carcinoma.

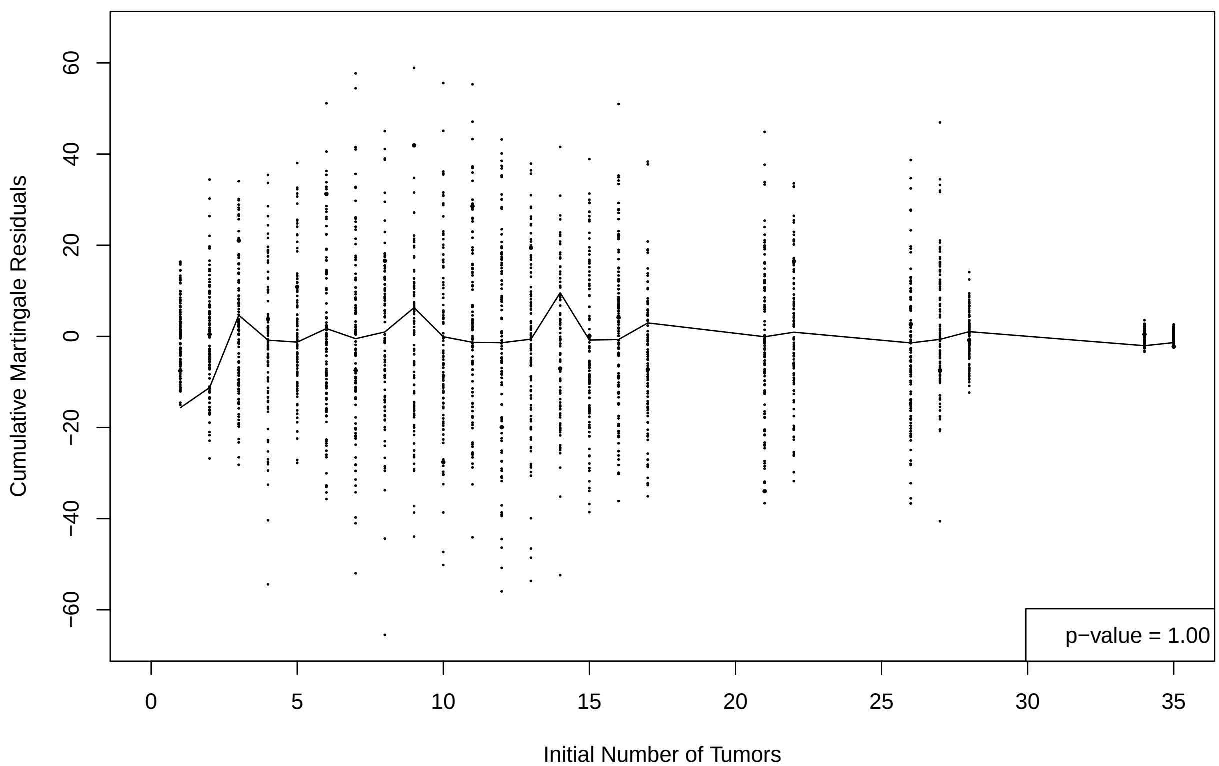


Figure B.4: Plot of cumulative martingale residuals in Multivariate Model A for squamous cell carcinoma.

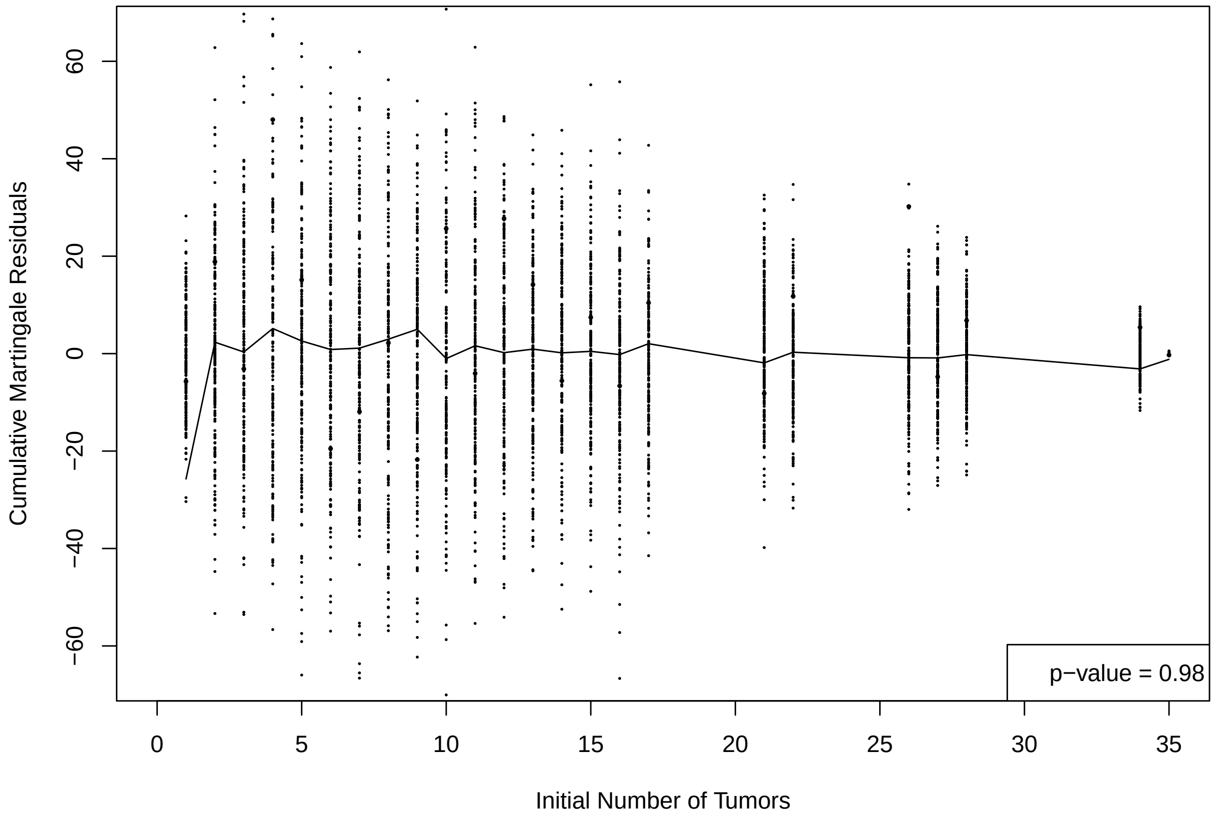


Figure B.5: Plot of cumulative martingale residuals in Multivariate Model B for basal cell carcinoma.

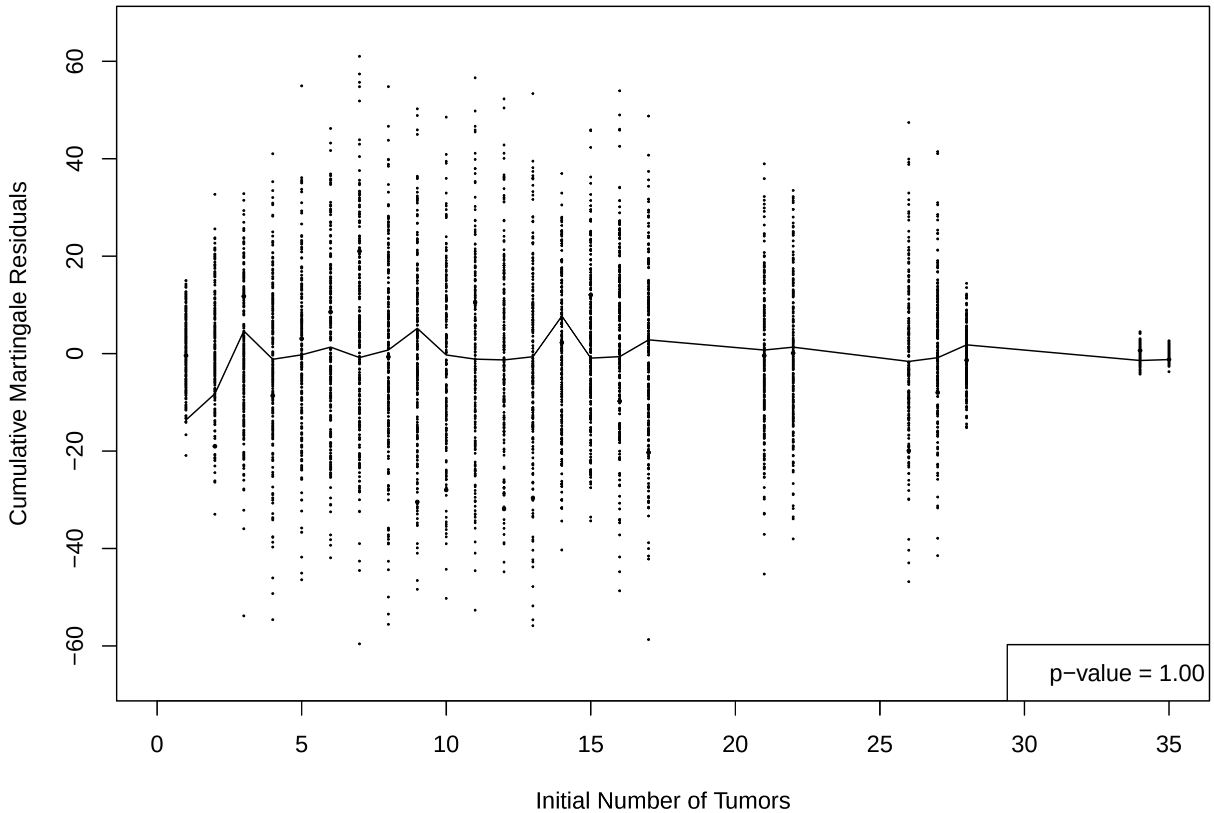


Figure B.6: Plot of cumulative martingale residuals in Multivariate Model B for squamous cell carcinoma.

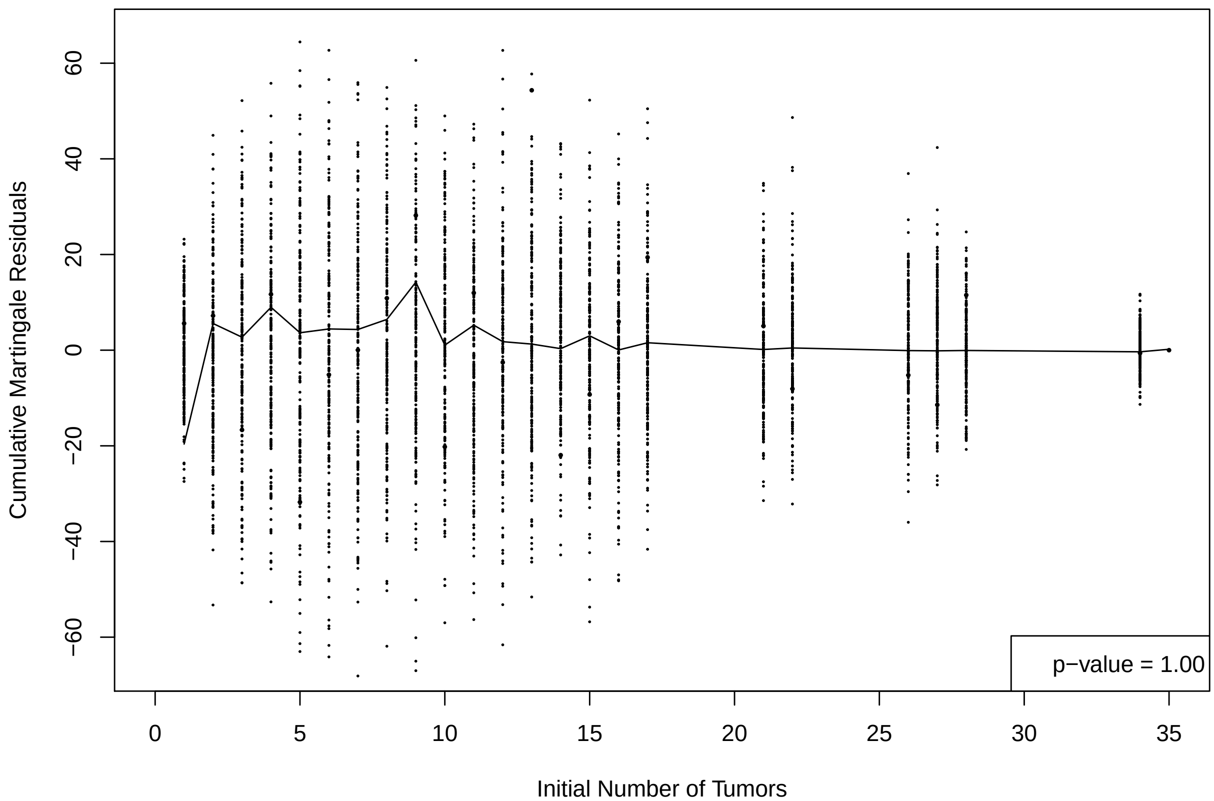


Figure B.7: Plot of cumulative martingale residuals in Multivariate Model C for basal cell carcinoma.

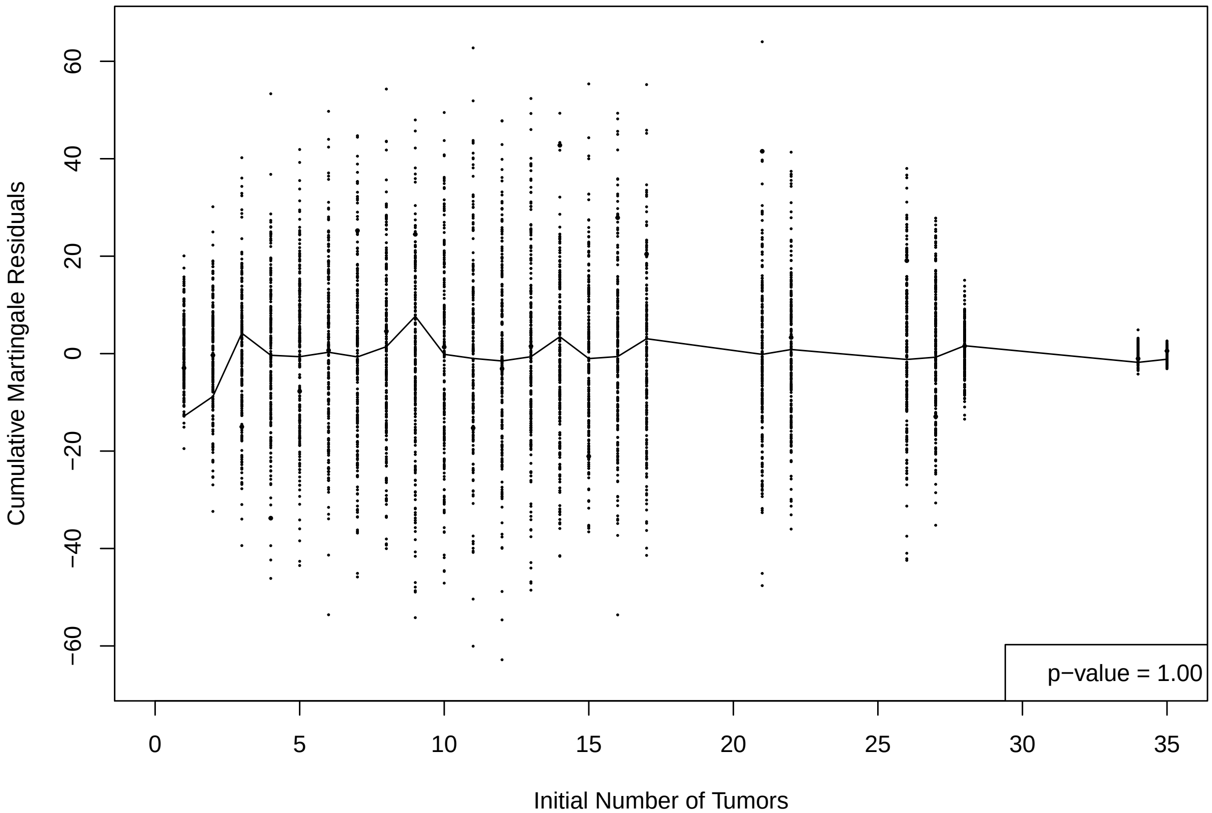


Figure B.8: Plot of cumulative martingale residuals in Multivariate Model C for squamous cell carcinoma.

Bibliography

[1] Andersen, Per Kragh, and Richard D. Gill. (1982), Cox's regression model for counting processes: a large sample study. *The Annals of Statistics*, 1100-1120.

[2] Bean, S. J. and Tsokos, C. P. (1980), Developments in non-parametric density estimation. *Int. Statist. Rev.*, 48, 267-287.

[3] Byar, D. P. (1980), The veterans administration study of chemoprophylaxis for recurrent stage i bladder tumours: comparisons of placebo, pyridoxine and topical thiotepa. *Bladder Tumors and Other Topics in Urological Oncology*, Springer, Boston, MA, 363-370.

[4] Cheng, S. C., and L. J. Wei. (2000), Inferences for a semiparametric model with panel data. *Biometrika*, 87.1, 89-97.

[5] Cook, Richard J., and Jerald Lawless. (2007), *The statistical analysis of recurrent events*, Springer Science and Business Media.

[6] Cox, D. R. (1972), Regression models and life tables (with discussion). *J. Roy. Statist. Soc. B*, 34, 187-220.

[7] Fan, Jianqing, and Runze Li. (2002), Variable selection for Cox's proportional hazards model and frailty model. *The Annals of Statistics*, 30.1, 74-99.

[8] 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.

[9] He, X., Tong, X., Sun, J. and Cook, R.J. (1982), Regression analysis of multivariate panel count data. *Biostatistics*, 9.2, 234-248.

[10] He, Xin, Xingwei Tong, and Jianguo Sun. (2009), Semiparametric analysis of panel count data with correlated observation and follow-up times. *Lifetime Data Analysis*, 15.2, 177-196.

[11] 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.

[12] Hu, X. Joan, Stephen W. Lagakos, and Richard A. Lockhart. (2009), Marginal analysis of panel counts through estimating functions. *Biometrika*, 96.2, 445-456.

[13] Jongbloed, G. (1998), The iterative convex minorant algorithm for nonparametric estimation *J. Comput. Graph. Statist*, 7, 310-321.

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

[15] Li, N., Park, D. H., Sun, J. and Kim, K. (2011), Semiparametric transformation models for multivariate panel count data with dependent observation process. *Canadian Journal of Statistics*, 39.3, 458-474.

[16] Liang, Ye, Yang Li, and Bin Zhang. (2018), Bayesian nonparametric inference for panel count data with an informative observation process. *Biometrical Journal*, 60.3, 583-596.

[17] Lin, Danyu Y., Lee-Jen Wei, and Zhiliang Ying. (1993), Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika*, 80.3, 557-572.

[18] Nelson, W. (1995), Confidence limits for recurrence data: Applied to cost or number of product repairs. *Technometrics*, 37, 147-157.

[19] Ramlau-Hansen, Henrik. (1983), Smoothing counting process intensities by means of kernel functions. *The Annals of Statistics*, 453-466.

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

[21] Sun, Jianguo, and L. J. Wei. (2000), Regression analysis of panel count data with covariate-dependent observation and censoring times. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62.2, 293-302.

[22] Sun, Jianguo, Xingwei Tong, and Xin He. (2007), Regression analysis of panel count data with dependent observation times. *Biometrics*, 63.4, 1053-1059.

[23] Sun, J. and Zhao, X. (2013), *Statistical Analysis of Panel Count Data*, Springer, New York.

[24] Ullah, Shahid, Tim J. Gabbett, and Caroline F. Finch. (2014), Statistical modelling for recurrent events: an application to sports injuries. *British Journal of Sports Medicine*, 48.17, 1287-1293.

[25] Wand, M. P. and Jones, M. C. (1995), *Kernel smoothing Chapman & Hall*, CRC/London.

[26] Wang, Mei-Cheng, Jing Qin, and Chin-Tsang Chiang. (2001), Analyzing recurrent event data with informative censoring. *Journal of the American Statistical Association*, 96.455, 1057-1065.

[27] Wedel, M., Desarbo, W. S., Bult, J. R. and Ramaswamy, V. (1993), A latent class Poisson regression model for heterogeneous count data. *Journal of Applied Econometrics*, 8, 397-411.

[28] Wei, Lee-Jen, Danyu Y. Lin, and Lisa Weissfeld. (1989), Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, 84.408, 1065-1073.

[29] Wellner, Jon A., and Ying Zhang. (1998), Large sample theory for an estimator of the mean of a counting process with panel count data. *Technical Report*.

[30] 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.

[31] Jon A. Wellner, and Ying Zhang. (2007), Two Likelihood-Based Semiparametric Estimation Methods for Panel Count Data with Covariates. *The Annals of Statistics*, 35.5, 2106-2142.

[32] Yao, Bin, Lianming Wang, and Xin He. (2016), Semiparametric regression analysis of panel count data allowing for within-subject correlation. *Computational Statistics and Data Analysis*, 97, 47-59.

[33] Zeng, Donglin, and D. Y. Lin. (2021), Maximum likelihood estimation for semiparametric regression models with panel count data. *Biometrika*, 108.4, 947-963.

[34] Zhang, H., Zhao, H., Sun, J., Wang, D., and Kim, K. (2013), Regression analysis of multivariate panel count data with an informative observation process. *Journal of Multivariate Analysis*, 119, 71-80.