Evaluating Time-varying Effect in Single-type and Multi-type Semi-parametric Recurrent Event Models

Chen Chen

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Statistics

Feng Guo, Chair
Pang Du
Yili Hong
Marion R. Reynolds, Jr.

September 10, 2015
Blacksburg, Virginia

Key Words: Frailty Model, Generalized Linear Mixed Model, Multi-type Recurrent Event, Naturalistic Driving Study, Penalized B-Spline, Proportional Intensity Function, Stratification, Time-varying Coefficient, Transportation Safety.

Evaluating Time-varying Effect in Single-type and Multi-type Semi-parametric Recurrent Event Models

Abstract

This dissertation aims to develop statistical methodologies for estimating the effects of time-fixed and time-varying factors in recurrent events modeling context. The research is motivated by the traffic safety research question of evaluating the influence of crash on driving risk and driver behavior. The methodologies developed, however, are general and can be applied to other fields. Four alternative approaches based on various data settings are elaborated and applied to 100-Car Naturalistic Driving Study in the following Chapters.

Chapter 1 provides a general introduction and background of each method, with a sketch of 100-Car Naturalistic Driving Study. In Chapter 2, I assessed the impact of crash on driving behavior by comparing the frequency of distraction events in per-defined windows. A count-based approach based on mixed-effect binomial regression models was used.

In Chapter 3, I introduced intensity-based recurrent event models by treating number of Safety Critical Incidents and Near Crash over time as a counting process. Recurrent event models fit the natural generation scheme of the data in this study. Four semi-parametric models are explored: Andersen-Gill model, Andersen-Gill model with stratified baseline functions, frailty model, and frailty model with stratified baseline functions. I derived model estimation procedure and and conducted model comparison via simulation and application.

The recurrent event models in Chapter 3 are all based on proportional assumption, where effects are constant. However, the change of effects over time is often of primary interest. In Chapter 4, I developed time-varying coefficient model using penalized B-spline function to approximate varying coefficients. Shared frailty terms was used to incorporate correlation within subjects. Inference and statistical test are also provided. Frailty representation was proposed to link time-varying coefficient model with regular frailty model.

In Chapter 5, I further extended framework to accommodate multi-type recurrent events with time-varying coefficient. Two types of recurrent-event models were developed. These models incorporate correlation among intensity functions from different type of events by correlated frailty terms. Chapter 6 gives a general review on the contributions of this dissertation and discussion of future research directions.

To my family.

Acknowledgments

First and foremost I would like to thank my advisor Dr.Feng Guo, for his illuminating guidance, generous support, and invaluable encouragement. It has been an honor to be his Ph.D. student. I am fortunate enough to have the opportunity to work with him. I sincerely appreciate my committee members, Drs. Pang Du, Yili Hong, and Marion R. Reynolds, Jr., for serving on my committees, for direction, feedback, and assistance.

I would also like to thank Dr. Jeffrey B. Birch, for his valuable advice for my teaching skills and giving constant help and encouragement. It's been a great journey to work with students from various disciplines and help them understand and practice statistic. I want to thank Dr. Eric A. Vance and Dr. Christopher T. Franck particularly. My skills in statistical collaboration and consulting has been greatly improved while working for LISA.

Thanks to all the professors, including but not limit to, Drs. John P. Morgan, Inyoung Kim, Xinwei Deng, and Scotland C. Leman for their inspiring courses and support. I would also like to thank all of the staff of the Department of Statistics, especially Tonya Pruitt and Betty Higginbotham, for their generous help.

My thanks also go to all the talented graduate students in Dr. Guo's research group, Youjia Fang, Qing Li, Yi Liu, Dengfeng Zhang, for the extensive collaboration and discussion on statistical issues that widened my vision. Thank you all my friends.

Lastly, my most special thanks go to my parents and my husband, Yiming Peng, for their endless love and support. None of this would be possible without my family.

Contents

List of Figures viii

List of Tables ix

1 Introduction 1

1.1 Background . . . 1

1.1.1 Introduction to Analysis of Recurrent Event Data . . . 1

1.1.2 Event Frequency Approach . . . 2

1.1.3 Time-to-event Data Approach . . . 2

1.1.4 Time-varying Coefficient Recurrent Event Model . . . 3

1.1.5 Multi-type Time-to-event Data . . . 4

1.2 Motivated Example: The 100-Car Naturalistic Driving Study . . . 4

1.3 Overview . . . 5

Bibliography 6

2 Evaluating the Impact of Crashes on Driving Behavior 8

2.1 Introduction . . . 8

2.2 Baseline Sampling Design and Data Collection . . . 10

2.3 Modeling Baseline Distraction Using Mixed Binomial Regression . . . 14

2.4 Discussion . . . 15

Bibliography 16

3 Assessing Influence of Crash on Driving Risk Using Semi-parametric Recurrent Events Model 18

3.1 Introduction . . . 18

3.2 100-Car NDS Data Setting for Recurrent Events Model . . . 19

3.3 Semi-Parametric Recurrent Events Models . . . 20

3.3.1 Andersen-Gill Model . . . 20

3.3.2 Stratified A-G model . . . 21

3.3.3 Shared Frailty Model . . . 22

3.3.4 Stratified Shared Frailty Model . . . 23

3.3.5 Model Fitting: Cox-Snell Residual . . . 23

3.4 Simulation Study . . . 24
3.4.1 Simulation setup . . . 24
3.4.2 Simulation results . . . 25
3.5 Application in 100-Car NDS . . . 28
3.6 Conclusion and Discussion . . . 33

Bibliography **34**

4 Inference on Semi-parametric Frailty Model with Time-varying Coefficient **36**
4.1 Introduction . . . 36
4.2 Time-varying coefficient model . . . 38
4.2.1 Penalized B-Spline estimation with alternative penalty matrices . . . 39
4.2.2 Double penalized partial likelihood and parameter estimation . . . 40
4.2.3 Stratified time-varying coefficient model . . . 42
4.3 Statistical inference . . . 42
4.3.1 Asymptotic distribution of maximum DPPL estimator . . . 42
4.3.2 The frailty model representation . . . 44
4.3.3 Inference on smoothing parameter and variance component . . . 45
4.3.4 Computation . . . 45
4.3.5 Test of time-varying coefficient . . . 46
4.4 Application to 100-Car Naturalistic Driving Study . . . 48
4.5 Simulation study . . . 49
4.5.1 Simulation setup . . . 50
4.5.2 Simulation result . . . 51
4.6 Conclusion and discussion . . . 57

Bibliography **59**

5 Multi-type Recurrent Event Model with Time-varying Coefficients **61**
5.1 Introduction . . . 61
5.2 Multi-type recurrent event model with time-varying coefficients . . . 62
5.2.1 Penalized B-Spline estimation . . . 64
5.2.2 Double penalized partial likelihood . . . 64
5.2.3 The frailty model representation . . . 65
5.2.4 Maximum DPPL estimation . . . 67
5.3 Statistical inference . . . 67
5.3.1 Inference on smoothing parameter and variance component . . . 67

5.3.2 Computation 69
5.3.3 Statistical Test 69
5.4 Application to 100-Car Naturalistic Driving Study 70
5.5 Simulation study 73
5.5.1 Simulation setup 73
5.5.2 Simulation result 74
5.6 Conclusion and discussion 79

Bibliography **81**

6 General Conclusions **83**
6.1 Conclusion and Contribution 83
6.2 Future Work 84

List of Figures

2.1 Baseline data collection. 11
2.2 Crash interval histogram. 12
2.3 Baseline sampling scheme. 12
2.4 Ratio of distraction rate after vs. before by gender. 13
2.5 Crash impact on distraction proportion. 15

3.1 Data collection structure. 20
3.2 Coverage probability comparison I. 25
3.3 Coverage probability comparison II. 26
3.4 Cox-Snell residual plots for SCI. 31
3.5 Cox-Snell residual plots for NC. 32
3.6 Baseline intensity rate estimation of SCI(left) and NC(right). 32

4.1 Data collection structure. 48
4.2 Crash influence on SCI and NC: $\widehat{\beta}(t)$, solid; 95% pointwise confidence interval, dashed; fixed effect estimation, dotted. 49
4.3 Simulation results for logistic function from stratified data 52
4.4 Simulation results for piecewise polynomial function from stratified data . . 52
4.5 Setting I: simulation results for two time-varying coefficient functions from stratified data 55

5.1 Data collection structure. 70
5.2 Crash influence on SCI and NC: $\widehat{\beta}(t)$, solid; 95% poitwise confidence interval, dashed. 72
5.3 Simulation result: positive correlation. 76
5.4 Simulation result: negative correlation. 77

List of Tables

1.1 Number of driver and driving time across groups in 100-Car NDS. 5

2.1 Definition of distraction. 10
2.2 Baseline sample distribution. 13

3.1 SCI/NC rate by groups 19
3.2 Simulation result I: k=(.95,1.1,1.22) c=(0.2,0.17,0.15). 27
3.3 Simulation result II: k=(1,1,1) c=(0.2,0.2,0.2). 27
3.4 Crash effect estimation on SCI. 28
3.5 Crash effect estimation on NC. 29
3.6 Distribution of extreme Cox-Snell residuals on SCI. 30

4.1 Parametric coefficient estimates of time-varying model applied to 100-Car NDS 50
4.2 Simulation results for parametric coefficients estimates in time-varying coefficient model 53
4.3 Empirical power/ type I error of tests for covariate effects 54
4.4 Simulation results for parametric coefficients estimates in two effects model . 56
4.5 Empirical power/ type I error of tests for covariate effects 57

5.1 Parametric coefficient estimates of 100-Car NDS 73
5.2 Simulation results for parametric coefficients estimates in one time-varying effect model 78
5.3 Empirical power/ type I error of tests for time-varying and time-fixed effects 79

Chapter 1 Introduction

1.1 Background

1.1.1 Introduction to Analysis of Recurrent Event Data

Recurrent event models focus on situations where events occur more than once per subject over follow-up time. It arises in fields such as medicine, public health, social sciences, reliability, and transportation safety research. For example, the repeated occurrence of heart attacks can be treated as recurrent events. Another example is the time to multiple failures of a product. The system which generating such data are referred as recurrent event process and can be analyzed by counting process models.

The literature on the statistical analysis of recurrent events has grown rapidly and a variety of methods has been developed. The approaches can be classified into two major categories based on the type of outcomes: frequency and time-to-event. For frequency approach, Poisson or negative binomial regression model is often used to evaluate if the rates of events over a predetermined study period differ by treatment [13]. Longitudinal modeling methods can be used to to discover pattern over time by dividing study period into non-overlapping intervals. In this setup, each subject is associated with multiple observations. The correlation within subject is taken into account by either generalized linear mixed models or generalized estimating equations (GEE) models[11].

Time-to-event data include not only frequency but also the time to each occurrence. Event occurrences are treated as a realization of counting process and modeled through by the intensity function. The relationship between covariates or treatment (fixed or time-varying) to event occurrence can be evaluated by comparing corresponding intensities. Objectives in analyzing recurrent data also include identifying variation across a population of processes after adjusted by potential covariates.

The frequency approach and the time-to-event approach are closely related with each other. Long time interval associates with low rate/risk and thus fewer events in a fixed time window. Short intervals correspond to high risk and more events are expected. As a matter of fact, frequency data models have assumptions of underlying distribution of time between events. For example, the standard Poisson regression has the underlying assumption that time intervals are exponentially distributed.

1.1.2 Event Frequency Approach

Poisson and negative binomial regression models are two widely applied approaches for modeling event frequency. They are natural fit for count observations.

For Poisson regression models, the probability of subject i having y_i events per time unit is given by:

$$Pr(y_i) = \exp(-\lambda_i)\frac{\lambda_i^{y_i}}{y_i!}, \tag{1.1}$$

where λ_i is the expected number of events. A logarithm link function is often used to link λ with explanatory variables as $\lambda_i = \exp(\boldsymbol{X}_i'\boldsymbol{\beta})$; where $\boldsymbol{X}_i$ is a vector of explanatory variables and $\boldsymbol{\beta}$ is a vector of estimable parameters.

Negative binomial regression is an extension of Poisson regression with assumption that λ_i in (1.1) follows a gamma distribution. The negative binomial model can be expressed by rewriting the function form as:

$$\lambda_i = \exp[\boldsymbol{X}_i'\boldsymbol{\beta} + \epsilon_i], \tag{1.2}$$

where $\exp(\epsilon_i)$ follows gamma distribution error term with mean 1 and variance α. This extra error term allows differentiation between mean and variance.

In the context where two types of outcome are possible, binomial regression models can be used to explaining the probability of a specific event. The model is given as follows:

$$Pr(y_i) = \binom{n_i}{y_i} p_i^{y_i}(1-p_i)^{n_i - y_i}$$
$$log(\frac{p_i}{1-p_i}) = \boldsymbol{X}_i'\boldsymbol{\beta},$$

where y_i is the number of events for subject i; p_i is the probability of an event. n_i is the number of baseline samples.

Estimation and inference of model parameters is usually achieved by maximum likelihood estimation with iterative reweighted least square method [15].

1.1.3 Time-to-event Data Approach

Time-to-event data consist of time to each event and to the end of study. Cox's proportional hazards model [8] is a commonly used model for time-to-event data. It provides reliable estimates of survival times and relative risk associated with covariates. Several extensions have been proposed for circumstances where events may happen more than once [2, 16, 20]. The model assume that no more than one event may happen at a given time and events occur in continuous time.

The models are built on counting process models and intensity functions is the key for

modeling and statistical inference. The general recurrent event model setup is discussed as follows. For a single recurrent event process starting at $t = 0$, let $0 < t_1 < t_2 < t_3 < \cdots$ denote event times, where t_k is the time of k-th event. The associated counting process $\{N(t),\, 0 \leq t\}$ records the cumulative number of events generated by the process. Specifically, $N(t) = \sum_{k=1}^{\infty} I(t_k \leq t)$ is the number of events occurring over the time interval $(0, t\,]$. Intensity function of the process gives instantaneous probability of an event occurring at t and is mathematically defined as

$$\lambda(t|H(t)) = \lim_{\triangle t \to 0} \frac{Pr[\triangle N(t) = 1|H(t)]}{\triangle t}, \tag{1.3}$$

where $\triangle N(t) = N(t + \triangle t^-) - N(t^-)$ is the number of events in the interval $[t, t + \triangle t]$ and $H(t) = \{N(s),\, 0 \leq s < t\}$ denotes the history of the process at time t [7].

The above modeling framework is general and encompass various data structure, models, dependence structures. Three types of models are commonly used: overall intensity models, frailty models, and marginal hazard models [9]. Overall intensity model is also well known as Poisson Process model or Andersen-Gill model (A-G model) [2]. It describes situations where events occur randomly in such a way that the numbers of events in non overlapping time intervals are statistically independent. Frailty model is an extension of A-G model that considers conditional intensity given the unobservable random effect. This model can address the correlation within subjects Hougaard [12]. Marginal model provides a basis for the development of robust methods of inference because it doesn't involve assumptions as that of Poisson Process. When the correlation within subject is unknown or not of interest, the marginal hazard model approach which models the 'population-averaged' covariate effects has been widely used [20].

1.1.4 Time-varying Coefficient Recurrent Event Model

The models discussed in Section 1.1.3 are based on the constant covariates coefficients assumption. It is convenient for estimation and interpretation but the assumption may not be satisfied in real data which could lead to bias in the coefficient estimates. In addition, evaluating time-varying effect could be of prime interest in some studies. To incorporate time-varying coefficients into intensity function, two techniques have been proposed and widely used: the kernel-weighted partial likelihood method and spline models.

Kernel-weighted partial likelihood method is proposed by Fan et al. [10] and further developed by Cai and Sun [6]. Coefficients are estimated locally based on the partial likelihood in a window around each time point where models are built for intensity function.Cai et al. [4], Cai et al. [5] and Sun et al. [18] elaborated this approach using marginal hazard

framework.

Spline functions aim to approximate the functional form of varying coefficients. A common form is piecewise polynomials functions that satisfying continuity constraints at the knots. Zucker and Karr [23] developed a penalized partial likelihood approach . The penalty function was designed to make the estimator smooth and thereby reduce variance. More recently, Amorim et al. [1] expanded spline technique to the filed of recurrent events data. Yu et al. [21] used a Gaussian frailty model to describe the intensity, accommodating both time-varying and time-constant coefficients. Penalized spline method and Laplace approximation were used to estimate coefficients.

Part of this dissertation focuses on developing a semi-parametric recurrent event model with time-varying coefficients. Penalized B-spline function is adopted to estimate the time varying effects. I jointly estimate variance component and smoothing parameter by using profile likelihood method. The proposed approach can be relatively easily implemented by fitting a frailty model.

1.1.5 Multi-type Time-to-event Data

Several types of related recurrent events could occur in the same period of time. For example, in a transportation study, a safety-related event may be classified into several levels according to severity. It is a challenging yet intriguing to model all events simultaneously. In the context of multi-type recurrent event, Cai and Schaubel [3] proposed a class of semi-parametric marginal means/rates models, with a general relative risk form on the censored event processes of interest. Wang et al. [19] and Zhu et al. [22] developed approaches with an arbitrary structure for both the relationship between the recurrent events and the terminal event and the effect of covariates on the terminal event. In these studies, dependency between events is considered but not of interest. Frailty provides a convenient tool to incorporate dependence and heterogeneity. Sankaran and Anisha [17] extended shared frailty model to recurrent event data with multiple cases for gap time distributions. In Mazroui et al. [14], two types of recurrent events with dependent terminal events were jointly modeled.

1.2 Motivated Example: The 100-Car Naturalistic Driving Study

Methodologies developed in the dissertation are motivated by the objective of evaluating influence of crash on driving risk and distraction behavior. Data for the analysis were from the 100-Car NDS. NDS is a novel approach for traffic research characterized by instrumenting participant vehicles with multi-channel video cameras, high-precision kinematic sensors, GPS, and radar sensors and continuous data collection for a extended period of time. The

NDS provides detailed information under natural driving movement and provide an unique opportunity for evaluate the impact of crash on driver behavior. The 100-Car NDS was the first instrumented vehicle study undertaken with the primary purpose of collecting large-scale NDS naturalistic driving data. Data were collected from 241 primary participants in northern Virginia. About $2,000,000$ vehicle miles and $43,000$ hours of driving are recorded in total. This study used data from 107 primary drivers. Crash distribution across drivers by their demographic information is given in Table 1.1.

Table 1.1: Number of driver and driving time across groups in 100-Car NDS.

		Age group					
Total number of crash		< 30 (49 subjects)		$31 \sim 55$ (44 subjects)		> 55 (14 subjects)	
		F	M	F	M	F	M
0	Number of driver	11	15	10	26	4	7
	Average driving hour	255	317	236	378	264	332
1	Number of driver	10	3	2	4	1	2
	Average driving hour	344	345	448	431	32	360
2	Number of driver	4	6	.	2	.	.
	Average driving hour	415	359	.	489	.	.

1.3 Overview

The rest of this dissertation is organized as follows. In Chapter 2, I used mixed Binomial model to compare distraction rate before and after crash using 100-Car NDS. Four intensity-based recurrent event models are introduced in Chapter 3 with model assessment and comparison. In Chapter 4, I extended the research to explore time-varying coefficient models, which allow to examine the functional form of crash influence over time. Chapter 5 considers a general platform of multi-type recurrent event model with time-varying coefficient and dependent terminal event. Detailed derivation for two types of recurrent events is highlighted. Chapter 6 provides a general review on the contributions of this dissertation, as well as discussion for future research directions.

Bibliography

[1] L. D. Amorim, J. Cai, D. Zeng, and M. L. Barreto. Regression splines in the time-dependent coefficient rates model for recurrent event data. *Statistics in Medicine*, 27 (28):5890–5906, 2008.

[2] P. K. Andersen and R. D. Gill. Cox's Regression Model for Counting Processes: A Large Sample Study. *The Annals of Statistics*, 10(4):1100–1120, 1982.

[3] J. Cai and D. E. Schaubel. Marginal means/rates models for multiple type recurrent event data. *Lifetime Data Anal.*, 10(2):121–138, 2004.

[4] J. Cai, J. Fan, H. Zhou, and Y. Zhou. Hazard models with varying coefficients for multivariate failure time data. *The Annals of Statistics*, 35(1):324–354, 2007.

[5] J. Cai, J. Fan, J. Jiang, and H. Zhou. Partially linear hazard regression with varying coefficients for multivariate survival data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):141–158, 2008.

[6] Z. Cai and Y. Sun. Local linear estimation for time-dependent coefficients in cox's regression models. *Scandinavian Journal of Statistics*, 30(1):93–111, 2003.

[7] R. J. Cook and J. F. Lawless. *The statistical analysis of recurrent events*. Statistics for biology and health. Springer, 2007.

[8] D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[9] J. Fan and J. Jiang. *Non- and semi-parametric modeling in survival analysis*, volume 1 of *Front. Stat.* 2009.

[10] J. Fan, I. Gijbels, and M. King. Local likelihood and local partial likelihood in hazard regression. *The Annals of Statistics*, 25:1661–1690, 1997.

[11] D. Hedeker and R. D. Gibbons. *Longitudinal data analysis*. Wiley Series in Probability and Statistics. Wiley, 2006.

[12] P. Hougaard. *Analysis of Multivariate Survival Data*. Statistics for Biology and Health. Springer, 2000.

[13] A. Jahn-Eimermacher. Comparison of the andersen-gill model with poisson and negative binomial regression on recurrent event data. *Computational Statistics & Data Analysis*, 52(11):4989–4997, 2008.

[14] Y. Mazroui, S. Mathoulin-Plissier, G. MacGrogan, V. Brouste, and V. Rondeau. Multivariate frailty models for two types of recurrent events with a dependent terminal event: Application to breast cancer data. *Biometrical Journal*, 55(6):866–884, 2013. ISSN 1521-4036.

[15] J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972.

[16] R. L. Prentice, B. J. Williams, and A. V. Peterson. On the regression analysis of multivariate failure time data. *Biometrika*, 68(2):373–379, 1981.

[17] P. G. Sankaran and P. Anisha. Shared frailty model for recurrent event data with multiple causes. *Journal of Applied Statistics*, 38(12):2859–2868, 2011.

[18] L. Sun, X. Zhou, and S. Guo. Marginal regression models with time-varying coefficients for recurrent event data. *Statistics in Medicine*, 30(18):2265–2277, 2011.

[19] M.-C. Wang, J. Qin, and C.-T. Chiang. Analyzing recurrent event data with informative censoring. *Journal of the American Statistical Association*, 96(455):1057–1065, 2001.

[20] L. J. Wei, D. Y. Lin, and L. Weissfeld. Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions. *Journal of the American Statistical Association*, 84(408):1065–1073, 1989.

[21] Z. Yu, L. Liu, D. M. Bravata, L. S. Williams, and R. S. Tepper. A semiparametric recurrent events model with time-varying coefficients. *Statistics in Medicine*, 32(6): 1016–1026, 2013.

[22] L. Zhu, J. Sun, X. Tong, and D. Srivastava. Regression analysis of multivariate recurrent event data with a dependent terminal event. *Lifetime Data Analysis*, 16(4):478–490, 2010.

[23] D. M. Zucker and A. F. Karr. Nonparametric survival analysis with time-dependent covariate effects: A penalized partial likelihood approach. *The Annals of Statistics*, 18 (1):329–353, 1990.

Chapter 2 Evaluating the Impact of Crashes on Driving Behavior

2.1 Introduction

Driver behavior is a critical contributing factor to traffic safety. It has been estimated that more than 90% of crashes are associated with driver errors [15]. A study by Curry et al. [3] concluded that 95.6% of all teen-involved serious crashes were due to driver error.

Studies have shown that driving risk is strongly associated with demographic and personality characteristics [6]. Within individual level, Chipman [2], Kaneko and Jovanis [9], Waller et al. [16] has revealed reduced driving risk with increased driving experience . Previous studies on driver's experience are typically based on measures from years of driving and/or mileage traveled (e.g.,Levy [10], Waller et al. [16]). For example, Waller et al. [16] used length since licensure as the measurement of experience. Kaneko and Jovanis [9] considered years of experience of drivers from a national less-than-truckload firm as a factor. Experiences based on years or mileage of driving includes the effects of many factors. For example, drivers are more mature and doing fewer risky behaviors, or they learn skill and be able to deal with more complex situations, or they learn lesson from drastic crash events. Using time and amount of driving as measurement of experience is commonly associated with age-related changes [1]. Thus, with experience measures it is difficult to isolate the effects of a specific associated factor.

One hypothesis is that crash experience decreases the frequency of risky driving behavior. The rationale is that drivers learned from their collision events (crash) and change their behavior correspondingly, thus reduce driving risk. From a psychological point of view, Lucas [12] showed that drivers who had been involved in a motor vehicle accident reported significantly greater worries about driving than did drivers who had not been in an accident.

Research on this topic is limited. Lin et al. [11] considered association between crash experience and risk-taking path among students in Taiwan. Crash experience was measured in terms of crash history prior to the study, crash frequency, time elapsed since the last crash, and crash severity respectively. There was no significant association observed. af Wåhlberg [1] conducted a study of bus drivers for about three years. Repeated measurements of speed change behaviors were compared between drivers with no crashes and drivers who had at least one crash. A steady decline in speed change was observed within the crash group over time but not related to their crashes. The no-crash group showed a similar pattern. Rajalin and Summala [14] studied the effect of fatal accidents on surviving drivers' subsequent driving

behavior based on self-reported driving behavior. The study showed that car drivers typically returned to their 'normal' driving within a few months, while heavy-vehicle drivers tended to be more cautious in terms of mileage of driving. The above studies focused on various populations and adopted different measurement of driving behavior (risk-taking score, speed change, and amount of driving). Many are based on self-reported data. Currently, however, there is no established study about the relationship between crash experience and driving risk using naturalistic driving data.

Naturalistic driving study (NDS) provides an innovative way to access traffic safety and driving behavior data [4, 5, 13] and thus makes exploring the relationship between crash experience and driving behavior and risk accessible. Participant vehicles are instrumented with data acquisition systems (DASs) that include cameras and various sensors to continuously monitor the driving process. The video images and kinematic measures can provide not only the exact driving behavior, vehicle kinematic, and driving environmental information, but also the sequence and precise time for each sub-event.

In NDS the video recordings can be used to assess driver behaviors that were difficult to retrieve before. In the present study, distraction pattern is evaluated as driving behavior. A high frequency of distraction and/or more complex non-driving-related tasks indicates more distracted driving behavior. To be specific, secondary tasks, such as communications, entertainment, information gathering, and navigation not required to drive, have been used to measure distraction. The secondary tasks can be categorized into three levels: complex (C), moderate (M), and simple (S), based on whether the task requires multi-step, multiple eye glances away from the forward roadway, and/or multiple button presses. Detailed categorization of distraction can be found in Table 2.1[7].

Table 2.1: Definition of distraction.

Simple Secondary Tasks	Moderate Secondary Tasks	Complex Secondary Tasks
1.Adjusting radio	1.Talking/listening to handheld device	1.Dialing a handheld device
2.Adjusting other devices integral to the vehicle	2.Handheld device-other	2.Locating/reaching/answering handheld device
3.Talking to passenger in adjacent seat	3. Inserting/retrieving CD	3. Operating a personal digital assistant (PDA)
4. Talking/Singing: no passenger present	4. Inserting/retrieving cassette	4. Viewing a PDA
5. Drinking	5. Reaching for object (not handheld device)	5. Reading
6. Smoking	6. Combing or fixing hair	6. Animal/object in vehicle
7. Lost in thought	7. Other personal hygiene	7. Reaching for a moving object
8. Other simple tasks	8. Eating	8. Insect in vehicle
	9. Looking at external object	9. Applying makeup

In this study, driving behavior is measured by the probability of distraction in randomly selected baseline samples occurring over a specific time period. According to the visual and manual demand of the secondary task, each baseline can be classified into four categories: (1) no distraction, (2) simple secondary task, (3) moderate secondary task, and (4) complex secondary task. Moderate and complex secondary tasks are used as indicators of high-risk behavior and used the percentage of baselines with moderate and complex secondary tasks to measure the likelihood of a driver to engage in high-risk behavior. Drivers were considered to have driven more cautiously if a larger proportion of moderate and complex secondary tasks occurred before a crash than after. The rest of this chapter is organized as follows. Section 2.2 introduces the sampling design and data collection structure. Application of mixed binomial regression model is presented in Section 2.3 to estimate crash impact. Section 2.4 contains conclusion and some discussion for future research.

2.2 Baseline Sampling Design and Data Collection

Driving is a continuous process and distraction behavior changes constantly. Although NDS makes the entire driving record accessible, it is not possible to reduce all video recordings by visual inspection and keep track of distraction. Guo and Hankey [7] proposed an analysis framework based on a case-cohort approach. Under their analysis framework, a random sampling scheme for baseline reduction led to approximation of odds ratio risk rate ratio. This random sampling scheme is stratified by drivers, and the number of samples for each driver is proportional to the valid moving hours or miles traveled. The random samples also

present the general behavior of drivers under normal driving conditions, and thus can be used to evaluate driver distraction. In 100-Car NDS, 11,466 baseline samples were incorporated from two sources. The first source is an existing baseline sample from a previous National Surface Transportation Safety Center for Excellence (NSTSCE) project [7], which contained 10,952 baseline samples. To increase the sample size close to crash time, 514 additional baseline samples were reduced within a 30-hour window around the crashes, which brought the total sample size to 882 within this window. Among all baseline samples, 44% involve various levels of distraction, and 40% are categorized as moderate or complex distractions.

The objective of the study is to examine difference of distraction behavior before and after crash. We proposed a before-after comparison through evaluation of the probability of high risky behavior. This approach requires a predefined a window", e.g., 10 hours of driving time. For each crash, the predefined before-and-after windows are considered as a matched pair, and the number of moderate and complex secondary tasks as well as the total number of baselines were evaluated. Data collection for this approach is illustrated in Figure 2.1.

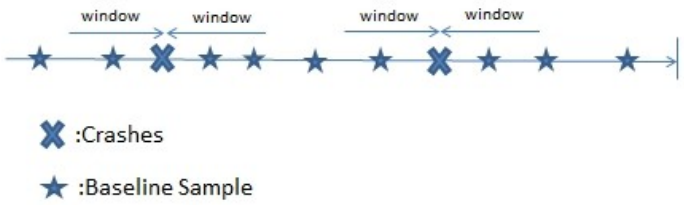


Figure 2.1: Baseline data collection.

An appropriate length of window needs to be carefully selected. If crash experience acts as a short stimulation for drivers and only affects driving behavior temporarily, a large window size will mask the effect of the crash by including non-influenced data. On the other hand, a small window size will not be able to capture enough event data and thus lose power to evaluate crash impact.

Window size selection is constrained by overlapping problem, which refers to a situation where the time interval between two accidents is less than the window size. Ideally, window size is chosen to be smaller than the shortest time interval between two consecutive accidents for one driver. However, in the 100-Car NDS, one driver experienced two collisions within 1 hour of driving, which makes the idea situation unsatisfied (one hour is too short to observe any event). Figure 2.2 shows a histogram of the time gaps between two consecutive crashes conducted by the same driver. Over 75% of two successive crashes for one driver were at least 30 hours apart, so the current study begins with a window size of 15 hours. Crash

effect based on other window sizes will be investigated later.

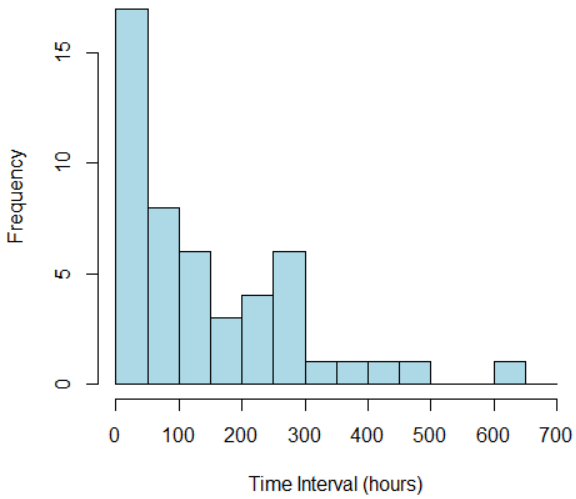


Figure 2.2: Crash interval histogram.

An existing sample of 10,952 baselines was randomly sampled from the 100-Car data [8]. For each baseline, a rigorous data reduction protocol was used to extract driver behavior information. Among these baselines samples, only 952 fall in the 30-hour window around observed crashes. For the purpose of comparing distraction before and after a crash, we reduced an additional 514 baselines within a 30-hour driving window around crashes. With these additional data, the final sampling scheme, as illustrated in Figure 2.3, was as follows: Two samples were randomly selected within a 2-hour window before and after a crash, and two were randomly selected in the 25-hour window. For the rest of the 30 hours, two samples were randomly selected in every 5-hour window.

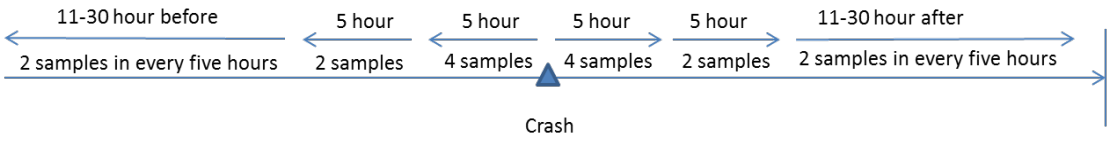


Figure 2.3: Baseline sampling scheme.

Table 2.2 lists the total number of baseline samples across window sizes. There is a small proportion of double-counted samples in the table, due to overlapping windows between

crashes. Baseline samples could be included in an after-crash window while being included in the before window of the next crash. Thus, the total number of baseline in Table 2.2 is larger than the total number of unique baselines that are identified from data reduction. Sampling rate, defined as the number of samples per hour, shows consistency between the before and after windows.

Table 2.2: Baseline sample distribution.

Window size (hours)	Total sample before	Total sample after	Sampling rate before	Sampling rate after
5	218	196	44	39
10	336	310	34	31
15	450	432	30	29
20	559	535	28	27
25	660	634	26	25
30	773	730	26	24
35	864	785	25	22
40	949	855	24	21
45	1003	901	22	20
50	1076	970	22	19
55	1136	1031	21	19
60	1184	1086	20	18

Figure 2.4 shows the ratio between the moderate and complex distraction proportions before and after a crash. The results indicate that drivers' engagement in moderate and complex secondary tasks tends to be lower after crashes, especially within a 15-hour driving time window. However, this decreasing effect tends to diminish over time.

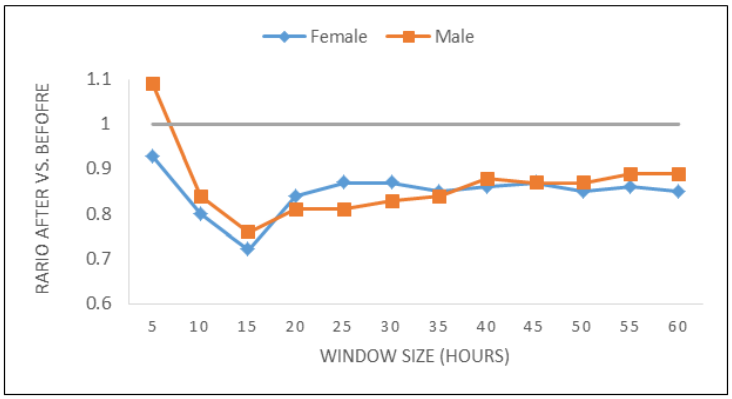


Figure 2.4: Ratio of distraction rate after vs. before by gender.

2.3 Modeling Baseline Distraction Using Mixed Binomial Regression

A formal statistical inference was conducted to investigate the impacts of crashes on driver distraction. For each crash, the number of baselines in which drivers engaged in moderate and complex secondary tasks during the before and after period are considered as a matched pair. A crash effect is observed if a larger probability of moderate and complex secondary tasks occurs before a crash than after. Mixed binomial regression models are used to evaluate the factors that affect the probability of distraction. Mixed binomial regression model is adopted to (1) incorporate the correlation among observations from the same driver, and (2) to adjust for confounding effects, e.g. gender, through modeling. Gender effect allows distraction behavior between male and female to be discerned. The model is given as follows:

$$Pr(Y_{ijk}) = \binom{n_{ijk}}{y_{ijk}} p_{ijk}^{y_{ijk}} (1 - p_{ijk})^{n_{ijk} - y_{ijk}}$$
$$log(\frac{p_{ijk}}{1 - p_{ijk}}) = \beta_0 + \beta_1 x_{ijk} + \beta_2 gender_i + \epsilon_i,$$

where:

1. Y_{ijk} is the total number of moderate and complex distractions for subject $i, i = 1, \cdots, 107$ in the before-after window of number j-th crash. Frequency follows a binomial distribution.

2. n_{ijk} is the total number of baseline samples in the associated window.

3. p_{ijk} is the probability of a moderate or complex distraction.

4. x_{ijk} indicates whether the distraction happens before or after a crash: $x_{ijk} = 0$ if it happens before, otherwise $x_{ijk} = 1$.

5. β_1 is crash effect and β_2 indicates gender distinction.

6. ϵ_i is a normally distributed random effect of mean 0 associated with subject i.

Results indicated that the percentage of baselines where drivers engaged in complex secondary tasks dropped after crashes. The maximum decrease occurred in the 15-hour window, with odds ratio = 0.54; 95% CI $[0.32, 0.93]$. Crash impact on distraction was also explored with varying window sizes, as shown in Figure 2.5. Distraction probability decreased after a crash, especially in the initial 15 hours. The difference diminished as window size increased and became negligible after 50 hours. This result suggests that drivers tend to

engage less in distractions during the initial period after a crash but return to regular behavior after a certain time period. The confidence band is relatively wide, which is primarily due to the relatively small sample size.

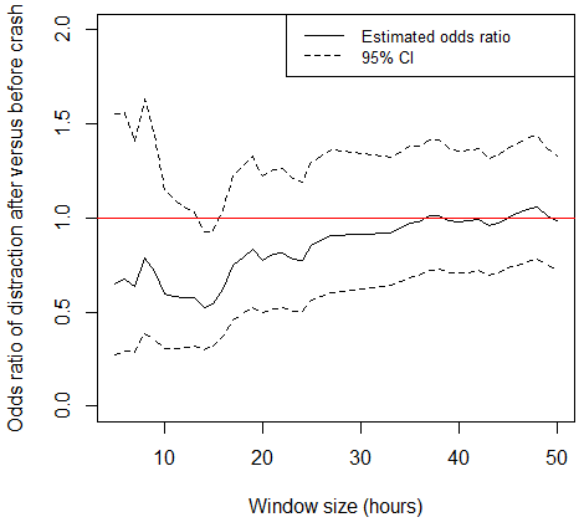


Figure 2.5: Crash impact on distraction proportion.

2.4 Discussion

This study evaluated the influence of crash on driver distraction behavior using 100-Car NDS data. Driving behavior was measured by secondary driving tasks. Crash influence on driving behavior was evaluated with a count-based approach using a mixed binomial regression model.The results indicate that drivers' engagement in moderate and complex secondary tasks tends to be lower after crashes, especially within a 15-hour driving time window. This decreasing effect tends to diminish over time.

There is some limitations of this count-based before-after comparison. Although it is straightforward, arbitrarily defined window lead to overlapping issue. Some baselines samples are studied twice. In addition, it's not easy to measure time-varying pattern of crash effect using this approach. In the next three chapters, we propose to evaluate the data from a different angle and solve the problem of defining window.

Bibliography

[1] A. af Wåhlberg. Changes in driver celeration behavior over time : do drivers learn from collisions? *Transportation Research Part F*, 15(5):471–479, 2012.

[2] M. L. Chipman. The role of exposure, experience and demerit point levels in the risk of collision. *Accident Analysis & Prevention*, 14(6):475 – 483, 1982.

[3] A. E. Curry, J. Hafetz, M. J. Kallan, F. K. Winston, and D. R. Durbin. Prevalence of teen driver errors leading to serious motor vehicle crashes. *Accident Analysis & Prevention*, 43(4):1285 – 1290, 2011.

[4] T. A. Dingus, S. Klauer, V. L. Neale, A. Petersen, S. E. Lee, J. Sudweeks, M. A. Perez, J. Hankey, D. Ramsey, S. Gupta, C. Bucher, Z. R. Doerzaph, J. Jermeland, and R. Knipling. The 100-car naturalistic driving study, phase ii–results of the 100-car field experiment. U.S. Dept. of Transportation Report DOT-HS-810-593, 2006.

[5] G. M. Fitch, S. A. Soccolich, F. Guo, J. McClafferty, Y. Fang, R. L. Olson, M. A. Perez, R. J. Hanowski, J. M. Hankey, and T. A. Dingus. The impact of hand-held and hands-free cell phone use on driving performance and safety-critical event risk. Technical Report DOT HS 811 757, 2013.

[6] F. Guo and Y. Fang. Individual driver risk assessment using naturalistic driving data. *Accident Analysis & Prevention*, 61:3–9, 2013.

[7] F. Guo and J. Hankey. Modeling 100-car safety events: A case-based approach for analyzing naturalistic driving data. Report No. 09-UT-006, 2009.

[8] F. Guo, S. G. Klauer, J. M. Hankey, and T. A. Dingus. Near-crashes as crash surrogate for naturalistic driving studies. *the Transportation Research Record:Journal of the Transportation Research Board*, 2147:66–74, 2010.

[9] T. Kaneko and P. P. Jovanis. Multiday driving patterns and motor carrier accident risk: A disaggregate analysis. *Accident Analysis & Prevention*, 24(5):437 – 456, 1992.

[10] D. T. Levy. Youth and traffic safety: The effects of driving age, experience, and education. *Accident Analysis & Prevention*, 22(4):327 – 334, 1990.

[11] D. Y. Lin, L. J. Wei, I. Yang, and Z. Ying. Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 62(4):711–730, 2000.

[12] J. L. Lucas. Drivers psychological and physical reactions after motor vehicle accidents. *Transportation Research Part F: Traffic Psychology and Behaviour*, 6(2):135 – 145, 2003.

[13] M. C. Ouimet, T. G. Brown, F. Guo, S. G. Klauer, B. G. Simons-Morton, Y. Fang, S. E. Lee, C. Gianoulakis, and T. A. Dingus. Higher crash and near-crash rates in teenaged drivers with lower cortisol response: an 18-month longitudinal, naturalistic study. *JAMA pediatrics*, 168(6):517–522, 2014.

[14] S. Rajalin and H. Summala. What surviving drivers learn from a fatal road accident. *Accident Analysis & Prevention*, 29(3):277– 283, 1997.

[15] J. R. Treat. A study of precrash factors involved in traffic accidents. *HSRI Research Review*, 10(6):35, 1980.

[16] P. F. Waller, M. R. Elliott, J. T. Shope, T. E. Raghunathan, and R. J. Little. Changes in young adult offense and crash patterns over time. *Accident Analysis & Prevention*, 33(1):117 – 128, 2001.

Chapter 3 Assessing Influence of Crash on Driving Risk Using Semi-parametric Recurrent Events Model

3.1 Introduction

Compared to the event frequency based approach, examining time to event provides an alternative way to evaluate covariate effect. In this case, no predefined time interval is required. It is also consistent with the natural data generation procedure and could be more informative in estimating covariates' effects.

In NDS, crash is a common indication of driving risk. In addition, several types of safety-related events were identified through kinematic signatures of the vehicle and confirmed through visual inspection for video recordings, such as: near crash (NC), and safety critical incident (SCI). A *crash* is defined as any contact with an object, either moving or fixed, at any speed in which kinetic energy is measurably transferred or dissipated. Crashes include a participants vehicle making contact with other vehicles, roadside barriers, and objects on or off the roadway, pedestrians, cyclists, or animals. A *NC* is defined as any circumstance requiring a rapid, evasive maneuver by the participant (or his/her vehicle) or any other vehicle, pedestrian, cyclist, or animal to avoid a crash. The crashes and near-crashes were identified through a multiple-step process of automatic trigger identification followed by visual confirmation by experts as described in Dingus et al. [6]. A *SCI* is defined as an unexpected events resulting in a close call or requiring fast action (evasive maneuver) on the part of a driver to avoid a crash Dingus et al. [6]. These safety related events represent non-desired safety conditions that should be avoided and are widely used in the literature as surrogate of crash for measuring driving risk Guo et al. [7] and are adopted in this paper.

NCs and SCIs occur at much higher frequency than crashes. Therefore, recurrent events modeling technique is needed. This technique has been commonly used in clinical trial and manufacture industry (e.g. Andersen and Gill [1], Lin et al. [9], Wei et al. [14]). Andersen and Gill [1] introduced a counting process model with the Cox [5] type of intensity function. The model assumes events occur randomly such that the numbers of events in non overlapping time intervals are statistically independent. Lin et al. [9] proved robustness of the inference by relaxing the Poisson-type assumption in Andersen and Gill [1]. Nielsen et al. [12] worked on an intensity function depending on unobservable quantities–frailties. Wei et al. [14] proposed a stratified model to analyze multivariate failure time data without correlation assumption. But in transportation, studies have more focused on duration models, which measure the

conditional probability of a crash happening given the history from the most recent crash(e.g., Chang and Jovanis [2], Jovanis and Chang [8], Lord and Mannering [10],Chung [3]). The processes of crash, NC, and SCI were rarely been studied as recurrent event.

This chapter focuses on investigating the influence of crashes on driving risk. We proposed to evaluate the influence of crashes on the intensity of SCI/NC by treating number of SCIs and NCs over time as a counting process. Four semi-parametric recurrent events models are compared and applied to 100-Car Naturalistic Driving Study. The objective is to evaluate whether drivers are more cautious in terms of whether SCI/NC rate will decrease after crash. Furthermore, we are also interested in whether male and female respond to crash differently.

The rest of this chapter is organized as follows. Section 3.2 introduces the 100-Car NDS in terms of measurement of driving risk. Section 3.3 explains the alternative models along with their assumptions and model fitting evaluation. A simulation study of model performance is summarized in Section 3.4. The application of models to the 100-Car NDS is presented in Section 3.5. Section 3.6 contains concluding remarks and some discussion.

3.2 100-Car NDS Data Setting for Recurrent Events Model

As discussed above, SCIs and NCs were used to measure driving risk. Average rate of both SCI and NC by gender, age, and total crash are explored in Table 3.1. Event rate is calculated as number of events per hours of driving. In general, SCIs occur 8 to 10 times more frequently than NCs across all groups of drivers. Higher SCI/NC rates are associated with drivers experienced more crashes.

Table 3.1: SCI/NC rate by groups

Total number of crash		Age group					
		< 30(49 subjects)		$31 \sim 55$(44 subjects)		> 55(14 subjects)	
		F	M	F	M	F	M
0	SCI rate*	0.18	0.15	0.15	0.11	0.04	0.1
	NC rate*	0.02	0.02	0.02	0.01	0.01	0.02
1	SCI rate*	0.27	0.26	0.41	0.22	0.16	0.14
	NC rate*	0.03	0.02	0.04	0.03	0.06	0.01
≤ 2	SCI rate*	0.52	0.33	.	0.27	.	.
	NC rate*	0.05	0.03	.	0.02	.	.

*: number of event per hour per driver

Data setting is shown in Figure 3.1. where each horizontal line represents the driving record of one driver. Drivers were subject to different numbers of crashes, NCs, and SCIs

at different time points throughout study. Thus, it was important to record all timestamps. We focused on the actual driving time. Non-driving time when the vehicle was not in use was excluded. As illustrated in the Figure, each driving period was divided into several phases based on relationship with crashes: before the first crash (coded as 0), between the first and second crash (coded as 1), and after the second crash (coded as 2). Driving period will be taken into account as covariate, working as an external and independent factor on SCI intensity. To account for potential confounding and interacting effects, gender and the age of the driver when first enrolled in the study are also evaluated.

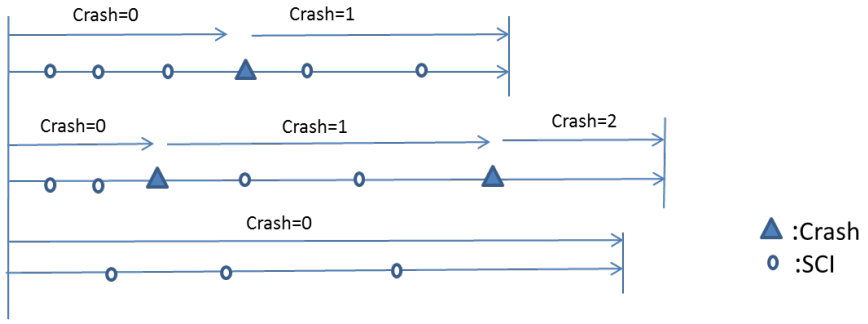


Figure 3.1: Data collection structure.

3.3 Semi-Parametric Recurrent Events Models

The concept of intensity functions and counting process has been introduced in 1. In this chapter, we focused on four alternative models, including an Andersen-Gill (A-G) model, a stratified A-G model, a frailty model, and a stratified frailty model.

3.3.1 Andersen-Gill Model

A Poisson process model, or Andersen-Gill model, is commonly used in the recurrent event literature [1]. It describes situations where events occur randomly in such a way that the numbers of events in non overlapping time intervals are statistically independent. The overall intensity function of Poisson Process is:

$$\lambda_i(t|\boldsymbol{z}_i(t)) = \lambda_0(t)\exp[\boldsymbol{z}_i'(t)\boldsymbol{\beta}], \tag{3.1}$$

where $i = 1, \cdots, m$ represents individual subject. Baseline intensity $\lambda_0(t)$ is a nonnegative integrable function. $\boldsymbol{z}_i(t)$ is a vector of fixed or time-varying external covariates associated with subject i and acts multiplicatively on the baseline. $\boldsymbol{\beta}$ is a vector of regression parameters

of the same length as $\boldsymbol{z}_i(t)$.

As can be seen from definition, the A-G model assumes that the probability of an event in $(t, t+\triangle t)$ may depend on t but not on $H(t)$. Cumulative intensity, denoted as $\mu(t) = \int_0^t \lambda(t)dt$, is continuous and finite for all $t > 0$ and can be explained as the average number of events occurring in time period $(0, t\ ]$.

The estimation of coefficients $(\widehat{\boldsymbol{\beta}})$ can be derived from maximizing log partial likelihood [4], which is given as follows:

$$log(PL) = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left\{\boldsymbol{z}_i(t)'\boldsymbol{\beta} - \sum_{k\in R_{ij}}\exp[\boldsymbol{z}_k(t)'\boldsymbol{\beta}]\right\}. \tag{3.2}$$

Where n_i is the total number of events of subject i, R_{ij} consists of all subjects who are at risk at given t_{ij}. The baseline function can be estimated by inserting $\widehat{\boldsymbol{\beta}}$ into log partial likelihood and given in (3.3).

$$\widehat{\lambda}_0(t) = \frac{\sum_{i=1}^{m} Y_i(t)dN_i(t)}{\sum_{i=1}^{m} Y_i(t)\exp[\boldsymbol{z}_i(t)'\widehat{\boldsymbol{\beta}}]}, \tag{3.3}$$

where $Y_i(t)$ indicates whether subject i is under study at time t, and $\sum_{i=1}^{m} Y_i(t)dN_i(t)$ is the total number of events at t.

3.3.2 Stratified A-G model

It is likely that subjects are sampled from subgroups of individuals with varying intensity functions. An effective way to accommodate this situation is to stratify the baseline function into strata. It is assumed that baseline functions vary among strata while coefficients remain the same. The stratification model is given bellow:

$$\lambda_{ri}(t|\boldsymbol{z}'_{ri}(t)) = \lambda_{0r}(t)\exp[\boldsymbol{z}'_{ri}(t)\boldsymbol{\beta}], \tag{3.4}$$

where $r = 1, \cdots, R$ indicates stratum level, $\lambda_{0r}(t)$ is the baseline function for stratum r, and $\boldsymbol{z}_{ri}(t)$ is the corresponding covariates vector for subject i in strata r, $1 \le i \le m_r$.

Estimation procedure for stratified model is very similar as that of A-G model, with log partial likelihood:

$$log(PL_{str}) = \sum_{r=1}^{R}\sum_{i=1}^{m_r}\sum_{j=1}^{n_{ri}}\left\{\boldsymbol{z}_{ri}(t)'\boldsymbol{\beta} - \sum_{k\in R_{rij}}\exp[\boldsymbol{z}_{rk}(t)'\boldsymbol{\beta}]\right\}. \tag{3.5}$$

3.3.3 Shared Frailty Model

In applications involving multiple subjects, heterogeneity is often apparent and requires consideration. Heterogeneity describes, conditioning on covariates, variation among individual intensity rate functions. In another word, there is more within-individual variation in event occurrence than is accounted for by a Poisson process. To capture the relation of the correlated observations, it has been considered that those event times of one subject share an unobserved effect [11, 12]. This shared individual random effect accounts for the variation beyond conditioning on covariates.

The shared frailty model assigns a random effect, u_i, $i = 1, \cdots, M$ to each subject acting multiplicately on the Poisson intensity model. Then, the intensity function is formed accordingly as:

$$\lambda_i(t|u_i, \boldsymbol{z}_i(t)) = u_i \lambda_0(t) \exp[\boldsymbol{z}_i'(t)\boldsymbol{\beta}], \tag{3.6}$$

where the random terms $u_i, \cdots, u_m$ are independent and identically distributed (i.i.d) with mean 1 (manually defined) and distribution function $G(u)$. Frailty gives the interpretation that individuals with $u_i > 1$ tend to occur at a faster rate. There are several choices for distribution $G(u)$, including Gamma, inverse Gaussian, and lognormal [15]. In this chapter, lognormal distribution is of primary use. If one specifies $\gamma_i = log(u_i)$, then $\gamma_i \sim N(0, \sigma)$ and 3.6 turns to be:

$$\lambda_i(t|\gamma_i, \boldsymbol{z}_i(t)) = \lambda_0(t) \exp[\boldsymbol{z}_i'(t)\boldsymbol{\beta} + \gamma_i], \tag{3.7}$$

For the shared frailty model, there are two commonly used methods to obtain $\widehat{\beta}$. One is expectation-maximization (EM) algorithm and the other is maximizing penalized partial log-likelihood. Therneau et al. [13] have proved that solution to lognormal shared frailty models by the E-M algorithm is closely linked to PPL estimation. The logarithm of PPL is given below:

$$log(PPL) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left\{ \boldsymbol{z}_i(t)'\boldsymbol{\beta} - \sum_{k \in R_{ij}} \exp[\boldsymbol{z}_k(t)'\boldsymbol{\beta}] \right\} - \frac{1}{2\sigma^2} \sum_{i=1}^{M} \boldsymbol{\gamma}'\boldsymbol{\gamma}, \tag{3.8}$$

where $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \cdots, \boldsymbol{\gamma}_m)$.

The maximization of this approximate likelihood is a doubly iterative process that alternates between the following two steps:

1. For a fixed value of σ^2, find the best covariates estimation by maximizing the penalized partial log likelihood, Log(PPL)

2. For fixed values of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, calculate the REML estimation of $\widehat{\sigma}^2 = \frac{\widehat{\gamma}'\widehat{\gamma}+trace(H_{22}^{-1})}{m}$ in which H_{22}^{-1} is the inverse of the second derivative matrix associated with frailty terms.

The estimation for baseline function is given as (3.9), which is the same as the A-G model with an offset of estimated random effects.

$$\widehat{\lambda}_0(t) = \frac{\sum_{i=1}^{m} Y_i(t) dN_i(t)}{\sum_{i=1}^{m} Y_i(t) \exp[\boldsymbol{z}_i(t)'\widehat{\boldsymbol{\beta}} + \hat{\gamma}_i]} \tag{3.9}$$

3.3.4 Stratified Shared Frailty Model

The stratified shared frailty model encloses both varying baseline function and among individuals variation and as a combination of the (3.4) and (3.7) as follows:

$$\lambda_{ri}(t|\boldsymbol{z}_{ri}(t), \gamma_{ri}) = \lambda_{0r}(t) \exp[\boldsymbol{z}_{ri}(t)'\boldsymbol{\beta} + \gamma_{ri}]. \tag{3.10}$$

where r indicates stratum level with baseline function of $\lambda_{0r}(t)$, and $z_{ri}(t)$ is covariates vector for subject $i,\ 1 \leq i \leq m_r$.

The estimation for β is by maximizing PPL, similar as that of previous shared frailty model. Baseline estimation for stratum r is given below:

$$\widehat{\lambda}_{0r}(t) = \frac{\sum_{i=1}^{m_r} Y_{ri}(t) dN_{ri}(t)}{\sum_{i=1}^{m_r} Y_{0r}(t) \exp[\boldsymbol{z}_{ri}(t)'\widehat{\boldsymbol{\beta}} + \widehat{\gamma}_{ri}]}, \tag{3.11}$$

where m_r is the number of subjects in strata r and $Y_{ri}(t)$ is an indicator of whether subject i in strata r is still under study at time t.

3.3.5 Model Fitting: Cox-Snell Residual

Cox-Snell residuals are useful for checking the overall fit of the final model [4]. For case of several processes $i = 1, \cdots, M$, with intensity $\lambda_i(t)$, Cox-Snell residuals are defined as

$$r_{ij} = \int_{t_{i,j-1}}^{t_{i,j}} \widehat{\lambda}_i(s) ds, \tag{3.12}$$

where $j = 1, \cdots, n_i + 1$. $t_{i,0}$. t_{i,n_i+1} are the start and stop times for subject i. $\widehat{\lambda}_i(t)$ is the estimated intensity rate. If the model is correct then r_{ij} should behave like a censored sample from a unit exponential distribution. Thus a plot of the estimated cumulative intensity rate of the residuals versus the residuals should be a straight line through the origin with a slope of 1 [4].

3.4 Simulation Study

3.4.1 Simulation setup

We conducted a simulation study to evaluate the performance of the alternative approaches. The simulation setup is analogous to the real situation and described below:

1. The driving time for 50 subjects was generated from a normal distribution with a mean of 335 and a standard deviation of 160, which was estimated from the 100-Car Study Data.

2. For each subject, up to two crashes were generated based on the intensity function:

$$\lambda_i(t) = \frac{1}{150}. \tag{3.13}$$

   The rate $\frac{1}{150}$ was selected based on the crash rate estimated from 100-Car data. The rate implies that on average one crash will occur for every 150 hours of driving. Other baseline rates were also evaluated and the model performance was robust to the change in baseline rate. Gender and other external factors were not considered in the simulation. Crash intensity was restricted to be constant over time. Crashes were considered to occur independently. If a simulated crash occurred later than the driver's study time, the crash would be censored.

3. After generating censor time and crash time, the intensity function for each driver is defined as follows:

$$\begin{aligned}\lambda_{ri}(t) =& c_r t^{k_r-1} \times \exp\Big[\beta_1(sex_{ri} = M) + \beta_2(I_{ri}(t) = 1) + \beta_3(I_{ri}(t) = 2) \\ &+ \beta_4(I_{ri}(t) = 1)(sex_{ri} = M) + \beta_5(I_{ri}(t) = 2)(sex_{ri} = M) + \gamma_{ri}\Big],\end{aligned} \tag{3.14}$$

   where

   (a) $r = 1, 2$, or 3, indicates stratum level. Drivers in level 1 do not experience any crashes, drivers in level 2 have only one crash, and drivers in level 3 have two crashes.

   (b) Baseline functions depend on two parameters, c and r, which can vary from stratum to stratum, as denoted by c_r and k_r. $k_r > 1$ indicates that the SCI rate increases over time. $k_r = 1$ corresponds to a constant rate, while $k_r < 1$ means a decreasing rate.

(c) $I_{ri}(t)$ is a time-varying crash indicator. It takes a value of 1 when t is between the first and second crash and 2 when t is larger than the second crash time.

(d) β_1 is gender effect; β_2 is the first crash effect for female drivers; β_3 is the second crash effect for female drivers; $\beta_2 + \beta_4$ is the first crash effect for male drivers; $\beta_3 + \beta_5$ is the second crash effect for male drivers.

(e) $\gamma_{ri} \sim N(0, \sigma), r = 1, \cdots, 3; i = 1, \cdots, 50$ are independent frailty terms.

3.4.2 Simulation results

In order to cover a sufficient large range of parameter space, 24 settings with different baseline parameter combinations, as well as various combinations of gender and crash effects were tested. In each setting, 500 repeats were generated and two models were implemented: a stratified frailty model and a frailty model. Because of space limits, we only provide results for selected scenarios.

Figure 3.2 shows a coverage probability (CP) comparison between two models, where three strata share the same shape parameter, k (set as 1), but have different scale parameters, c. Seven setting results with assorted combinations of c are presented. Both models perform well, with an average CP around 95% and small bias (1% to 3% difference). The stratified frailty model does not show a great benefit over the frailty model because the variation among strata is proportional, and thus can be explained through frailty terms. Figure 3.2 shows results from another seven settings where three strata share the same scale parameter, c (set as 0.2), but different shape parameters, k. The stratified frailty model retains a CP of around 95% while the frailty model performs poorly, with the CP being as low as 20%.

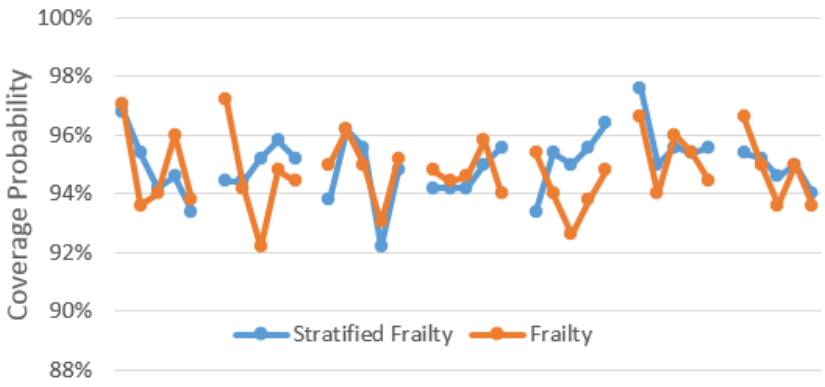


Figure 3.2: Coverage probability comparison I.

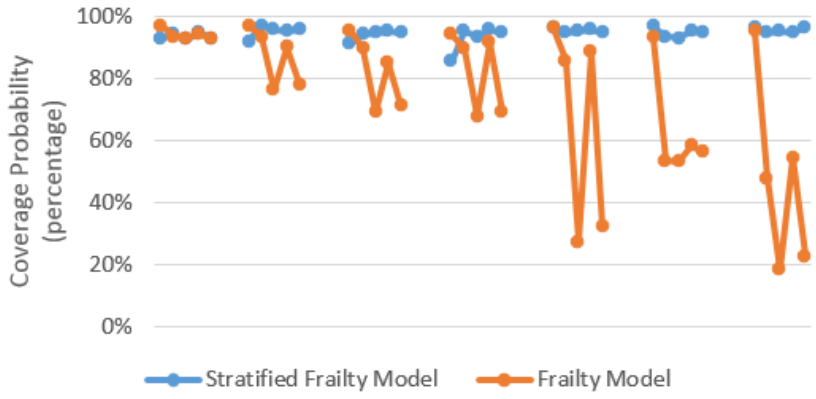


Figure 3.3: Coverage probability comparison II.

To evaluate model performance at different levels of standard deviation for frailty terms, two settings are shown in Table 3.2: one has frailty terms follows $N(0, 0.5)$, the other has frailty terms follow $N(0, 1)$. Baseline functions are set to be the same. This variance of frailty terms represents a degree of heterogeneity among subjects. With higher levels of heterogeneity, we observed larger bias and empirical standard error for fixed effect estimation (β_1). Bias and empirical standard error for other effects remain similar. Compared to the low CP of the frailty model, the stratified frailty model has around a 95% CP for all effects. In Table 3.3, the performance of the stratified frailty model was tested when there was no stratification. Three baseline functions are set with the same c and k. Although the stratified frailty model is more complicated than the situation requires, the 95% CP on average indicates a credible and stable estimation.

Table 3.2: Simulation result I: k=(.95,1.1,1.22) c=(0.2,0.17,0.15).

		Stratified Frailty Model				Frailty Model			
Parameter	True value	Bias	SE $^{\%}$	SEM$^{\triangle}$	CP*	Bias	SE $^{\%}$	SEM$^{\triangle}$	CP*
β_1	-0.2	0.002	0.159	0.173	95.8	0.004	0.198	0.205	95.6
β_2	0	-0.001	0.059	0.062	94.8	0.04	0.06	0.059	89.4
β_3	-0.6	0.003	0.078	0.078	95.6	0.115	0.085	0.074	66.6
$\beta_2+\beta_4$	-0.2	0.002	0.069	0.068	94.2	0.046	0.071	0.066	87
$\beta_3+\beta_5$	-0.7	0.007	0.082	0.083	94	0.124	0.089	0.079	66.4
σ	0.5	-0.036	NA	NA	NA	0.121	NA	NA	NA
β_1	-0.2	0.039	0.323	0.301	91.6	-0.011	0.698	0.372	92.8
β_2	0	-0.003	0.056	0.055	96	0.025	0.06	0.052	88.8
β_3	-0.6	0.005	0.083	0.069	94.2	0.095	0.079	0.065	67.7
$\beta_2+\beta_4$	-0.2	-0.002	0.061	0.06	95.2	0.095	0.079	0.065	67.7
$\beta_3+\beta_5$	-0.7	0.009	0.087	0.073	95	0.101	0.083	0.069	65.1
σ	1	-0.063	NA	NA	NA	0	NA	NA	NA

%: Empirical standard error

$\triangle$: Mean of standard error

*: Coverage probability

Table 3.3: Simulation result II: k=(1,1,1) c=(0.2,0.2,0.2).

		Stratified Frailty Model				Frailty Model			
Parameter	True value	Bias	SE $^{\%}$	SEM$^{\triangle}$	CP*	Bias	SE $^{\%}$	SEM$^{\triangle}$	CP*
β_1	-0.2	-0.006	0.228	0.265	96.8	-0.007	0.218	0.237	97
β_2	0	-0.003	0.086	0.086	95.4	-0.004	0.082	0.08	93.6
β_3	-0.6	-0.005	0.121	0.116	94.2	-0.007	0.106	0.104	94
$\beta_2+\beta_4$	-0.2	-0.006	0.1	0.097	94.6	-0.005	0.092	0.091	96
$\beta_3+\beta_5$	-0.7	0.003	0.129	0.125	93.4	0.002	0.115	0.113	93.8
σ	0.75	-0.036	NA	NA	NA	-0.021	NA	NA	NA

%: Empirical standard error

$\triangle$: Mean of standard error

*: Coverage probability

In summary, stratified frailty model is capable of accommodating possible variation among groups without losing power of the test for effects of interest. If subjects behave differently with various intensity functions, aggregating them together will mask the effect of covariates in the individual level.

3.5 Application in 100-Car NDS

In this section, four models are applied to 100-Car NDS. Three covariates are incorporated in to model: gender (G), age at the driver first enrolled in the study, and crash effect based on relationship with crashes (0 for before first crash, 1 for between first and second crash, and 2 for after second crash). In order to test and estimate each crash effect, it is considered as categorical variable. Since the study lasted for one year, age is considered to be constant.

Estimation of crash effects on SCI using A-G model, stratified A-G model, shared frailty model, and stratified shared frailty model is elaborated in Table 3.4. The stratified frailty model yields an intensity rate ratio of 0.8 (95% CI [0.693, 0.971]) between after the time of the first crash and before the first crash for male drivers. There is no significant first crash influence on female drivers, with an intensity rate ratio of 1.115 (95% CI $[0.96, 1.296]$). Unlike first crash effect, SCI risk drops sharply after the second crash for both female and male drivers, with a corresponding rate ratio of 0.432 (95% CI $[0.342, 0.547]$) and 0.472 (95% CI $[0.377, 0.59]$) respectively. The shared frailty model shows a comparable second crash effect as the stratified frailty model. Post-crash intensity is significantly lower. However, in terms of first crash effect, it is not significant for male drivers. Other than that, intensity rate increases substantially. Models without frailty terms have larger standard errors, thus wider confidence interval on the estimation, which leads to non-significant results.

Table 3.4: Crash effect estimation on SCI.

Model	Contrast	Estimate	Intensity Rate Ratio	CI*		Pr>ChiSq
A-G	1 vs. 0 Female	0.228	1.256	0.745	2.119	0.393
	2 vs. 1 Female	0.526	1.693	1.083	2.645	0.021
	1 vs. 0 Male	0.162	1.176	0.712	1.943	0.528
	2 vs. 1 Male	0.269	1.309	0.778	2.201	0.310
Stratified A-G	1 vs. 0 Female	-0.053	0.948	0.552	1.628	0.847
	2 vs. 1 Female	0.056	1.058	0.691	1.619	0.795
	1 vs. 0 Male	-0.382	0.683	0.378	1.233	0.206
	2 vs. 1 Male	-0.065	0.937	0.572	1.533	0.795
Shared frailty	1 vs. 0 Female	0.145	1.156	1.017	1.314	0.027
	2 vs. 1 Female	-0.385	0.681	0.57	0.812	$< .0001$
	1 vs. 0 Male	-0.030	0.970	0.836	1.126	0.691
	2 vs. 1 Male	-0.337	0.714	0.598	0.852	0.000
Stratified frailty	1 vs. 0 Female	0.109	1.115	0.960	1.296	0.155
	2 vs. 1 Female	-0.839	0.432	0.342	0.547	$< .0001$
	1 vs. 0 Male	-0.199	0.820	0.693	0.971	0.021
	2 vs. 1 Male	-0.751	0.472	0.377	0.590	$< .0001$

*: Confidence Limits of Intensity Rate Ratio

Table 3.5 lists estimations of first crash effect on NC based on four models. NCs are observed much less frequently compared to SCI events. Four female drivers experienced two or more crashes in the study, as shown in Table 3. After careful examination, only one of them had more than one NC recorded after the second crash. The rest have the second crash as their last driving record. Consequently, the estimation of the second crash effect for female drivers depends heavily on one single driver, which may lead to an individual crash influence rather than a population-wise effect. Thus, we decided to use time up to the second crash only and evaluate the first crash influence on NC. As indicated by the stratified frailty model, the intensity rate after first crash is 0.52 times (95% CI $[0.314, 0.874]$) the before-crash intensity rate for male drivers. Female drivers do not show a significant decreasing trend after a crash. A similar crash influence was found for SCI data. Estimation of σ is 0.93 based on REML estimation.

The shared frailty model proposes a different crash influence compared to the stratified frailty model. Neither male nor female drivers reveal a remarkably lower post-crash intensity. Models without frailty terms have larger standard errors, and thus wider confidence bands on the estimation, which leads to non-significant results.

Table 3.5: Crash effect estimation on NC.

Model	Contrast	Estimate	Intensity Rate Ratio	CI*		Pr>ChiSq
A-G	1 vs. 0 Female	0.206	1.228	0.712	2.120	0.460
	1 vs. 0 Male	0.284	1.329	0.816	2.163	0.253
Stratified A-G	1 vs. 0 Female	-0.359	0.698	0.378	1.289	0.251
	1 vs. 0 Male	-0.545	0.58	0.329	1.024	0.060
Shared frailty	1 vs. 0 Female	0.222	1.249	0.815	1.913	0.308
	1 vs. 0 Male	-0.140	0.870	0.558	1.355	0.537
Stratified frailty	1 vs. 0 Female	-0.117	0.890	0.551	1.438	0.633
	1 vs. 0 Male	-0.646	0.524	0.314	0.874	0.013

*: Confidence Limits of Intensity Rate Ratio

Figure 3.4 and Figure 3.5 show the Cox-Snell residual plots of four intensity based models for residuals of SCI and NC respectively. As mentioned in Section 3.3, if model is specified correct, Cox-Snell residuals follow exponential one distribution.

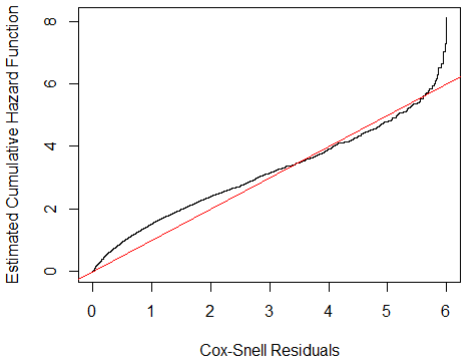
In current study for SCI, distribution of residuals is heavy tailed compared to exponential one distribution. For residuals larger than 6, the percentage varies from 1% to 2% from model to model. The probability associated with large (> 6) Cox-Snell residuals is supposed to be 0.25% for exponential one distribution. These extreme large residuals will lead to a departure of fitting from a straight line. Long intervals between two SCI events are the major source of extreme residuals, such as a 20-hour gap compared to a 5-hour gap on average. We have

Table 3.6: Distribution of extreme Cox-Snell residuals on SCI.

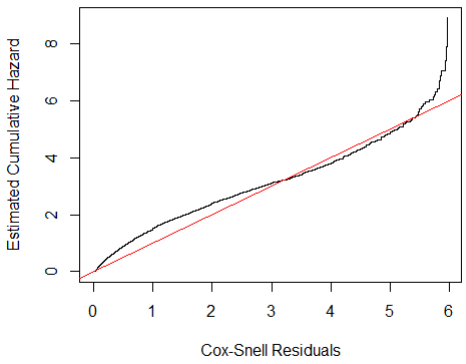
Quartile	96%	96.5%	97%	97.5%	98%	98.5%	99%	99.5%	100%
A-G	3.90	4.16	4.69	5.38	6.04	7.34	9.67	16.16	88.95
Stratified A-G	3.89	4.21	4.59	5.06	5.85	7.14	9.01	15.40	71.00
Frailty	3.77	4.02	4.32	4.64	5.12	5.68	6.76	8.73	36.73
Stratified frailty	3.71	3.97	4.21	4.52	4.88	5.43	6.37	8.11	34.72

examined those long gaps and there is a possibility that this was caused by missing event identification during the data reduction process. For this reasons, we present the distribution of extremely large residuals in Table 3.6 and set the upper limit of the residual to 6. It can be shown that the model fitting is reasonably well for majority of the data points.

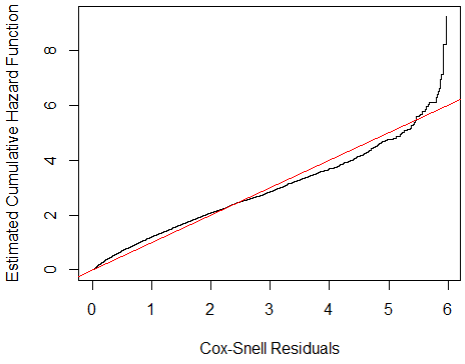
For NC, residual plots of stratified frailty model is much closer to a straight line comparing to other three models, indicates the best model fitting. In order to evaluate difference among strata, estimated baseline intensity functions of SCI and NC are explored by stratified frailty model in Figure 3.6. It can be concluded intensity rate among different stratum is not identical, which supports the idea of stratification.



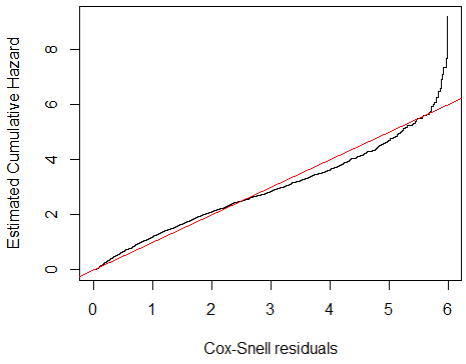
(a) A-G model



(b) Stratified A-G model

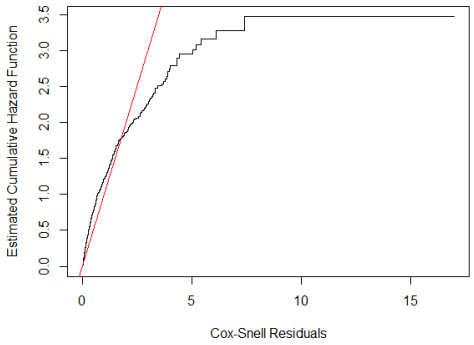


(c) Frailty Model

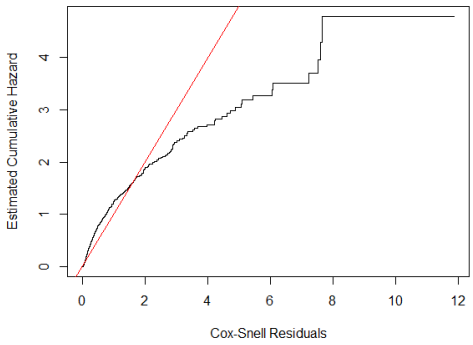


(d) Stratified Frailty Model

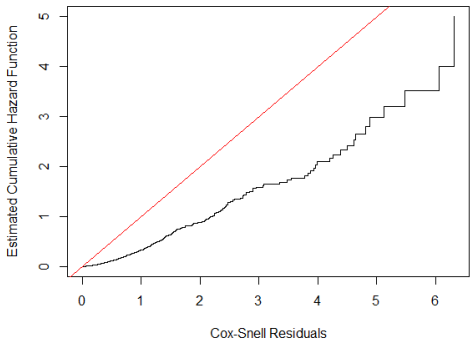
Figure 3.4: Cox-Snell residual plots for SCI.



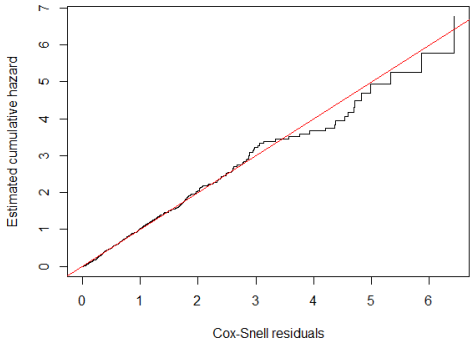
(a) A-G model



(b) Stratified A-G model

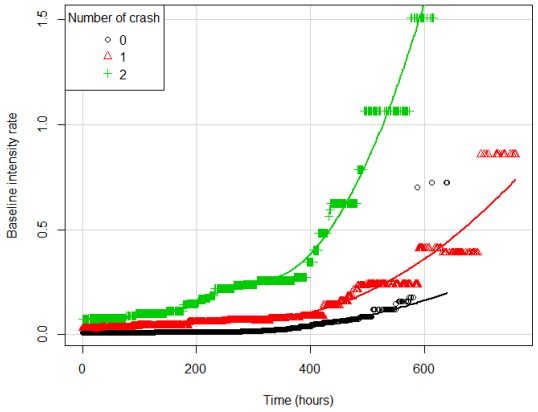


(c) Frailty Model



(d) Stratified Frailty Model

Figure 3.5: Cox-Snell residual plots for NC.



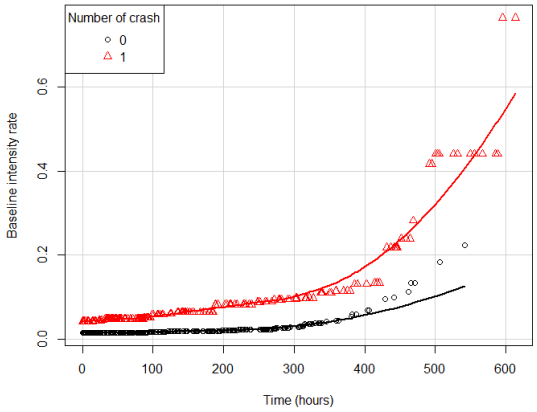


Figure 3.6: Baseline intensity rate estimation of SCI(left) and NC(right).

3.6 Conclusion and Discussion

This study evaluated the influence of crashes on driving risk using 100-Car NDS data. The results suggest that crashes have a positive effect on driver behavior with lower SCI intensity after crashes. Drivers might either learn from the crashes experience or be more attentive while driving, which is reflected in the reduced SCI intensity within a short window after crashes. In addition, the study indicates that female and male drivers showed different response to crashes and the number of crashes also influence driver behavior. Male drivers responded to both the first crash and the second crash with a lower SCI intensity after each crash. Females showed no significant response to the first crash but did show a decreased SCI intensity after the second crash. These findings provide crucial information for understanding driver's response to dramatic safety events and can be critical for development safety education programs and safety counter measures.

We evaluated and compared four intensity-based recurrent models based on the characteristics of the data. Safety outcomes, including SCIs and NCs, are used as markers for driving risk. The simulation study demonstrated that the stratified frailty model is capable of accommodating possible variation among groups without losing power to test for effects of interest. If subjects behave differently among various levels, aggregating them together will mask the effect at the individual level. We also observed robust performance of the stratified frailty model when subjects are not from different levels. Application the models to the data suggest that the stratified frailty model fits the context of the study and provides the best model fitting for the data.

There are a couple of limitations of this study. First, the individual driver risk variation might be confounded with the observed effect. Second, the study is based on a relative small number of crashes with mild crash severity. With larger NDS data sets becoming available, such as the Second Strategic Highway Research Program (SHRP 2) Naturalistic Driving Study, more concrete evidence will be available on the influence of crashes on driver behavior and potentially the influence of crashes by severity.

Bibliography

[1] P. K. Andersen and R. D. Gill. Cox's Regression Model for Counting Processes: A Large Sample Study. *The Annals of Statistics*, 10(4):1100–1120, 1982.

[2] H. Chang and P. P. Jovanis. Formulating accident occurrence as a survival process. *Accident Analysis & Prevention*, 22(5):407 – 419, 1990.

[3] Y. Chung. Development of an accident duration prediction model on the korean freeway systems. *Accident Analysis & Prevention*, 42(1):282 – 289, 2010.

[4] R. J. Cook and J. F. Lawless. *The statistical analysis of recurrent events*. Statistics for biology and health. Springer, 2007.

[5] D. R. Cox. Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[6] T. A. Dingus, S. Klauer, V. L. Neale, A. Petersen, S. E. Lee, J. Sudweeks, M. A. Perez, J. Hankey, D. Ramsey, S. Gupta, C. Bucher, Z. R. Doerzaph, J. Jermeland, and R. Knipling. The 100-car naturalistic driving study, phase ii–results of the 100-car field experiment. U.S. Dept. of Transportation Report DOT-HS-810-593, 2006.

[7] F. Guo, S. G. Klauer, J. M. Hankey, and T. A. Dingus. Near-crashes as crash surrogate for naturalistic driving studies. *the Transportation Research Record:Journal of the Transportation Research Board*, 2147:66–74, 2010.

[8] P. P. Jovanis and H. Chang. Disaggregate model of highway accident occurrence using survival theory. *Accident Analysis & Prevention*, 21(5):445 –458, 1989.

[9] D. Y. Lin, L. J. Wei, I. Yang, and Z. Ying. Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 62(4):711–730, 2000.

[10] D. Lord and F. Mannering. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice*, 44(5):291 – 305, 2010.

[11] C. A. McGilchrist and C. W. Aisbett. Regression with frailty in survival analysis. *Biometrics*, 47:461–466, 1991.

[12] G. G. Nielsen, R. D. Gill, P. K. Andersen, and T. I. A. Sørensen. A Counting Process Approach to Maximum Likelihood Estimation in Frailty Models. *Scandinavian Journal of Statistics*, 19(1), 1992.

[13] T. M. Therneau, P. M. Grambsch, and V. S. Pankratz. Penalized survival models and frailty. *Journal of Computational and Graphical Statistics*, 12(1):156–175, 2003.

[14] L. J. Wei, D. Y. Lin, and L. Weissfeld. Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions. *Journal of the American Statistical Association*, 84(408):1065–1073, 1989.

[15] A. Wienke. *Frailty Models in Survival Analysis*. Chapman & Hall/CRC Biostatistics Series. CRC Press, Hoboken, 2010.

Chapter 4 Inference on Semi-parametric Frailty Model with Time-varying Coefficient

4.1 Introduction

Recurrent events data are often encountered in longitudinal study in many disciplines, such as biomedical research, public health, and engineering. It has been widely studied using marginal models and frailty models by assuming risk factors' effects are constant over time [2, 13, 20]. As constant effect assumption may not hold in reality, it is an important extension to evaluate time-varying coefficients. Among varieties of approaches, semiparametric time-varying coefficients model has received much attention over past decades [6, 19, 21]. By assuming covariate effects on the logarithm of the hazard function as an unknown function over time, researchers are able to explore temporal effects of the covariates on the failure time.

To incorporate time-varying coefficients in hazard function, two type of techniques have been widely studied. One is kernel-weighted partial likelihood and the other is spline-based model. Kernel-weighted partial likelihood was proposed by Fan et al. [10] and further developed by Cai and Sun [6]. Coefficients are estimated locally based on the partial likelihood in a window around each time point where models are build for intensity function. Tian et al. [19] constructed pointwise and simultaneous confidence intervals for the regression parameters over a properly chosen time interval via a simple resampling technique. They also derived a prediction method for future survival with any specific set of covariates. Cai et al. [4] and Cai et al. [5] elaborated such question under marginal hazard framework.

Spline provides an approximating function of the interested coefficient(s). Early studies were Cox-based models defined for univariate time-to-event and were proposed to detect nonlinear covariate effect, not specifically for time-varying effect [15]. Zucker and Karr [23] used a penalized partial likelihood approach, where the penalty function was designed to make estimates smooth and thereby reduce variance. Sleeper and Harrington [17] developed a survival model based on data from a clinical trial using regression splines. More recently, Amorim et al. [1] expanded spline technique to the filed of recurrent events data. Yan and Huang [21] offered an adaptive group lasso method that not only selects important variables but also selects between time-independent and time-varying specifications. Sun et al. [18] formulated a class of semiparametric marginal rates models, which incorporate a mixture of time-varying and time-independent parameters, to analyze recurrent event data. Yu et al.

[22] proposed a spline method to estimate coefficients and a Gaussian frailty to characterize the correlation among recurrent events.

A major motivation of this paper is to study how crash experience influences driving risk. In the previous Chapter, we showed driving risk was reduced after crash with amount of decrease varies between gender. To explore and examine the pattern of influence over time is an appealing question. Based on personal driving experience, the conjecture is that drivers are more cautious following a crash and less careful over time. A presumable consequence is that driving risk decreases first after crash and gradual increase as time pass by. Klauer et al. [12] stated driving pattern and risk tended to remain stable for mature drivers. It suggests driving risk may eventually return to the same level as that in pre-crash period. The primary data is from 100-Car Naturalistic Driving Study (NDS). Safety-related events, such as near-crashes (NCs) and safety-critical incidents (SCIs) were recorded. They represent undesired safety conditions to be avoided, thus are used as measurement of driving risk. We evaluate the influence of crashes on the intensity of SCIs and NCs by treating SCIs and NCs as counting processes.

There have been few references on the inference of smoothing parameters for time-varying effects. In this chapter, we use Gaussian frailty model for recurrent events to accommodate correlation among events within subject. Penalized B-spline is used to approximate time-varying coefficient. Variance components and smoothing parameters are estimated jointly by maximizing profile likelihood. We propose to estimate time-varying coefficient through regular frailty model by reparameterization. Besides easy implementation, it allows us to make systematic inference on all components, including smoothing parameters. The reason of using spline other than kernel-weighted partial likelihood is the unique double time systems in the 100-Car data. Crashes occur at different time points across subjects. It motives evaluation crash influence as a function of time after crash. On the other hand, baseline intensity may change over study time, which is another time system.

The remainder of this Chapter is organized as follows. Section 4.2 introduces the time-varying coefficient model and estimation procedure. In Section 4.3 we discuss inference of coefficients and variance components, following by hypothesis tests. We illustrate the method in Section 4.4 by applying it to 100-Car data. Simulation studies, including model performance in finite samples and comparisons between alternative penalty matrices, are summarized in Section 4.5. Finally, we conclude with discussion in Section 4.6.

4.2 Time-varying coefficient model

In a longitudinal study, suppose individual i is observed over $[0, T_i]$, $i = 1, \cdots, m$, where $t = 0$ indicates start of the event process. Let n_i be the total number of events individual i encounters during the observed time period. All time points, $0 < t_{i,1} < \cdots < t_{i,n_i} \leq T_i$, are recorded. We consider event times are independent of T_i. We formulate intensity function by a semiparametric function including both time-dependent coefficients and time-constant coefficients as follows:

$$\lambda_i\Big[t \mid \boldsymbol{z}_i(t), \boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{\gamma}\Big] = \lambda_0(t) \exp\Big[\boldsymbol{z}_i'(t)\boldsymbol{\beta}(t) + \boldsymbol{x}_i'(t)\boldsymbol{\alpha} + \boldsymbol{w}_i'\boldsymbol{\gamma}\Big]. \tag{4.1}$$

$\lambda_0(t)$ is a nonparametric baseline function. $\boldsymbol{z}_i(t)$ and $\boldsymbol{x}_i(t)$ are two vectors of time-varying (or constant) explanatory variables with dimension p and q respectively. Coefficients $\boldsymbol{\beta}(t)$ is a vector composed by p functions of times, $\boldsymbol{\beta}(t) = (\beta_1(t), \cdots, \beta_p(t))'$. $\boldsymbol{\alpha}(t)$ is a vector of time-constant coefficients. $\boldsymbol{w}_i$ explains the correlation structure among events for subject i. $\boldsymbol{\gamma}$ is a vector of random effects following $\mathcal{N}(\mathbf{0}, \Sigma(\boldsymbol{\theta}))$. $\boldsymbol{\theta}$ is a vector of unknown parameters of variance components. Model (4.1) adapts various study designs because one can specify a flexible covariance structure. One special case is shared frailty model, where each element of $\boldsymbol{\gamma}$ independently follow $\mathcal{N}(0, \theta^{\frac{1}{2}})$.

Following Yu et al. [22] the loglikelihood, conditional on covariates and frailty terms, is presented below:

$$\begin{aligned} l_c\Big[\lambda_0(\cdot), \boldsymbol{\alpha}, \boldsymbol{\beta}(\cdot) \mid \boldsymbol{\gamma}\Big] &= \sum_{i=1}^{m} l_i\Big[\lambda_0(\cdot), \boldsymbol{\alpha}, \boldsymbol{\beta}(\cdot) \mid \boldsymbol{\gamma}\Big] \\ &= \sum_{i=1}^{m} \Bigg\{ \sum_{j=1}^{n_i} \Big[log\lambda_0(t_{ij}) + \boldsymbol{z}_i(t_{ij})'\boldsymbol{\beta}(t_{ij}) + \boldsymbol{x}_i(t_{ij})'\boldsymbol{\alpha} + \boldsymbol{w}_i'\boldsymbol{\gamma}\Big] \\ &\quad - \Lambda_0(T_i) \exp\Big[\boldsymbol{z}_i(T_i)'\boldsymbol{\beta}(T_i) + \boldsymbol{x}_i(T_i)'\boldsymbol{\alpha} + \boldsymbol{w}_i'\boldsymbol{\gamma}\Big]\Bigg\}, \end{aligned} \tag{4.2}$$

where $\Lambda_0(t) = \int_0^t \lambda(u)du$. Estimation of $\boldsymbol{\beta}(t)$ and $\boldsymbol{\alpha}$ requires marginal log likelihood after integrating out frailty terms:

$$\begin{aligned} l_m &= \log \int \prod_{i=1}^{m} L_i\Big[\lambda_0(t), \boldsymbol{\beta}(t), \boldsymbol{\alpha} \mid \boldsymbol{\gamma}\Big] \times f(\boldsymbol{\gamma})d\boldsymbol{\gamma} \\ &= \log \int \frac{1}{(2\pi)^{\frac{m}{2}}|\Sigma(\boldsymbol{\theta})|^{\frac{1}{2}}} \exp\left\{\sum_{i=1}^{m} l_i\Big[\lambda_0(\cdot), \boldsymbol{\alpha}, \boldsymbol{\beta}(\cdot) \mid \boldsymbol{\gamma}\Big] - \frac{\boldsymbol{\gamma}'\Sigma(\boldsymbol{\theta})^{-1}\boldsymbol{\gamma}}{2}\right\} d\boldsymbol{\gamma}. \end{aligned} \tag{4.3}$$

It involves formulation of nonparametric time-varying coefficients and integration over frailty terms. In the next two subsections, we shall first discuss how to construct penalized B-spline estimates of $\boldsymbol{\beta}(t)$. Then we introduce approximation for marginal likelihood.

4.2.1 Penalized B-Spline estimation with alternative penalty matrices

We consider penalized B-spline to estimate time-varying coefficients. B-splines, introduced by DeBoor [8], is a popular type of regression splines consisting several piecewise polynomial functions. Time-varying coefficient $\beta_l(t), l = 1, \cdots, p$ is approximated by standard B-spline basis functions with dimension d. Curry and Schoenberg [7] implied any spline function can be written as a unique linear combination of the elements in the basis $B_k(t), k = 1, \cdots, d$. The expression of $\beta_l(t)$ is given as:

$$\beta_l(t) = \sum_{k=1}^{d} \eta_{lk} B_k(t) = B(t)' \boldsymbol{\eta}_l. \tag{4.4}$$

Let $\boldsymbol{\eta} = (\boldsymbol{\eta}_1', \cdots, \boldsymbol{\eta}_p')'$. After replacing the scalar covariate $\beta_l(t)$ by d-dimensional vector, model (4.1) turns into:

$$\lambda_i\Big[t|\boldsymbol{z}_i(t), \boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{\gamma}\Big] = \lambda_0(t) \exp\Big[\boldsymbol{z}_i(t)' \otimes B(t)'\boldsymbol{\eta} + \boldsymbol{x}_i'(t)\boldsymbol{\alpha} + \boldsymbol{w}_i'\boldsymbol{\gamma}\Big]. \tag{4.5}$$

It doesn't include time-varying coefficient anymore. Once basis functions are determined, $B(t)$ and $\boldsymbol{\eta}$ take place of $\boldsymbol{\beta}(\cdot)$ as unknown parameter and can be analyzed by standard procedure.

Dimension d is the summation of number of knots n and order r. n and r are of choice by researchers and have influence on the result. Higher order and more knots give more flexibility to better approach the true function. The downside of many knots and high order is overfitting. Cubic splines (order 4) has been widely used among literatures [11, 22] and is adopted in this study. In terms of knots selection, a widely used technique is to use a relatively large number of knots and add a penalty term for each spline function (e.g. Gray [11], Lin and Zhang [14], O'Sullivan [15], Yu et al. [22]). It gives penalized marginal likelihood:

$$l_{pm} = l_m\Big[\lambda_0(t), \boldsymbol{\alpha}, \boldsymbol{\eta}, \boldsymbol{\theta}\Big] - \frac{1}{2}\sum_{l=1}^{p} \lambda_l \boldsymbol{\eta}_l' P \boldsymbol{\eta}_l, \tag{4.6}$$

where $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_p)$ is a vector of smoothing parameters to control the tradeoff between the goodness of fit and the smoothness of the estimated functions. Smaller smoothing parameter place less constrain and the estimated coefficient function is more flexible. P is

a nonnegative definite smoothing matrix. There are two types of P being commonly applied, one is second derivative of basis functions $P = \int \boldsymbol{B}^{(2)}(t)\boldsymbol{B}^{(2)}(t)'dt$ [15]. With this, $\boldsymbol{\eta}_l P \boldsymbol{\eta}_l$ is the second derivative of the fitted curve. The other one is proposed by Eilers and Marx [9] as finite differences of the coefficients of adjacent B-splines. For example, $\boldsymbol{\eta}' P \boldsymbol{\eta} = \sum_{l=3}^{d} (\eta_l - 2\eta_{l-1} + \eta_{l-2})^2$ is the order of two penalty terms. The corresponding P is

$$\begin{pmatrix} 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & -2 & 1 \end{pmatrix}. \tag{4.7}$$

A strong connections has been derived between these two alternatives. Eilers and Marx [9] showed using a difference penalty on the coefficients simplified the procedure of maximization penalized likelihood. Comparison between them is conducted through simulation study in Section 4.5.

4.2.2 Double penalized partial likelihood and parameter estimation

With normally distributed frailties, there is no closed form for (4.6). We borrowed the idea of Laplace approximation from Breslow and Clayton [3], Ripatti and Palmgren [16], and Yu et al. [22]. The penalized marginal loglikelihood is approximated as:

$$l_{pm} \approx -\frac{1}{2} log|\Sigma(\boldsymbol{\theta})| - log|K''(\tilde{\boldsymbol{\gamma}})| + l_c\Big[\lambda_0(\cdot), \boldsymbol{\alpha}, \boldsymbol{\eta} \mid \tilde{\boldsymbol{\gamma}}\Big] - \frac{1}{2}\tilde{\boldsymbol{\gamma}}'\Sigma(\boldsymbol{\theta})^{-1}\tilde{\boldsymbol{\gamma}} - \frac{1}{2}\sum_{l=1}^{p} \lambda_l \boldsymbol{\eta}_l' P \boldsymbol{\eta}_l, \tag{4.8}$$

where $K(\boldsymbol{\gamma}) = -l_c\Big[\lambda_0(\cdot), \boldsymbol{\alpha}, \boldsymbol{\eta} \mid \boldsymbol{\gamma}\Big] + \frac{1}{2}\boldsymbol{\gamma}'\Sigma(\boldsymbol{\theta})^{-1}\boldsymbol{\gamma}$. $\tilde{\boldsymbol{\gamma}}$ is the solution to the first derivative $K'(\boldsymbol{\gamma}) = \mathbf{0}$ and $K''(\cdot)$ is the second derivative of $K(\cdot)$ with respect to $\boldsymbol{\gamma}$.

If $\boldsymbol{\theta}$ is fixed, the first term of (4.8) is constant with respect to $\boldsymbol{\alpha}, \boldsymbol{\eta}$. Ripatti and Palmgren [16] showed through simulation study that ignoring the second term in (4.8) resulted some information loss, but not too much to influence the estimation precision. It did simplify estimation procedure. Furthermore, if $\boldsymbol{\gamma}$ and $\boldsymbol{\eta}$ are considered as fixed effects parameters, (4.8) is double penalized log likelihood. The first term put penalty on extreme values of $\boldsymbol{\gamma}$ and the second is for penalizing smoothness of the time-varying coefficients functions. l_{pm} is then approximated by

$$l_{pm} \approx l_c\Big[\lambda_0(\cdot), \boldsymbol{\alpha}, \boldsymbol{\eta} \mid \tilde{\boldsymbol{\gamma}}\Big] - \frac{1}{2}\tilde{\boldsymbol{\gamma}}'\Sigma(\boldsymbol{\theta})^{-1}\tilde{\boldsymbol{\gamma}} - \frac{1}{2}\sum_{l=1}^{p}\lambda_l\boldsymbol{\eta}_l'P\boldsymbol{\eta}_l. \tag{4.9}$$

The first term of (4.9) is full log likelihood for a Cox model with given frailty terms. It can be maximized using penalized fixed effects partial likelihood (PPL). Double penalized partial likelihood (DPPL) is of form:

$$\begin{aligned}
&\sum_{i=1}^{m}\sum_{j=1}^{n_i}\Big\{\boldsymbol{z}_i(t_{ij})'\otimes B(t_{ij})'\boldsymbol{\eta} + \boldsymbol{x}_i(t_{ij})'\boldsymbol{\alpha} + \boldsymbol{w}_i'\boldsymbol{\gamma} \\
&- log\sum_{k\in R_{rij}}\exp\Big[\boldsymbol{z}_k(t_{ij})'\otimes B(t_{ij})'\boldsymbol{\eta} + \boldsymbol{x}_k(t_{ij})'\boldsymbol{\alpha} + \boldsymbol{w}_k'\boldsymbol{\gamma}\Big]\Big\} - \frac{1}{2}\boldsymbol{\gamma}'\Sigma(\boldsymbol{\theta})^{-1}\boldsymbol{\gamma} - \frac{1}{2}\sum_{l=1}^{p}\lambda_l\boldsymbol{\eta}_l'P\boldsymbol{\eta}_l.
\end{aligned} \tag{4.10}$$

The parameter estimate is obtained by maximizing (4.10). Note that smoothing parameters $\lambda_l, l = 1, \cdots, p$ shall be specified. Detail of smoothing parameter estimation is discussed in section 4.3. Differentiating 4.10 with respect to $(\boldsymbol{\alpha}', \boldsymbol{\eta}', \boldsymbol{\gamma}')'$ yields estimating equations as below:

$$\begin{aligned}
&\sum_{i=1}^{m}\sum_{j=1}^{n_i}\boldsymbol{z}_i(t_{ij})\otimes B(t_{ij}) - \frac{\exp\Big[\boldsymbol{z}_i(t_{ij})'\otimes B(t_{ij})'\boldsymbol{\eta} + \boldsymbol{x}_i'(t_{ij})\boldsymbol{\alpha} + \boldsymbol{w}_i'\boldsymbol{\gamma}\Big]\boldsymbol{z}_i(t_{ij})\otimes B(t_{ij})}{\sum\limits_{k\in R_{rij}}\exp\Big[\boldsymbol{z}_k(t_{ij})'\otimes B(t_{ij})'\boldsymbol{\eta} + \boldsymbol{x}_k(t_{ij})'\boldsymbol{\alpha} + \boldsymbol{w}_k'\boldsymbol{\gamma}\Big]} - \sum_{l=1}^{p}\lambda_l P\boldsymbol{\eta}_l = \mathbf{0} \\
&\sum_{i=1}^{m}\sum_{j=1}^{n_i}\boldsymbol{x}_i(t_{ij}) - \frac{\exp\Big[\boldsymbol{z}_i(t_{ij})'\otimes B(t_{ij})'\boldsymbol{\eta} + \boldsymbol{x}_i'(t_{ij})\boldsymbol{\alpha} + \boldsymbol{w}_i'\boldsymbol{\gamma}\Big]\boldsymbol{x}_i(t_{ij})}{\sum\limits_{k\in R_{rij}}\exp\Big[\boldsymbol{z}_k(t_{ij})'\otimes B(t_{ij})'\boldsymbol{\eta} + \boldsymbol{x}_k(t_{ij})'\boldsymbol{\alpha} + \boldsymbol{w}_k'\boldsymbol{\gamma}\Big]} = \mathbf{0} \\
&\sum_{i=1}^{m}\sum_{j=1}^{n_i}\boldsymbol{w}_i - \frac{\exp\Big[\boldsymbol{z}_i(t_{ij})'\otimes B(t_{ij})'\boldsymbol{\eta} + \boldsymbol{x}_i'(t_{ij})\boldsymbol{\alpha} + \boldsymbol{w}_i'\boldsymbol{\gamma}\Big]\boldsymbol{w}_i}{\sum\limits_{k\in R_{rij}}\exp\Big[\boldsymbol{z}_k(t_{ij})'\otimes B(t_{ij})'\boldsymbol{\eta} + \boldsymbol{x}_k(t_{ij})'\boldsymbol{\alpha} + \boldsymbol{w}_k'\boldsymbol{\gamma}\Big]} - \Sigma(\boldsymbol{\theta})^{-1}\boldsymbol{\gamma} = \mathbf{0}
\end{aligned} \tag{4.11}$$

Estimating equations in (4.11) can be solved by Newton-Raphson algorithm or other numerical methods. After obtaining $\widehat{\boldsymbol{\eta}}$, time-varying coefficient $\hat{\beta}_l(t)$ is estimated by $\hat{\boldsymbol{\eta}}_l'\boldsymbol{B}(t)$. Baseline intensity function estimation is:

$$\widehat{\lambda}_0(t) = \frac{\sum_{i=1}^{m}Y_i(t)dN_i(t)}{\sum_{i=1}^{m}Y_i(t)\exp\Big[\boldsymbol{z}_i(t)'\otimes B(t)'\hat{\boldsymbol{\eta}} + \boldsymbol{x}_i'(t)\hat{\boldsymbol{\alpha}} + \boldsymbol{w}_i'\hat{\boldsymbol{\gamma}}\Big]}, \tag{4.12}$$

where $Y_i(t)$ is a process recording whether subject i is at risk at time t.

4.2.3 Stratified time-varying coefficient model

In some studies, it is likely that subjects are sampled from subgroups of individuals with varying intensity functions. An effective way to accommodate this situation is to stratify the baseline functions into strata. It is assumed that baseline functions vary among strata while coefficients remain the same. The stratification model is of the following form:

$$\lambda_{ri}\Big[t \mid \boldsymbol{z}_{ri}(t), \boldsymbol{x}_{ri}, \boldsymbol{w}_{ri}, \boldsymbol{\gamma}\Big] = \lambda_{0r}(t) \exp\Big[\boldsymbol{z}'_{ri}(t)\boldsymbol{\beta}(t) + \boldsymbol{x}'_{ri}(t)\boldsymbol{\alpha} + \boldsymbol{w}'_{ri}\boldsymbol{\gamma}\Big], \tag{4.13}$$

where $r = 1, \cdots, R$ indicates stratum level. $i = 1, \cdots, m_r$ represents the i-th subject in stratum r. $\lambda_{0r}(t)$ is the baseline function for stratum r.

The estimating procedure for the stratified model is maximizing stratified DPPL:

$$\sum_{r=1}^{R}\sum_{i=1}^{m_r}\sum_{j=1}^{n_{ri}}\Big\{\boldsymbol{z}_{ri}(t_{rij})' \otimes B(t_{rij})'\boldsymbol{\eta} + \boldsymbol{x}_{ri}(t_{rij})'\boldsymbol{\alpha} + \boldsymbol{w}'_{ri}\boldsymbol{\gamma}$$
$$- log \sum_{k \in R_{rij}} \exp\Big[\boldsymbol{z}_k(t_{ij})' \otimes B(t_{rij})'\boldsymbol{\eta} + \boldsymbol{x}_k(t_{rij})'\boldsymbol{\alpha} + \boldsymbol{w}'_k\boldsymbol{\gamma}\Big]\Big\} - \frac{1}{2}\boldsymbol{\gamma}'\Sigma(\boldsymbol{\theta})^{-1}\boldsymbol{\gamma} - \frac{1}{2}\sum_{l=1}^{p}\lambda_l\boldsymbol{\eta}'_l P\boldsymbol{\eta}_l. \tag{4.14}$$

We estimate each baseline intensity using subjects of that stratum through form of below:

$$\widehat{\lambda}_{0r}(t) = \frac{\sum_{i=1}^{m_r} Y_{ri}(t)dN_{ri}(t)}{\sum_{i=1}^{m_r} Y_{ri}(t)\exp\Big[\boldsymbol{z}_{ri}(t)' \otimes B(t)'\hat{\boldsymbol{\eta}} + \boldsymbol{x}'_{ri}(t)\hat{\boldsymbol{\alpha}} + \boldsymbol{w}'_{ri}\hat{\boldsymbol{\gamma}}\Big]}. \tag{4.15}$$

4.3 Statistical inference

In this section, first we show the maximum DPPL estimator is not an unbiased estimator. But in some cases, bias converges to 0. In addition, we propose to estimate the DPPL estimators by fitting a frailty model. Frailty model representation provides a foundation for joint estimation procedure of smoothing parameters and variance component. It makes estimation procedure implemented easily by existing techniques.

4.3.1 Asymptotic distribution of maximum DPPL estimator

To approximate the distribution of the estimator, two additional assumptions beyond regularity conditions are required: the knots locations and number of parameters are held fixed as the sample size n increases; the family of models include the true distribution [11]. We

write the two penalty terms in (4.10) in a matrix format:

$$DPPL = PL - \frac{1}{2}\boldsymbol{\varphi}' \begin{pmatrix} 0 & 0 & \cdots & 0 & 0 \\ 0 & \lambda_1 P & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \lambda_p P & 0 \\ 0 & 0 & \cdots & 0 & \Sigma(\boldsymbol{\theta})^{-1} \end{pmatrix} \boldsymbol{\varphi}, \tag{4.16}$$

where $\boldsymbol{\varphi} = (\boldsymbol{\alpha}', \boldsymbol{\eta}', \boldsymbol{\gamma}')'$. We use R to denote the big penalty matrix for convenience. In general, as n increases, the magnitude of the contribution from the partial likelihood increase while the P is fixed. Thus the smoothing parameter $\boldsymbol{\lambda}$ will also need to increase at a rate of $O(n)$ to keep the degree of smoothing the same.

Let $\boldsymbol{\varphi}_0$ be the vector of true unknown parameters and $\hat{\boldsymbol{\varphi}}$ be the maximum DPPL estimator. For given $\boldsymbol{\lambda}$ and $\Sigma(\boldsymbol{\theta})$, score function and second derivative matrix are

$$S_p(\boldsymbol{\varphi}) = \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\varphi}} dppl = S(\boldsymbol{\varphi}) - R\boldsymbol{\varphi} \tag{4.17}$$

$$H_p(\boldsymbol{\varphi}) = \frac{\mathrm{d}^2}{\mathrm{d}\boldsymbol{\varphi} d\boldsymbol{\varphi}'} dppl - R = H(\boldsymbol{\varphi}) - R, \tag{4.18}$$

where $S(\boldsymbol{\varphi})$ is the score function of partial likelihood without penalty and $H(\boldsymbol{\varphi})$ is the second derivative matrix without penalty. Using standard arguments based on first-order expansion, under mild regularity conditions, it can be seen that:

$$\sqrt{n}(\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}_0) \approx -H_p(\boldsymbol{\varphi}_0)^{-1} S_p(\boldsymbol{\varphi}_0). \tag{4.19}$$

And

$$\frac{\sqrt{n} S_p(\boldsymbol{\varphi}_0)}{n} \xrightarrow{d} N(-\frac{R\boldsymbol{\varphi}_0}{\sqrt{n}}, I(\boldsymbol{\varphi}_0))$$
$$-\frac{I_p(\boldsymbol{\varphi}_0)}{n} \xrightarrow{p} I(\boldsymbol{\varphi}_0) + \frac{R}{n},$$

where $I(\boldsymbol{\varphi}_0)$ is the Fisher information matrix for partial likelihood. With above asymptotic convergence, the asymptotic distribution of the maximum DPPL estimator is given below:

$$\sqrt{n}(\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}_0) \xrightarrow{d} N\Big\{ -\Big[I(\boldsymbol{\varphi}_0) + \frac{R}{n}\Big]^{-1} \frac{R\boldsymbol{\varphi}_0}{\sqrt{n}}, \Big[I(\boldsymbol{\varphi}_0) + \frac{R}{n}\Big]^{-1} I(\boldsymbol{\varphi}_0) \Big[I(\boldsymbol{\varphi}_0) + \frac{R}{n}\Big]^{-1} \Big\}. \tag{4.20}$$

The estimation bias of time-varying coefficients depends on $\lambda_l P \boldsymbol{\eta}_l$. There are two special cases where bias is negligible. The first case is $\lambda_l P \boldsymbol{\eta}_l = \mathbf{0}$. This indicates the penalty

functions do not induce any bias in the estimate. The covariance matrix can be estimated by $-H_p(\hat{\boldsymbol{\varphi}})^{-1}H(\hat{\boldsymbol{\varphi}})H_p(\hat{\boldsymbol{\varphi}})^{-1}$. The second case is when $\lambda_l P\boldsymbol{\eta}_l$ is not $\mathbf{0}$ but the amount of smoothing is decrease as $n \to \infty$. The covariance matrix can be estimated similarly as the first case.

4.3.2 The frailty model representation

Lin and Zhang [14] showed connection between generalized additive mixed model and generalized linear mixed model by reparameterizing the spline coefficients as a combination of fixed and random effects. Using similar idea, we propose to rewrite $\boldsymbol{\eta}_l$ in (4.5) as

$$\boldsymbol{\eta}_l = \mathbf{1}\beta_{0l} + \boldsymbol{b}\beta_{1l} + A\boldsymbol{a}_l, \tag{4.21}$$

where $\mathbf{1}$ is a vector of 1 with length d and $A = L(L'L)^{-1}$. L is a $d \times (d-2)$ full row rank matrix satisfying $P = LL'$. $\boldsymbol{b}$ is a $d \times 1$ vector that satisfies $\boldsymbol{b}'\mathbf{1} = 0$ and $\boldsymbol{b}'A = \mathbf{0}$. In another word, $\boldsymbol{b}$ is the orthogonal complement of the space composed by $\mathbf{1}$ and A. Using the identity $\boldsymbol{\eta}_l' P\boldsymbol{\eta}_l = \boldsymbol{a}_l'\boldsymbol{a}_l$, DPPL (4.10) turns into:

$$\begin{aligned}
&\sum_{i=1}^{m}\sum_{j=1}^{n_i}\Big\{\boldsymbol{z}_i(t_{ij})' \otimes B(t_{ij})'\boldsymbol{\eta} + \boldsymbol{x}_i(t_{ij})'\boldsymbol{\alpha} + \boldsymbol{w}_i'\boldsymbol{\gamma} \\
&- log\sum_{k\in R_{rij}} \exp\Big[\boldsymbol{z}_k(t_{ij})' \otimes B(t_{ij})'\boldsymbol{\eta} + \boldsymbol{x}_k(t_{ij})'\boldsymbol{\alpha} + \boldsymbol{w}_k'\boldsymbol{\gamma}\Big]\Big\} - \frac{1}{2}\boldsymbol{\gamma}'\Sigma(\boldsymbol{\theta})^{-1}\boldsymbol{\gamma} - \frac{1}{2}\boldsymbol{a}'\Lambda^{-1}\boldsymbol{a},
\end{aligned} \tag{4.22}$$

where $\boldsymbol{a} = (\boldsymbol{a}_1', \cdots, \boldsymbol{a}_p')'$ and $\Lambda = diag\Big(\frac{1}{\lambda_1}\boldsymbol{I}, \cdots, \frac{1}{\lambda_p}\boldsymbol{I}\Big)$.

Plugging equation (4.21) into (4.1), (4.22) suggests that the maximum DPPL estimator can be obtained by fitting the following frailty model:

$$\lambda_i\Big[t \mid \boldsymbol{x}_i^{*'}(t), \boldsymbol{w}_i^{*'}, \boldsymbol{\gamma}^*\Big] = \lambda_0(t)\exp\Big[\boldsymbol{x}_i^{*'}(t)\boldsymbol{\beta} + \boldsymbol{w}_i^{*'}\boldsymbol{\gamma}^*\Big], \tag{4.23}$$

where $\boldsymbol{x}_i^*(t) = \Big(\boldsymbol{x}_i(t)',\ \boldsymbol{z}_i(t)' \otimes B(t)'\boldsymbol{I} \otimes \mathbf{1},\ \boldsymbol{z}_i' \otimes B(t)'\boldsymbol{I} \otimes \boldsymbol{b}\Big)'$, and $\boldsymbol{w}_i^* = \Big(\boldsymbol{z}_i(t)' \otimes B(t)'\boldsymbol{I} \otimes A,\ \boldsymbol{w}_i'\Big)'$. $\boldsymbol{\beta} = \Big(\boldsymbol{\alpha}',\ \boldsymbol{\beta}_0',\ \boldsymbol{\beta}_1'\Big)'$ is a vector of coefficients, with $\boldsymbol{\beta}_0 = (\beta_{01}, \cdots, \beta_{0p})'$ and $\boldsymbol{\beta}_0 = (\beta_{11}, \cdots, \beta_{1p})'$. $\boldsymbol{\gamma}^* = \Big(\boldsymbol{a}',\ \boldsymbol{\gamma}'\Big)'$. $\boldsymbol{a}$ and $\boldsymbol{\gamma}$ are independent random effects with distributions $\boldsymbol{a} \sim \mathcal{N}(\mathbf{0}, \Lambda)$ and $\boldsymbol{\gamma} \sim \mathcal{N}(\mathbf{0}, \Sigma(\boldsymbol{\theta}))$. The maximum DPPL estimator $\widehat{\beta}_l(t)$ is calculated as $B(t)'\mathbf{1}\widehat{\beta}_{0l} + B(t)'\boldsymbol{b}\widehat{\beta}_{1l} + B(t)'A\hat{\boldsymbol{a}}_l$, which is a linear combination of the maximum penalized partial likelihood estimator of the fixed effect and the random effects $\hat{\boldsymbol{a}}$ in Ripatti and Palmgren [16] .

4.3.3 Inference on smoothing parameter and variance component

Statistical inference the nonparametric functions $\boldsymbol{\beta}_l(t)$ relies on the estimation of smoothing parameter and the inference on variance parameter $\boldsymbol{\theta}$. In the previous section, we estimated time-varying coefficients as a linear combination of fixed effect and random effects. Smoothing parameters $\boldsymbol{\lambda}$ was treated as extra variance components. Thus inference on smoothing parameters can be conducted similarly to variance component in frailty model. Plugging maximum DPPL estimators into (4.6) results profile likelihood function for $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$:

$$l \approx -\frac{1}{2}log|\Lambda| - \frac{1}{2}log|\Sigma(\boldsymbol{\theta})| - \frac{1}{2}\log|K''(\hat{\boldsymbol{\gamma}^*})| - \frac{1}{2}\hat{\boldsymbol{\gamma}}\Sigma(\boldsymbol{\theta})^{-1}\hat{\boldsymbol{\gamma}} - \frac{1}{2}\hat{\boldsymbol{a}}\Lambda^{-1}\hat{\boldsymbol{a}}, \tag{4.24}$$

where K was derived in 4.2.2. Here we propose to use $K''_{DPPL}(\hat{\boldsymbol{\gamma}^*}) = (\partial^2 DPPL)/(\partial\boldsymbol{\gamma}^*\partial\boldsymbol{\gamma}^{*'})$. Estimating equations of $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ can be derived by taking the first derivative of (4.24). The corresponding Fisher information matrix can be derived by differentiating (4.24) twice and taking the expectation with respect to $\boldsymbol{\gamma}^{*'}$, similar to equation (8) in Ripatti and Palmgren [16].

In this Chapter, we focus on the frailties that are i.i.d with $\mathcal{N}(0, \theta^{\frac{1}{2}})$. Thus $\Sigma(\boldsymbol{\theta})$ is a diagonal matrix with elements θ. $\boldsymbol{w}_i$ is a sparse vector has value 1 at position of its order and 0 elsewhere. $\boldsymbol{w}_i'\boldsymbol{\gamma}$ turns to be γ_i. Solutions to estimating equations are:

$$\begin{aligned} \hat{\theta} &= \frac{\hat{\boldsymbol{\gamma}}'\hat{\boldsymbol{\gamma}} + tr(K''^{-1}_{DPPL}(\hat{\boldsymbol{\gamma}}))}{m} \\ \widehat{\lambda}_l &= \frac{d-2}{\hat{\boldsymbol{a}}_l'\hat{\boldsymbol{a}}_l + tr(K''^{-1}_{DPPL}(\hat{\boldsymbol{a}}_l))}, \; l = 1, \cdots, p. \end{aligned} \tag{4.25}$$

where $K''^{-1}_{DPPL}(\hat{\boldsymbol{\gamma}})$ is the submatrix of K''^{-1}_{DPPL} corresponding to $\boldsymbol{\gamma}$. $K''^{-1}_{DPPL}(\hat{\boldsymbol{a}})$ is the remaining submatrix of K''^{-1}_{DPPL}. Variance is estimated by the inverse of estimated Fisher information matrix:

$$\begin{aligned} Var(\hat{\theta}) &= 2\hat{\theta}^2\Big[m + \frac{tr(K''^{-1}_{DPPL}(\hat{\boldsymbol{\gamma}})K''^{-1}_{DPPL}(\hat{\boldsymbol{\gamma}}))}{\hat{\theta}^2} - \frac{2tr(K''^{-1}_{DPPL}(\hat{\boldsymbol{\gamma}}))}{\hat{\theta}^2}\Big]^{-1} \\ Var(\hat{\lambda}_l) &= 2\widehat{\lambda}_l^2\Big[3(d-2) + \widehat{\lambda}_l^2 tr\Big(K''^{-1}_{DPPL}(\hat{\boldsymbol{a}}_l)K''^{-1}_{DPPL}(\hat{\boldsymbol{a}}_l)\Big)\Big]^{-1}. \end{aligned} \tag{4.26}$$

4.3.4 Computation

The maximization of DPPL is done in two steps. First, an initial value for $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ is guessed. (4.11) is solved using the Newton-Raphson technique or other numerical methods by given $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$. Then $\widehat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{a}}$ are fixed at the values obtained and (4.25) is solved to find a new value for $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$. The two steps are iterated until convergence. Once convergence is

achieved, the cumulative baseline hazard can be estimated using (4.12), which is the same as the Andersen-Gill [2] model with an offset of estimated random effects.

4.3.5 Test of time-varying coefficient

The objective of this subsection is to elaborate three tests of particular interest: test for no effect, test for time-fixed effect, and test for linear time-varying effect. We start with a general linear hypothesis test:

$$H_0 : C(\boldsymbol{\alpha}', \boldsymbol{\eta}')' = \mathbf{0},\ H_a : C(\boldsymbol{\alpha}', \boldsymbol{\eta}')' \neq \mathbf{0} \tag{4.27}$$

where $(\boldsymbol{\alpha}', \boldsymbol{\eta}')'$ is the combined time-fixed coefficients and time-varying coefficients. C is a matrix with full row rank R. Gray [11] suggested a Wald test statistic of the form:

$$[C(\boldsymbol{\alpha}', \boldsymbol{\eta}')']'(CI_p^{-1}C')^{-1}[C(\boldsymbol{\alpha}', \boldsymbol{\eta}')'], \tag{4.28}$$

where $I_p = -H_p$ is negative penalized hessian matrix. Under H_0, the distribution of the test statistic is asymptotically

$$\sum_{l=1}^{R} \mu_l Z_l^2, \tag{4.29}$$

where Z_l's are independent standard normal random variables and μ_l's are the eigenvalues of the matrix $(CI_p^{-1}C')^{-1}CVC$. Thus the generalized degrees of freedom of the test statistic is defined as:

$$df = trace[(CI_p^{-1}C')^{-1}CVC], \tag{4.30}$$

where $V = -H_p^{-1}HH_P^{-1}$. We shall reject the null hypothesis if the test statistic exceeds critical value.

In the case of testing no effect, for example: the l-th covariate doesn't have effect on the intensity function, we are seeking evidence against $H_0 : \beta_l(t) = 0$. This null hypothesis can be represented as:

$$H_0 : \boldsymbol{\eta}_l = \mathbf{0},\ H_a : \boldsymbol{\eta}_l \neq \mathbf{0}. \tag{4.31}$$

C matrix for this test is an identity matrix.

Another hypothesis that is often of interest is constant effect. For example, the l-th covariate has the same magnitude of effect β_l on the intensity function over time. The formatted hypothesis $H_0 : \beta_l(t) = c$ where c is some constant and $H_a : \beta_l(t) \neq c$ can be

further represented as:

$$\begin{aligned} &H_0 : \eta_{l,1} = \cdots = \eta_{l,d} \\ &H_a : \text{at least one } \eta_{l,i} \text{ is not the same as others.} \end{aligned} \tag{4.32}$$

C matrix for this test is a $(d-1) \times d$ matrix as follows:

$$\begin{pmatrix} 1 & 0 & \cdots & 0 & -1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -1 \end{pmatrix}. \tag{4.33}$$

The last hypothesis is for linearity of time-varying coefficient. For instance, in order to test whether the effect of l-th covariate is a linear function over time, we can formated the null hypothesis as $H_0 : \beta_l(t) = c_0 + c_1 t$, where c_0 and c_1 are some constant numbers. It is more convenient if we test on the first derivative of $\beta_l(t)$ as a constant number as:

$$H_0 : \frac{\partial B(t)'}{\partial t}\boldsymbol{\eta} = c \tag{4.34}$$

DeBoor [8] gives a simple formula for derivatives of B-splines:

$$h \sum_{i=1}^{d} \eta_{li} B_i'(t,3) = -\sum_{i=1}^{d-1} \triangle\eta_{i+1} B_i(t,2), \tag{4.35}$$

where h is the difference between two consecutive knots. $\triangle\eta_{i+1} = \eta_{i+1} - \eta_i$ is the difference of two successive coefficients. $B_i(t,2)$ is a B-spline function of second order at value t. By using this relationship, the null hypothesis (4.34) is equivalent as testing $\triangle\eta_{l,1} = \cdots = \triangle\eta_{l,d-1}$. In this case C is a $(d-2) \times d$ matrix as follows:

$$\begin{pmatrix} 1 & -2 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -2 & 1 \end{pmatrix}. \tag{4.36}$$

Of all the above three test, it is essentially the same to conduct a test on (β_{0l}, β_{0l}, $\boldsymbol{a}_l$). The corresponding C matrix needs to be multiplied by $\begin{pmatrix} \mathbf{1} & \boldsymbol{b} & A \end{pmatrix}$.

4.4 Application to 100-Car Naturalistic Driving Study

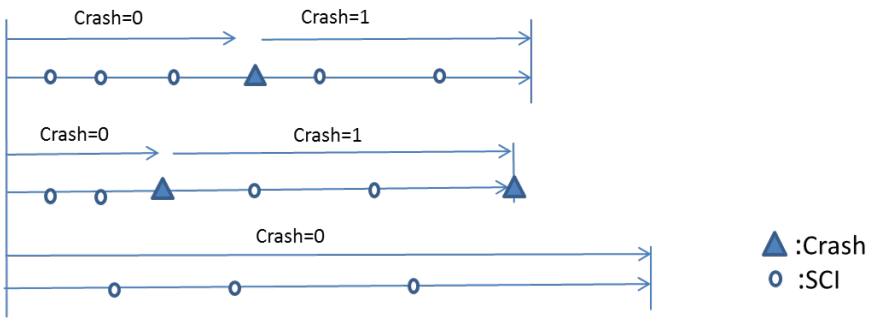


Figure 4.1: Data collection structure.

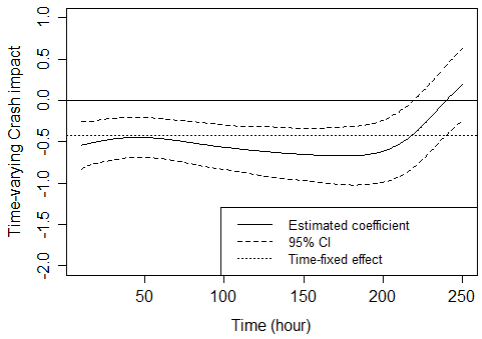
In this section, we apply the time-varying coefficient model to the 100-Car Naturalistic Driving Study (NDS). The objective is to evaluate the pattern of crash influence on driving risk over time. Data collection structure is shown in Figure 4.1. Each horizontal line represents record of one driver. Drivers are subject to different numbers of crashes, NCs, and SCIs at different time points throughout study. Thus, it is important to record all timestamps. We focus on the actual driving time. Non-driving time when the vehicle is not in use has been excluded. SCIs and NCs are treated as two counting processes and evaluated separately. The process of SCIs is explained here; a similar setting is used for NCs. Each driving period is divided into two phases based on its relationship with crashes: before the first crash (coded as 0) and between the first and second crash (coded as 1). Since this is a one-year study, there are only 12 drivers had two crashes. After careful review, only 4 of them had NCs after second crash. Thus we consider to evaluate the first crash effect only. Observations after second crashes are treated as censored. Driving period is taken into account as a covariate, working as an external and independent factor on SCI/NC intensity. To account for potential confounding and interacting effects, gender and age of the driver when first enrolled in the study are evaluated as time-fixed covariates. For SCI/NC, final model is given below:

$$\lambda_{ri}(t) = \lambda_{r0} \exp\Big[\beta_1 G_{ri} + \beta_2 Age_{2ri} + \beta_3(t - c_{ri})I_{ri}(t) + \beta_4(t - c_{ri})G_{ri}I_{ri}(t) + \gamma_{ri}\Big], \quad (4.37)$$

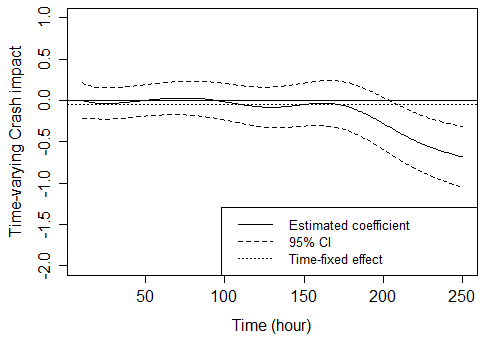
where $G_{ri} = 1$ for male driver and 0 for female driver. $\beta_3(t - c_{ri})$ is crash influence on female driver and $\beta_3(t - c_{ri}) + \beta_4(t - c_{ri})$ is crash influence on male driver.

Estimated crash influence on male drivers and female drivers are presented in Figure 4.2. In addition to time-varying model, we also fit a naive frailty model by taking crash influence as a time-fixed coefficient. As it shows from top two plots, SCI intensity drops after crash

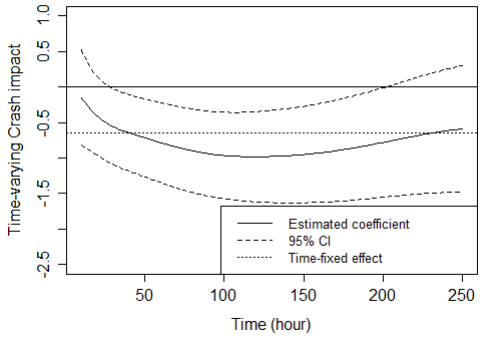
for male drivers and gradually increase after 200 hours. On contrary, SCI intensity stays the same after crash and then decrease after 200 hours for female drivers. Both effects are consistent with results from constant coefficient model. In the bottom two plots, male drivers experience lower driving risk after 50 hours of crash till 150 hours. No significant change after that. For female drivers, there is no significant decrease in NC intensity after crash. In summary, driving risk of male drivers tend to decrease after crash first and then increase. But we have not found similar pattern for female drivers. Compared with the naive analysis, the time-varying coefficient analysis provides more information.



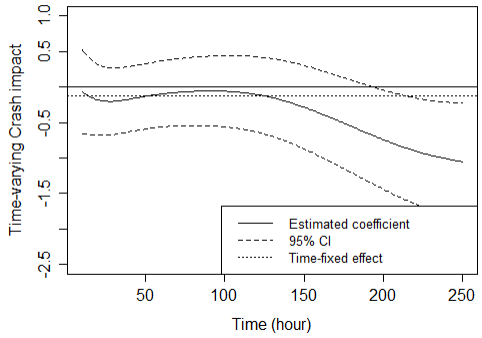
(a) Estimated crash influence on SCI male drivers



(b) Estimated crash influence on SCI female drivers



(c) Estimated crash influence on NC male drivers



(d) Estimated crash influence on NC female drivers

Figure 4.2: Crash influence on SCI and NC: $\widehat{\beta}(t)$, solid; 95% pointwise confidence interval, dashed; fixed effect estimation, dotted.

4.5 Simulation study

In this section, we conduct a simulation study to evaluate the performance of the proposed method. We simulate intensity models in two scenarios. Details will be provided later in

Table 4.1: Parametric coefficient estimates of time-varying model applied to 100-Car NDS

	Risk factor	Estimates	SE	Intensity rate ratio	p-value
NC	Gender	-0.173	0.118	0.811	0.144
	Age	-0.007	0.004	0.992	0.084
	σ^2	0.437	0.101		
SCI	Gender	-0.180	0.056	0.841	.001
	Age	-0.009	0.002	0.993	$< .001$
	σ^2	0.968	0.145		

this section. In each scenario, we examine both estimation precision and power of test for time-varying coefficient.

4.5.1 Simulation setup

The simulation procedure were designed to mimic the 100-Car Data. It consists generating censor time, crash time and events time for each subjects as described below.

1. Censor time is set as 4 for each subject.

2. For each subject, we generate a crash time c_i based on the following intensity function:

$$\lambda_i(t) = \frac{1}{2},\ t \leq 4. \tag{4.38}$$

   $\frac{1}{2}$ was selected based on the relationship between crash time and study period from 100-Car data, where we observed one crash in every 150 hours of driving and the average study time was around 300. The above intensity function assumes subjects performing similarly, without any discrepancy among subgroups (for example, gender difference). Crash intensity was restricted to be constant over time and crashes were considered to occur independently. If one subject had c_i that was greater than censor time, c_i would be censored. We used a time-varying indicator function, $I_i(t) = (t > c_i)$, to denote the relationship between time t and crash time.

3. Recurrent events time are then generated from the following intensity function:

$$\lambda_{ri}(t) = c_r t^{k_r - 1} \exp\left[\beta_1 x_{1ri} + \beta_2 x_{2ri} + \beta_3 (t - c_{ri}) I_{ri}(t) + \beta_4 (t - c_{ri}) x_{1ri} I_{ri}(t) + \gamma_{ri}\right], \tag{4.39}$$

   where:

   (a) $r = 1$ or 2 indicats stratum level. $r = 1$ represents no-crash group and $r = 2$ represent one-crash group.

(b) Baseline intensity functions follow Weibull distribution, where two parameters c and r play a critical role. They may vary from stratum to stratum, as denoted by c_r and k_r. $k_r > 1$ indicates an increasing rate over time. $k_r = 1$ refers to a constant rate, and $k_r < 1$ means a decreasing rate.

(c) c_{ri} is the crash time generated in the previous step.

(d) x_{1ri} is a binary covariate, with 50% probability to be 0/1.

(e) x_{2ri} is a continuous covariate following uniform distribution from -1 to 1.

(f) I_{ri} is considered a time-varying binary covariate. It takes a value of 1 when t is larger than the crash time.

(g) β_1 and β_2 are time-fixed effects, with value of 0.5, -0.3 respectively.

(h) $\beta_3(t - c_{ri})$ and $\beta_4(t - c_{ri})$ are two time-varying effects. Since c_{ri}'s are different across subjects, we focus evaluating those effects as functions of time after crash.

(i) $\gamma_{ri} \sim N(0, \sigma), r = 1, or 2;\ i = 1, \cdots, n_r$ are independent frailty terms. In most settings, σ is set as 0.5, representing a moderate heterogeneity.

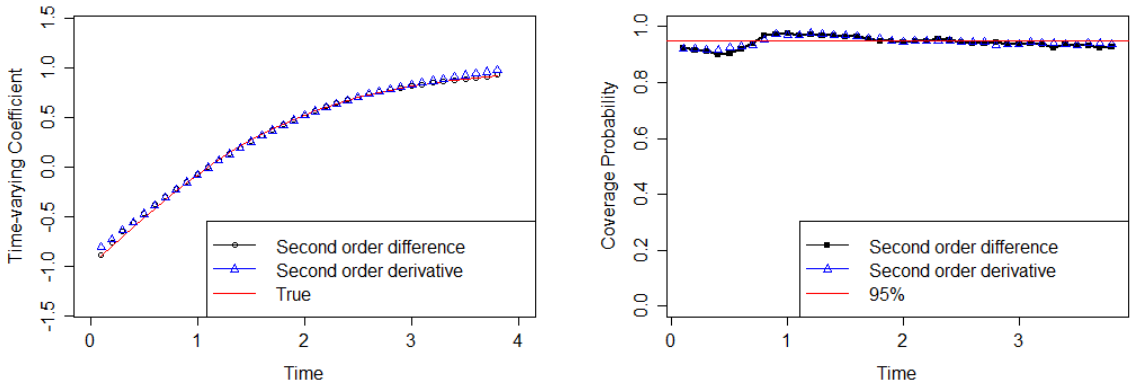
Two scenarios will be mainly discussed in this section. In the first scenario, interaction effect in (4.39) is not included, thus there is only one time-varying coefficient. In the second scenario, both time-varying effects are included and estimated. In order to cover a certain range of parameter space, 6 different settings of baseline parameters and sample size combinations are explored in each scenario. Within each setting, 500 realizations are generated and two models are implemented: a stratified time-varying coefficient model and a non-stratified model. Because of space limits, we only provide results from stratified time-varying coefficient model and selected settings.

4.5.2 Simulation result

4.5.2.1 Recurrent event model with one time-varying effect

In the first setting of one time-varying effect model, we generated time-varying coefficient $\beta_3(t) = \frac{4}{1+exp(-t)} - 3$ as shown in left panel of Figure 4.3. It is a logistic function gradually increasing from -0.9 to 0.9 over time period $(0, 4)$. We set the scale and shape parameters as $c = (1, 1.5)$ and $k = (1, 0.9)$. It indicates a constant rate of 1 event per time unit for subjects in stratum I. Subjects in stratum II experience decreasing rate over time. We generated data from 50 subjects which yields about 9 events per subject. Figure 4.3 shows simulation results over 500 replications. In the left panel, the solid red line is the true logistic function. The circles are average estimating effects based on second-corder difference penalty matrix

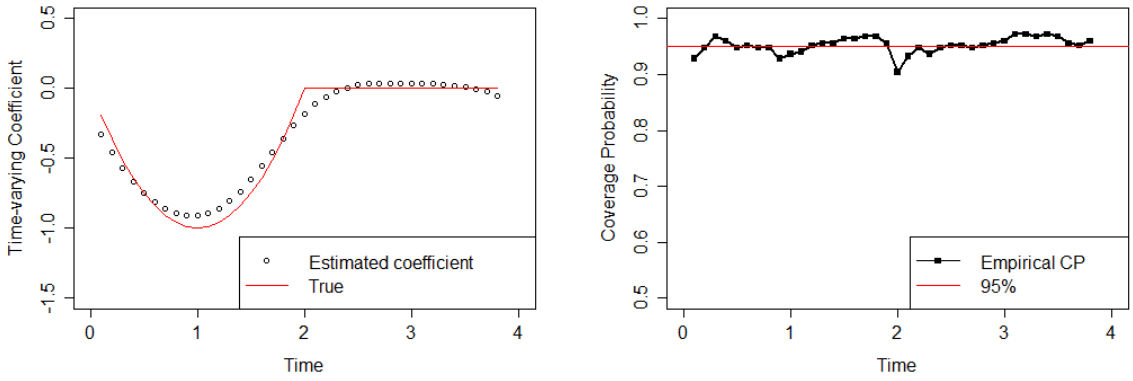
as mentioned in subsection 4.2.1 and the triangles are estimated effects based on second-order derivative penalty matrix. Both provides very precise estimation. Pointwise empirical coverage probability (CP) of two types of penalty matrices are compared in the right panel. Again, there performance are close with CP slightly less than nominal value 95%.



(a) Varying coefficient estimation (b) Pointwise empirical CP

Figure 4.3: Simulation results for logistic function from stratified data

To examine the performance of the proposed method when true effect has a stronger curvature, in the second setting, we set the effect to be $(t^2 - t) \times I(t < 2) + 0 \times I(t >= 2)$ as shown in the left panel of Figure 4.4. In the range of $(0, 2)$, it decreases first and increases to 0 and maintain 0 after 2. This function is selected because of out the hypothesis: crash effect decreases first and then increase. Simulation result shows close agreement between estimated effect and true function. Pointwise empirical CP is slightly smaller than 95%.



(a) Varying coefficient estimation (b) Pointwise empirical CP

Figure 4.4: Simulation results for piecewise polynomial function from stratified data

Table 4.2 lists results of time-fixed coefficients, variance component, and smoothing parameter from various settings. In setting I, the bias of two time-fixed coefficients are -0.02 and 0.01, which are about 4% and 3%. The empirical coverage probabilities of 95% confidence intervals using the estimated standard error both are 94%, which are close to nominal value. Bias for variance component σ^2 is -0.02 (8% bias). The estimated standard derivation of β_1, β_2, and σ^2 (SEM) are close to the empirical standard derivations (SE).

Table 4.2: Simulation results for parametric coefficients estimates in time-varying coefficient model

Parameter	True value	Mean	Bias	SE %	SEM△	CP*
Setting I: logistic function						
β_1	0.5	0.48	-0.02	0.18	0.18	94%
β_2	-0.3	-0.29	0.01	0.17	0.16	94%
θ	0.25	0.23	-0.02	0.06	0.08	92%
λ		38				
β_1	0.5	0.44	-0.06	0.29	0.29	94%
β_2	-0.3	-0.30	0.00	0.27	0.26	94%
θ	1	0.84	-0.16	0.21	0.21	81%
λ		97				
Setting II: Piecewise polynomial function						
β_1	0.5	0.48	-0.02	0.2	0.19	94%
β_2	-0.3	-0.29	0.01	0.17	0.17	94%
θ	0.25	0.23	-0.02	0.08	0.09	92%
λ		14				
β_1	0.5	0.5	0	0.29	0.30	94%
β_2	-0.3	-0.31	-0.01	0.27	0.27	94%
θ	1	0.85	-0.15	0.21	0.22	83%
λ		17				

%: Empirical standard error

△: Mean of standard error

*: Coverage probability

Table 4.3 presents power of proposed tests in subsection 4.3.5 under difference true functions. When true effect is a logistics function, power of test for no effect and constant effect is above 99%.

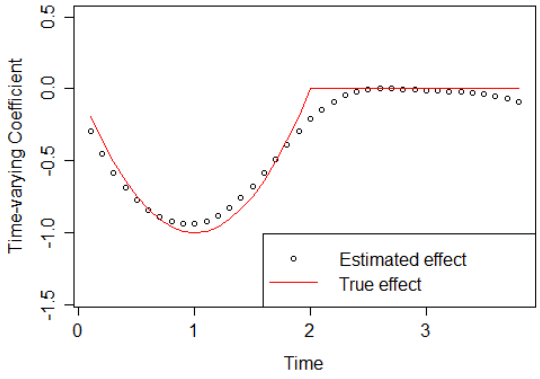
Table 4.3: Empirical power/ type I error of tests for covariate effects

	Null hypothesis		
True function	No effect	Constant	Linear
Logistic	99%	99%	98%
Piecewise Polynomial	98%	88%	84%

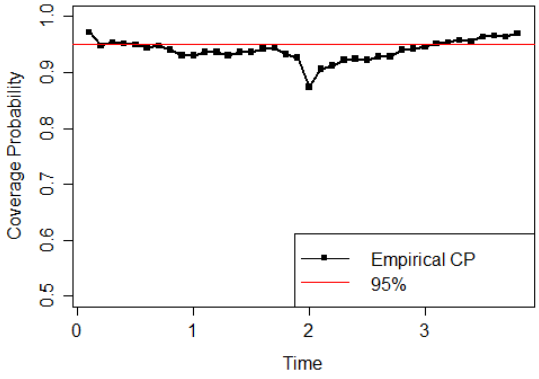
4.5.2.2 Two additive time-varying effect model

In the first setting of two time-varying effects model, we generated time-varying coefficients using previous piecewise polynomial function and logistic function. Baseline intensity functions are set to be difference across strata with $c = (1, 1.5)$ and $k = (1, 0.9)$. Each replication is simulated from 50 subjects.

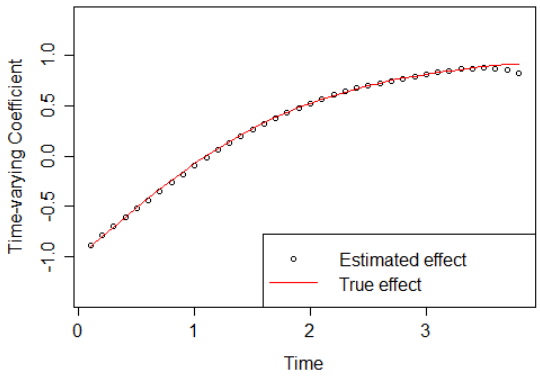
In the left panel of Figure 4.5, we observe close agreement between the true functions and average estimates for both coefficients. As given on the right panel, pointwise empirical coverage probabilities are close to 95% nominal value.



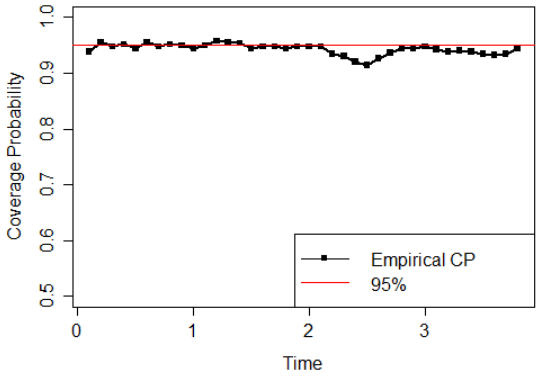
(a) Varying coefficient estimation for first effect



(b) Pointwise empirical CP for first effect



(a) Varying coefficient estimation for second effect



(b) Pointwise empirical CP for second effect

Figure 4.5: Setting I: simulation results for two time-varying coefficient functions from stratified data

Table 4.4 lists the simulation results for time-fixed coefficients, variance component, and smoothing parameters from various settings. Bias of two time-fixed coefficients are small (from 4% to 10%). The empirical coverage probabilities of 95% confidence intervals using the estimated standard error are 94% and 93%. Bias of variance component σ^2 is small (4% bias) when true standard deviation is small. It becomes larger (23%) when true standard deviation is larger. The estimated standard derivation of σ^2 (SEM) is close to the empirical standard derivation (SE).

Table 4.4: Simulation results for parametric coefficients estimates in two effects model

Parameter	True	Mean	Bias	SE$^{\%}$	SEM$^{\triangle}$	CP*
Setting I: $n = 100$, $\theta = 0.25$						
β_1	0.5	0.48	-0.02	0.16	0.16	95%
β_2	-0.3	-0.29	0.01	0.11	0.11	96%
θ	0.25	0.24	-0.01	0.03	0.06	99%
λ		(3.5, 3.4)				
Number of events per subject		8				
Setting II: $n = 50$, $\theta = 0.25$						
β_1	0.5	0.48	-0.02	0.25	0.23	94%
β_2	-0.3	-0.29	-0.01	0.17	0.16	93%
θ	0.25	0.24	-0.01	0.07	0.08	95%
λ		(7.4, 6)				
Number of events per subject		8				
Setting III: $n = 50$, $\theta = 1$						
β_1	0.5	0.45	-0.05	0.34	0.31	92%
β_2	-0.3	-0.28	0.02	0.26	0.25	94%
θ	1	0.77	-0.23	0.20	0.20	68%
λ		(5.5, 4.8)				
Number of events per subject		11				

%: Empirical standard error

$\triangle$: Mean of standard error

*: Coverage probability

Table 4.5 presents power of proposed test in subsection 4.3.5 under difference true functions. When true effect is a polynomial function, power of test for no effect and constant effect are about 78%. However, power for linear effect is 47%. The test for whether two coefficients model are equal has power about 87%. When true effect is constant at 0 , type I error for no effect and constant effect is about 14%, which is much higher than nominal value 5%.

Table 4.5: Empirical power/ type I error of tests for covariate effects

		Null hypothesis			
Setting	True function	No effect	Constant	Linear	Equality
$n = 50$, $\sigma = .5$	Polynomial	78%	78%	47%	87%
	Logistic	99%	98%	37%	
$n = 50$, $\sigma = .5$	Constant 0	14%	14%	14%	90%
	Logistic	99%	99%	37%	
$n = 50$, $\sigma = .5$	Constant -1	88%	10%	11%	99%
	Logistic	99%	98%	40%	
$n = 50$, $\sigma = .5$	Logistic	97%	97%	39%	9%
	Logistic	99%	98%	41%	

4.6 Conclusion and discussion

In this Chapter, we evaluated both time-varying and time-fixed effects in recurrent events model. Gaussian frailty terms are incorporated to accommodate correlation within subject. Penalized B-spline is adopted to approximate time-varying coefficient. We proposed to use regular frailty model framework to estimate time-varying coefficient by expressing time-varying coefficients as a linear combination of fixed effect and random effects. The smoothing parameter associated with time-varying coefficients is then treated as an extra variance component. We estimate nonparametric functions and jointly estimated smoothing parameters and variance component by using marginal partial likelihood. We studied the asymptotic distribution of proposed estimates. Penalty terms introduce bias to the estimation. But the bias is close to zero if sample size is relatively large. We used Wald-type test for several popular test: no effect, constant effect, linear effect, and equivalence test between two time-varying effects.

The simulation study indicates small bias of the proposed model and relatively good coverage probability for both time-fixed effects and time-varying effects. Power of tests vary from settings to settings. If the true function is similar to the one tested (e.g., true function is logistic and test function is linear), test is not very powerful. If they are quite different, we observed high power around 95%. We also compared two types of penalty matrices and did not find one is uniformly better than the other. Simulation study also demonstrated that the stratified frailty model is capable of accommodating possible variation among groups

without losing power to test for effects of interest. If subjects behave differently among various levels, aggregating them together will mask the effect at the individual level. We also observed robust performance of the stratified frailty model when subjects are not from different levels.

We applied the model to evaluate influence of crashes on driving risk using 100-Car NDS data. SCIs and NCs are used as markers for driving risk. Driving risk of male drivers tend to decrease after crash first and then increase. But we did not find similar for female drivers. Compared with the time-fixed model, the time-varying coefficient analysis provided deeper insight of the data. These findings provide crucial information for understanding drivers' response to dramatic driving events and can be critical for development safety education programs and safety counter measures.

There are a couple of limitations of this study. First, the individual driver risk variation might be confounded with the observed effect. Second, the study is based on a relative small number of crashes with mild crash severity. With larger NDS data sets becoming available, such as the Second Strategic Highway Research Program (SHRP 2) Naturalistic Driving Study, more concrete evidence will be available on the influence of crashes on driver behavior and potentially the influence of crashes by severity.

Bibliography

[1] L. D. Amorim, J. Cai, D. Zeng, and M. L. Barreto. Regression splines in the time-dependent coefficient rates model for recurrent event data. *Statistics in Medicine*, 27 (28):5890–5906, 2008.

[2] P. K. Andersen and R. D. Gill. Cox's Regression Model for Counting Processes: A Large Sample Study. *The Annals of Statistics*, 10(4):1100–1120, 1982.

[3] N. Breslow and D. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88:9–25, March 1993.

[4] J. Cai, J. Fan, H. Zhou, and Y. Zhou. Hazard models with varying coefficients for multivariate failure time data. *The Annals of Statistics*, 35(1):324–354, 2007.

[5] J. Cai, J. Fan, J. Jiang, and H. Zhou. Partially linear hazard regression with varying coefficients for multivariate survival data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):141–158, 2008.

[6] Z. Cai and Y. Sun. Local linear estimation for time-dependent coefficients in cox's regression models. *Scandinavian Journal of Statistics*, 30(1):93–111, 2003.

[7] H. Curry and I. Schoenberg. On plya frequency functions iv: The fundamental spline functions and their limits. *Journal dAnalyse Mathmatique*, 17(1):71–107, 1966.

[8] C. DeBoor. *A practical guide to splines*. Springer, 1978.

[9] P. Eilers and B. Marx. Flexible smoothing with B-splines and penalties. *Statistical Science*, 11:89–121, 1996.

[10] J. Fan, I. Gijbels, and M. King. Local likelihood and local partial likelihood in hazard regression. *The Annals of Statistics*, 25:1661–1690, 1997.

[11] R. J. Gray. Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association*, 87(420): 942–951, 1992.

[12] S. G. Klauer, F. Guo, B. G. Simons-Morton, M. C. Ouimet, S. E. Lee, and T. A. Dingus. The prevalence and risk of cell phone and other secondary tasks as observed in crashes

and near-crashes with novice and experienced drivers. *The New England Journal of Medicine*, 370:54–59, 2014.

[13] D. Y. Lin, L. J. Wei, I. Yang, and Z. Ying. Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 62(4):711–730, 2000.

[14] X. Lin and D. Zhang. Inference in generalized additive mixed modelsby using smoothing splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61 (2):381–400, 1999.

[15] F. O'Sullivan. Nonparametric estimation of relative risk using splines and cross-validation. *SIAM Journal on Scientific and Statistical Computing*, 9(3):531–542, 1988.

[16] S. Ripatti and J. Palmgren. Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics*, 56(4):1016–1022, 2000.

[17] L. A. Sleeper and D. P. Harrington. Regression splines in the cox model with application to covariate effects in liver disease. *Journal of the American Statistical Association*, 85 (412):941–949, 1990.

[18] L. Sun, X. Zhou, and S. Guo. Marginal regression models with time-varying coefficients for recurrent event data. *Statistics in Medicine*, 30(18):2265–2277, 2011.

[19] L. Tian, D. Zucker, and L. J. Wei. On the cox model with time-varying regression coefficients. *Journal of the American Statistical Association*, 100(469):172–183, 2005.

[20] L. J. Wei, D. Y. Lin, and L. Weissfeld. Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions. *Journal of the American Statistical Association*, 84(408):1065–1073, 1989.

[21] J. Yan and J. Huang. Model selection for cox models with time-varying coefficients. *Biometrics*, 68(2):419–428, 2012.

[22] Z. Yu, L. Liu, D. M. Bravata, L. S. Williams, and R. S. Tepper. A semiparametric recurrent events model with time-varying coefficients. *Statistics in Medicine*, 32(6): 1016–1026, 2013.

[23] D. M. Zucker and A. F. Karr. Nonparametric survival analysis with time-dependent covariate effects: A penalized partial likelihood approach. *The Annals of Statistics*, 18 (1):329–353, 1990.

Chapter 5 Multi-type Recurrent Event Model with Time-varying Coefficients

5.1 Introduction

It is common to observe different types of events occur simultaneously in a study. For example, in 100-Car NDS, there are three types of incidents related with driving risk from most to least severe: Crash, Near Crash, and Safety Critical Incident. In other field such as clinical study, it often involves patients who are exposed to different types of diseases that are possibly recurrent. Previous studies have been focused on single type recurrent event. For study includes more than one type, we model each type independently. This made interpretation clear and easy. But correlation across varied events is informative and potentially affect the legitimate of analysis. How to incorporate all types at the same time and borrow information from each other becomes an appealing topic.

In the context of multivariate recurrent event data, Cai and Schaubel [3] proposed a class of semi-parametric marginal means/rates models, with a general relative risk form on the censored event processes of interest. Wang et al. [12] and Zhu et al. [14] developed approaches with an arbitrary structure for both the relationship between recurrent events and the terminal event and the effect of covariates on the terminal event. In these studies, dependency between events is considered but not of interest. Frailty provides a convenient tool to incorporate dependence and heterogeneity. Sankaran and Anisha [10] extended shared frailty model to recurrent event data with multiple cases for gap time distributions. In Mazroui et al. [7], two types of recurrent events with dependent terminal event were jointly modeled.

Reference on evaluating time-dependent coefficients in multi-type recurrent event data is limited. Sun et al. [11] proposed marginal modeling approach and estimating equation based inference procedures. In the context of unobserved heterogeneity and dependence, there is no existing model to our best knowledge. Motivated by this, we propose a general frailty model which jointly analyze multiple types of recurrent events with both time-varying and time-fixed effects. Relationship among different types of events can be assessed from frailty terms. We use penalized B-spline to approximate time-varying coefficient. Variance components and smoothing parameters are estimated jointly by maximizing profile likelihood. We propose to estimate time-varying coefficient through multivariate frailty model by reparameterization. Besides easy implementation, it allows us to make systematic inference on all components.

The remainder of this Chapter is organized as follows. Section 5.2 introduces multi-type time-varying coefficient model and estimation procedure. In Section 5.3 we discuss inference of coefficients and variance components, followed by detailed hypothesis tests. We illustrate the method in Section 5.4 by applying it to 100-Car data. Simulation studies, including model performance in finite samples and power of statistical tests, are summarized in Section 5.5. Section 5.6 is devoted to concluding remarks and discussion.

5.2 Multi-type recurrent event model with time-varying coefficients

Assume there are K different types of recurrent events for each individual i, $i = 1, \cdots, m$. All types of events are observed till a censoring time T_i. In the present study, we assume a noninformative censor that is independent of any event process. Let $n_i^{(k)}$ be the total number of k-th type of events individual i encounters during the observed time period. Every time to event, $0 < t_{i1}^{(k)} < \cdots < t_{in_i^{(k)}}^{(k)} \leq T_i$, are recorded. This series of occurrence is treated as a counting process. An indicator function, $Y_i(t) = I(t \leq T_i)$, represents whether or not subject i is at risk at time t. Continuous assumption holds for all types of processes. This requires two events, no matter of the same type or different types, do not happen at the same time.

We model the intensity functions of recurrent events through a multivariate frailty model. It accomondates both time-dependent coefficients and time-constant coefficients as follows:

$$\lambda_i^{(k)}\left[t \mid \boldsymbol{\alpha}^{(k)}, \boldsymbol{\beta}^{(k)}(t), \boldsymbol{\gamma}\right] = \lambda_0^{(k)}(t) \exp\left[\boldsymbol{x}_i'(t)\boldsymbol{\alpha}^{(k)} + \boldsymbol{z}_i'(t)\boldsymbol{\beta}^{(k)}(t) + \boldsymbol{w}_i^{(k)'}\boldsymbol{\gamma}\right],\ k = 1, \cdots, K, \tag{5.1}$$

where $\lambda_0^{(k)}(t)$ is a nonparametric baseline function for k-th type of counting process. $\boldsymbol{x}_i(t)$ and $\boldsymbol{z}_i(t)$ are two vectors of time-dependent (or constant) covariates with dimension q and p respectively. $\boldsymbol{\alpha}^{(k)}$ is a vector of time-constant coefficient associated with $\boldsymbol{x}_i(t)$. Coefficients $\boldsymbol{\beta}^{(k)}(t)$ is a vector composed by p functions, $\boldsymbol{\beta}^{(k)}(t) = \left(\beta_1^{(k)}(t), \cdots, \beta_p^{(k)}(t)\right)'$. Note that the covariates could be different for various types of events. Without loss of generality, we assume they are the same. In the situation of different covariates, we can define a bigger set of covariate matrix and new vector of covariates. For example, if the covariates for the first type is $\boldsymbol{x}_i^1 = (x_{i1}, x_{i2})'$, for the second type is $\boldsymbol{x}_i^2 = (x_{i3}, x_{i4})'$, a comprehensive vector is defined as $\boldsymbol{x}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4})'$. Assume $\boldsymbol{\gamma}$ is a vector of frailty terms following multivariate normal distribution with mean $\mathbf{0}$ and covariance $\Sigma(\boldsymbol{\theta})$. Other choices, such as gamma distribution, can be derived similarly. The expression of $\boldsymbol{w}_i^{(k)}$ and distribution of $\boldsymbol{\gamma}$ are the key to correlation design. Taking two types of recurrent events with shared frailty

intensity function [4] as an example, model (5.1) has a specific form as below:

$$\begin{aligned}\lambda_i^{(1)}\Big[t \mid \boldsymbol{\alpha}^{(1)}, \boldsymbol{\beta}^{(1)}(t), \boldsymbol{\gamma}\Big] &= \lambda_0^{(1)}(t)\exp\Big[\boldsymbol{x}_i'(t)\boldsymbol{\alpha}^{(1)} + \boldsymbol{z}_i'(t)\boldsymbol{\beta}^{(1)}(t) + u_i\Big] \\ \lambda_i^{(2)}\Big[t \mid \boldsymbol{\alpha}^{(2)}, \boldsymbol{\beta}^{(2)}(t), \boldsymbol{\gamma}\Big] &= \lambda_0^{(2)}(t)\exp\Big[\boldsymbol{x}_i'(t)\boldsymbol{\alpha}^{(2)} + \boldsymbol{z}_i'(t)\boldsymbol{\beta}^{(2)}(t) + v_i\Big].\end{aligned} \tag{5.2}$$

u_i and v_i account for the unobserved heterogeneity, the inter-recurrence dependencies, and the dependency between different event types of subject i. Assuming all subjects are independent, each set of $(u_i, v_i)'$ is a realization of multivariate normal distribution with covariance: $\Sigma_i = \begin{pmatrix} \theta & \rho\sqrt{\theta\eta} \\ \rho\sqrt{\theta\eta} & \eta \end{pmatrix}$. Thus $\Sigma(\boldsymbol{\theta}) = \Sigma_i \otimes \boldsymbol{I}_m$. $\boldsymbol{I}_m$ is an identity matrix of length m and $\boldsymbol{\theta} = (\theta, \eta, \rho)'$. To be more specific, $\boldsymbol{\gamma} = (u_1, \cdots, u_m, v_1, \cdots, v_m)'$, $\boldsymbol{w}_1^{(1)} = (1, 0, \cdots, 0)'$, and $\boldsymbol{w}_1^{(2)} = (0, \cdots, 0, 1, 0, \cdots, 0)'$. Shared frailty model is a special case that has been widely studied [7, 13]. As a matter of fact, model (5.1) is very general and adapts various study designs since one can specify a flexible covariance structure.

One thing we like to note is that processes are jointly dependent through frailties only. Recurrent events are conditionally independent once frailty terms are given. Based on this, conditional loglikelihood is presented below:

$$\begin{aligned} l_c\Big[\lambda_0(\cdot), \boldsymbol{\alpha}, \boldsymbol{\beta}(\cdot) \mid \boldsymbol{\gamma}\Big] &= \sum_{k=1}^{K}\sum_{i=1}^{m} l_i^{(k)}\Big[\lambda_0^{(k)}(\cdot), \boldsymbol{\alpha}^{(k)}, \boldsymbol{\beta}(\cdot)^{(k)} \mid \boldsymbol{\gamma}\Big] \\ &= \sum_{k=1}^{K}\sum_{i=1}^{m}\Bigg\{\sum_{j=1}^{n_i^{(k)}}\Big[log\lambda_0^{(k)}(t_{ij}^{(k)}) + \boldsymbol{z}_i(t_{ij}^{(k)})'\boldsymbol{\beta}^{(k)}(t_{ij}^{(k)}) + \boldsymbol{x}_i'(t_{ij}^{(k)})\boldsymbol{\alpha}^{(k)} + \boldsymbol{w}_i^{'(k)}\boldsymbol{\gamma}\Big] \\ &\quad - \Lambda_0^{(k)}(T_i)\exp\Big[\boldsymbol{z}_i(\tau_i)'\boldsymbol{\beta}^{(k)}(\tau_i) + \boldsymbol{x}_i(\tau_i)'\boldsymbol{\alpha}^{(k)} + \boldsymbol{w}_i^{(k)'}\boldsymbol{\gamma}\Big]\Bigg\},\end{aligned} \tag{5.3}$$

where $\Lambda_0^{(k)}(t) = \int_0^t \lambda_0^{(k)}(u)du$. For convenience, we use $\boldsymbol{\lambda}_0(\cdot)$, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}(\cdot)$ to incorporate all the parameters associated with different types. It is not hard to find that (5.3) is the summation of all individual log likelihood by type. But marginal likelihood which is necessary for estimation of $\boldsymbol{\beta}(t)$ and $\boldsymbol{\alpha}$ can not be calculated from addition. We need to integrate out frailty terms:

$$\begin{aligned} l_m &= \log\int\prod_{k=1}^{K}\prod_{i=1}^{m} L_i^{(k)}\Big[\lambda_0^{(k)}(\cdot), \boldsymbol{\alpha}^{(k)}, \boldsymbol{\beta}^{(k)}(\cdot) \mid \boldsymbol{\gamma}\Big] \times f(\boldsymbol{\gamma})d\boldsymbol{\gamma} \\ &= \log\int\frac{1}{(2\pi)^{\frac{m}{2}}|\Sigma(\boldsymbol{\theta})|^{\frac{1}{2}}}\exp\left\{\sum_{k=1}^{K}\sum_{i=1}^{m} l_i^{(k)}\Big[\lambda_0^{(k)}(\cdot), \boldsymbol{\alpha}^{(k)}, \boldsymbol{\beta}^{(k)}(\cdot) \mid \boldsymbol{\gamma}\Big] - \frac{\boldsymbol{\gamma}'\Sigma(\boldsymbol{\theta})^{-1}\boldsymbol{\gamma}}{2}\right\}d\boldsymbol{\gamma}.\end{aligned} \tag{5.4}$$

It involves formulation of nonparametric time-varying coefficients and integration over frailty terms. In the next two subsections, we shall first discuss how to construct penalized B-spline estimates for $\boldsymbol{\beta}(t)$ and introduce approximation for marginal likelihood.

5.2.1 Penalized B-Spline estimation

In this subsection, we illustrate how to use penalized B-spline to approximate time-varying coefficients. Here we continue to use spline basis function that was introduced in Section 4.2.1. The expression of $\beta_l^{(k)}(t)$ is given below:

$$\beta_l^{(k)}(t) = B(t)^{(k)'}\boldsymbol{\eta}_l^{(k)}. \tag{5.5}$$

$B(t)^{(k)}$ collects spline basis functions of k-th event. If time-varying coefficients across types span on similar scale, it is easier to have a universal spline basis for all. Replacing the scalar covariate $\beta_l^{(k)}(t)$ by d-dimensional vector, model (5.1) turns into:

$$\lambda_i^{(k)}\Big[t \mid \boldsymbol{\alpha}^{(k)}, \boldsymbol{\eta}^{(k)}, \boldsymbol{\gamma}\Big] = \lambda_0^{(k)}(t)\exp\Big[\boldsymbol{z}_i(t)' \otimes B(t)^{(k)'}\boldsymbol{\eta}^{(k)} + \boldsymbol{x}_i'(t)\boldsymbol{\alpha}^{(k)} + \boldsymbol{w}_i^{(k)'}\boldsymbol{\gamma}\Big], \tag{5.6}$$

where $\boldsymbol{\eta}^{(k)} = (\boldsymbol{\eta}_1^{(k)'}, \cdots, \boldsymbol{\eta}_p^{(k)'})'$. It does not include time-varying coefficient anymore. Once basis functions are defined, $\boldsymbol{\eta}$ takes place of $\boldsymbol{\beta}(\cdot)$ as unknown parameter and can be analyzed by standard procedure. Among literatures about B-splines, a popular technique is to introduce a penalty term to control smoothness of the estimated functions. Penalized marginal loglikelihood is of the following form:

$$l_{pm} = l_m\Big[\boldsymbol{\lambda}_0(t), \boldsymbol{\alpha}, \boldsymbol{\eta}, \boldsymbol{\theta}\Big] - \frac{1}{2}\sum_{k=1}^{K}\sum_{l=1}^{p}\lambda_l^{(k)}\boldsymbol{\eta}_l^{(k)'}P^{(k)}\boldsymbol{\eta}_l^{(k)}. \tag{5.7}$$

$\lambda_l^{(k)}$ is a smoothing parameter. Smaller smoothing parameter places less constraint and make estimated coefficient function more flexible. $P^{(k)}$ is a nonnegative definite smoothing matrix. A common choice of $P^{(k)}$ is second derivative of basis functions $P^{(k)} = \int \boldsymbol{B}^{(k)(2)}(t)\boldsymbol{B}^{(k)(2)}(t)'dt$ [8]. For convenience of future reference, we denote $\boldsymbol{\lambda}^{(k)} = (\lambda_1^{(k)}, \cdots, \lambda_p^{(k)})'$.

5.2.2 Double penalized partial likelihood

With normally distributed frailties, there is no closed form for (5.7). We borrowed the idea of Laplace approximation from Breslow and Clayton [2], Ripatti and Palmgren [9], and Yu

et al. [13]. The penalized marginal loglikelihood is approximated as:

$$l_{pm} \approx -\frac{1}{2}log|\Sigma(\boldsymbol{\theta})| - log|K''(\tilde{\boldsymbol{\gamma}})| + l_c\Big[\lambda_0(\cdot), \boldsymbol{\alpha}, \boldsymbol{\eta} \mid \tilde{\boldsymbol{\gamma}}\Big] - \frac{1}{2}\tilde{\boldsymbol{\gamma}}'\Sigma(\boldsymbol{\theta})^{-1}\tilde{\boldsymbol{\gamma}} - \frac{1}{2}\sum_{k=1}^{K}\sum_{l=1}^{p}\lambda_l^{(k)}. \quad (5.8)$$

$K(\boldsymbol{\gamma}) = -l_c\Big[\lambda_0(\cdot), \boldsymbol{\alpha}, \boldsymbol{\eta} \mid \boldsymbol{\gamma}\Big] + \frac{1}{2}\boldsymbol{\gamma}'\Sigma(\boldsymbol{\theta})^{-1}\boldsymbol{\gamma}$. $\tilde{\boldsymbol{\gamma}}$ is the solution to the first derivative $K'(\boldsymbol{\gamma}) = \mathbf{0}$ and $K''(\cdot)$ is the second derivative of $K(\cdot)$ with respect to $\boldsymbol{\gamma}$.

If $\boldsymbol{\theta}$ is fixed, the first term in (5.8) is constant with respect to $\boldsymbol{\alpha}, \boldsymbol{\eta}$. Ripatti and Palmgren [9] showed through simulation study that ignoring the second term in (5.8) resulted some information loss, but not too much to influence the estimation precision. It did simplify estimation procedure. In addition, if $\boldsymbol{\gamma}$ and $\boldsymbol{\eta}$ are considered as fixed effects parameters, (5.8) is a double penalized log likelihood. The first term put penalty on extreme values of $\boldsymbol{\gamma}$ and the second is for penalizing smoothness of the time-varying coefficients functions. Maximizing l_{pm} is equivalent to maximizing:

$$l_c\Big[\lambda_0(\cdot), \boldsymbol{\alpha}, \boldsymbol{\eta} \mid \tilde{\boldsymbol{\gamma}}\Big] - \frac{1}{2}\tilde{\boldsymbol{\gamma}}'\Sigma(\boldsymbol{\theta})^{-1}\tilde{\boldsymbol{\gamma}} - \frac{1}{2}\sum_{k=1}^{K}\sum_{l=1}^{p}\lambda_l^{(k)}\boldsymbol{\eta}_l^{(k)'}P^{(k)}\boldsymbol{\eta}_l^{(k)}. \quad (5.9)$$

Conditional log likelihood can be maximized by partial likelihood (PPL). Finally, we derive parameters by maximizing double penalized partial likelihood (DPPL) as below:

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{j=1}^{n_i^{(k)}}\Big\{\boldsymbol{z}_i(t_{ij}^{(k)})'\boldsymbol{\beta}^{(k)}(t_{ij}^{(k)}) + \boldsymbol{x}_i(t_{ij}^{(k)})'\boldsymbol{\alpha}^{(k)} + \boldsymbol{w}_i^{(k)'}\boldsymbol{\gamma} \\
&- log\sum_{r\in R_{ij}^{(k)}}\exp\Big[\boldsymbol{z}_r(t_{ij}^{(k)})'\boldsymbol{\beta}^{(k)}(t_{ij}^{(k)}) + \boldsymbol{x}_r(t_{ij}^{(k)})'\boldsymbol{\alpha}^{(k)} + \boldsymbol{w}_r^{(k)'}\boldsymbol{\gamma}\Big]\Big\} \\
&- \frac{1}{2}\boldsymbol{\gamma}'\Sigma(\boldsymbol{\theta})^{-1}\boldsymbol{\gamma} - \frac{1}{2}\sum_{k=1}^{K}\sum_{l=1}^{p}\lambda_l^{(k)}\boldsymbol{\eta}_l^{(k)'}P^{(k)}\boldsymbol{\eta}_l^{(k)}.
\end{aligned} \quad (5.10)$$

Note that smoothing parameters λ_l, $l = 1, \cdots, p$ need to be specified beforehand. Detail of smoothing parameter estimation is discussed in Section 5.3.

5.2.3 The frailty model representation

Lin and Zhang [6] proposed to fit generalized additive mixed model by generalized linear mixed model. The key idea is to estimate spline coefficients as a combination of fixed and

random effects. Using similar idea, we propose to rewrite $\boldsymbol{\eta}_l^{(k)}$ in (5.6) as

$$\boldsymbol{\eta}_l^{(k)} = \mathbf{1}\beta_{0l}^{(k)} + \boldsymbol{b}^{(k)}\beta_{1l}^{(k)} + A^{(k)}\boldsymbol{a}_l^{(k)}, \tag{5.11}$$

where $\mathbf{1}$ is a vector of 1 with length d; $A^{(k)} = L^{(k)}(L^{(k)'}L^{(k)})^{-1}$. $L^{(k)}$ is a $d \times (d-2)$ full rank matrix satisfying $P^{(k)} = L^{(k)}L^{(k)'}$. $\boldsymbol{b}^{(k)}$ is a $d \times 1$ vector that satisfies $\boldsymbol{b}^{(k)'}\mathbf{1} = 0$ and $\boldsymbol{b}^{(k)'}A^{(k)} = \mathbf{0}$. In another word, $\boldsymbol{b}^{(k)}$ is the orthogonal complement of the space composed by $\mathbf{1}$ and $A^{(k)}$. Using the identity $\boldsymbol{\eta}_l^{(k)'}P^{(k)}\boldsymbol{\eta}_l^{(k)} = \boldsymbol{a}_l^{(k)'}\boldsymbol{a}_l^{(k)}$, Equation (5.10) turns into:

$$\begin{aligned}&\sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{j=1}^{n_i^{(k)}}\Big\{\boldsymbol{x}_i^*(t_{ij}^{(k)})'\boldsymbol{\alpha}^{*(k)} + \boldsymbol{w}_i^{*(k)'}\boldsymbol{\gamma}^* \\ &- log\sum_{r\in R_{ij}^{(k)}}\exp\Big[\boldsymbol{x}_r^*(t_{ij}^{(k)})'\boldsymbol{\alpha}^{*(k)} + \boldsymbol{w}_r^{*(k)'}\boldsymbol{\gamma}^*\Big]\Big\} - \frac{1}{2}\boldsymbol{\gamma}^{*'}\Sigma^{*-1}(\boldsymbol{\theta})\boldsymbol{\gamma}^*.\end{aligned} \tag{5.12}$$

where $\boldsymbol{x}_i^*(t)' = \Big(\boldsymbol{x}_i(t)',\, \boldsymbol{z}_i(t)' \otimes B(t)'\boldsymbol{I} \otimes \mathbf{1},\, \boldsymbol{z}_i' \otimes B(t)'\boldsymbol{I} \otimes \boldsymbol{b}^{(k)}\Big)$; $\boldsymbol{w}_i^{*(k)'} = \Big(\boldsymbol{z}_i(t)' \otimes B(t)'\boldsymbol{I} \otimes A^{(k)},\, \boldsymbol{w}_i^{(k)'}\Big)$. $\boldsymbol{\alpha}^{*(k)} = \Big(\boldsymbol{\alpha}^{(k)'},\, \boldsymbol{\beta}_0^{(k)'},\, \boldsymbol{\beta}_1^{(k)'}\Big)'$ is a vector of time-fixed coefficients, where $\boldsymbol{\beta}_0^{(k)} = (\beta_{01}^{(k)}, \cdots, \beta_{0p}^{(k)})'$ and $\boldsymbol{\beta}_1^{(k)} = (\beta_{11}^{(k)}, \cdots, \beta_{1p}^{(k)})'$. $\boldsymbol{\gamma}^* = \Big(\boldsymbol{a}',\, \boldsymbol{\gamma}'\Big)'$ is a vector of random effects. $\boldsymbol{a}' = \Big(\boldsymbol{a}_1^{(1)'}, \cdots, \boldsymbol{a}_p^{(K)'}\Big)$ follows a multivariate normal distribution with center $\mathbf{0}$ and covariance matrix $\Lambda = Diag\Big(\frac{1}{\lambda_1^{(1)}}\boldsymbol{I}_{d-2}, \cdots, \frac{1}{\lambda_p^{(K)'}}\boldsymbol{I}_{d-2}\Big)$.

Eqatuion (5.12) suggests that the maximum DPPL estimator can be obtained by fitting the following frailty model:

$$\lambda_i^{(k)}\Big[t \mid \boldsymbol{\alpha}^{*(k)}, \boldsymbol{\gamma}^*\Big] = \lambda_0^{(k)}(t)\exp\Big[\boldsymbol{x}_i^{*'}(t)\boldsymbol{\alpha}^{*(k)} + \boldsymbol{w}_i^{*'}\boldsymbol{\gamma}^*\Big]. \tag{5.13}$$

Each individual time-varying coefficient $\widehat{\beta}_l(t)$ can be calculated by:

$$B^{(k)}(t)'\mathbf{1}\widehat{\beta}_{0l}^{(k)} + B^{(k)}(t)'\boldsymbol{b}^{(k)}\widehat{\beta}_{1l}^{(k)} + B^{(k)}(t)'A^{(k)}\hat{\boldsymbol{a}}_l{}^{(k)}. \tag{5.14}$$

It is a linear combination of the Ripatti and Palmgren [9] maximum penalized partial likelihood estimator of the fixed effect and the random effects.

5.2.4 Maximum DPPL estimation

The estimated parameters are obtained by Newton-Raphson algorithm. Taking first derivative of equation (5.12) with respect to $(\boldsymbol{\alpha}^{*'}, \boldsymbol{\gamma}^{*'})'$ yields score functions as below:

$$\begin{aligned}
\frac{\partial DPPL}{\partial \boldsymbol{\alpha}^{*'}} &= \sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{j=1}^{n_i^{(k)}} \boldsymbol{x}_i^*(t_{ij}^{(k)}) - \frac{\exp\left[\boldsymbol{x}_i^*(t_{ij}^{(k)})'\boldsymbol{\alpha}^{*(k)} + \boldsymbol{w}_i^{*(k)'}\boldsymbol{\gamma}^*\right]\boldsymbol{x}_i^*(t_{ij}^{(k)})}{\sum_{r\in R_{ij}^{(k)}} \exp\left[\boldsymbol{x}_r^*(t_{ij}^{(k)})'\boldsymbol{\alpha}^{*(k)} + \boldsymbol{w}_r^{*(k)'}\boldsymbol{\gamma}^*\right]} \\
\frac{\partial DPPL}{\partial \boldsymbol{\gamma}^{*'}} &= \sum_{k=1}^{K}\sum_{i=1}^{m}\sum_{j=1}^{n_i^{(k)}} \boldsymbol{w}_i^* - \frac{\exp\left[\boldsymbol{x}_i^*(t_{ij}^{(k)})'\boldsymbol{\alpha}^{*(k)} + \boldsymbol{w}_i^{*(k)'}\boldsymbol{\gamma}^*\right]\boldsymbol{w}_i^*}{\sum_{r\in R_{ij}^{(k)}} \exp\left[\boldsymbol{x}_r^*(t_{ij}^{(k)})'\boldsymbol{\alpha}^{*(k)} + \boldsymbol{w}_r^{*(k)'}\boldsymbol{\gamma}^*\right]} - \Sigma^*(\boldsymbol{\lambda},\boldsymbol{\theta})^{-1}\boldsymbol{\gamma}^*,
\end{aligned} \tag{5.15}$$

where $\Sigma^*(\boldsymbol{\lambda},\boldsymbol{\theta}) = Diag(\Lambda,\Sigma)$. Note that the first score function is combination of estimating equations of fixed effects across types. Fitting K frailty models is a easy approach. However the second score function is more complicated because of correlation matrix. Given all parameters, baseline intensity function estimate is:

$$\widehat{\lambda}_0^{(k)}(t) = \frac{\sum_{i=1}^m Y_i(t) dN_i^{(k)}(t)}{\sum_{i=1}^m Y_i(t)\exp\left[\boldsymbol{x}_r^*(t_{ij}^{(k)})'\hat{\boldsymbol{\alpha}}^{*(k)} + \boldsymbol{w}_r^{*(k)'}\widehat{\boldsymbol{\gamma}}^*\right]}, \tag{5.16}$$

where $N_i^{(k)}(t)$ is a process counting the number of type k event happen to subject i before time t.

5.3 Statistical inference

5.3.1 Inference on smoothing parameter and variance component

Statistical inference of nonparametric functions $\boldsymbol{\beta}_l^{(k)}(t)$ relies on the smoothing parameters and variance parameter $\boldsymbol{\theta}$. In the previous section, we estimated time-varying coefficients as a linear combination of fixed effect and random effects. Smoothing parameters were treated as extra variance components. Thus inference on smoothing parameters can be conducted similarly to that on variance component in multivariate frailty model. Plugging maximum DPPL estimators into (5.8) results profile likelihood function of $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$:

$$l(\boldsymbol{\theta},\boldsymbol{\lambda}) \approx -\frac{1}{2}log|\Lambda| - \frac{1}{2}log|\Sigma(\boldsymbol{\theta})| - \frac{1}{2}\log|K''(\hat{\boldsymbol{\gamma}}^*)| - \frac{1}{2}\hat{\boldsymbol{\gamma}}\Sigma(\boldsymbol{\theta})^{-1}\hat{\boldsymbol{\gamma}} - \frac{1}{2}\hat{\boldsymbol{a}}\Lambda^{-1}\hat{\boldsymbol{a}}, \tag{5.17}$$

where K was derived in 5.2.2. Here we propose to use $K''_{DPPL}(\hat{\boldsymbol{\gamma}^*}) = (\partial^2 DPPL)/(\partial\boldsymbol{\gamma}^*\partial\boldsymbol{\gamma}^{*'})$ instead of K''. Estimating equations of $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ is derived by taking the first derivative of (5.17). The corresponding Fisher information matrix can be derived by differentiating (5.17) twice and taking the expectation with respect to $\boldsymbol{\gamma}^{*'}$, similar to equation (8) in Ripatti and Palmgren [9].

Estimating equations of $\boldsymbol{\theta}$ is a series of nonlinear functions:

$$\frac{1}{2}\Big[tr(\Sigma^{*-1}\frac{\partial\Sigma^*}{\partial\theta_i} + tr\Big(K''^{-1}_{DPPL}\frac{\partial\Sigma^{*-1}}{\partial\theta_i}\Big) + \hat{\boldsymbol{\gamma}^{*'}}\frac{\partial\Sigma^{*-1}}{\partial\theta_i}\hat{\boldsymbol{\gamma}^*}\Big] = 0, \tag{5.18}$$

where i denotes each parameter in $\boldsymbol{\theta}$. Solving nonlinear equations in (5.18) results estimations of the parameters. In the present study, we focus on the case of two types of recurrent event and the intensity function of each type is a shared frailty model (See Section 5.2 for detailed form of $\boldsymbol{w}_i^{(k)}$, $\boldsymbol{\gamma}$, and $\Sigma(\boldsymbol{\theta})$). With some derivation, simplified estimating equations are:

$$\begin{aligned}
&(1-\rho^2)m - \frac{tr(A)}{\theta} + \rho\frac{tr(B+C)}{2\sqrt{\theta\eta}} - \frac{\boldsymbol{u}'\boldsymbol{u}}{\theta} + \rho\frac{\boldsymbol{u}'\boldsymbol{v}}{\sqrt{\theta\eta}} = 0\\
&(1-\rho^2)m - \frac{tr(D)}{\eta} + \rho\frac{tr(B+C)}{2\sqrt{\theta\eta}} - \frac{\boldsymbol{v}'\boldsymbol{v}}{\eta} + \rho\frac{\boldsymbol{u}'\boldsymbol{v}}{\sqrt{\theta\eta}} = 0\\
&(1-\rho^2)m - \frac{tr(A)}{\theta} - \frac{tr(D)}{\eta} + \frac{(1+\rho^2)tr(B+C)}{2\rho\sqrt{\theta\eta}} - \frac{\boldsymbol{u}'\boldsymbol{u}}{\theta} - \frac{\boldsymbol{v}'\boldsymbol{v}}{\eta} + \frac{(1+\rho^2)\boldsymbol{u}'\boldsymbol{v}}{\rho\sqrt{\theta\eta}} = 0
\end{aligned} \tag{5.19}$$

where A is the upper left block of submatrix of K''^{-1}_{DPPL} corresponding to $\boldsymbol{u}$; B is the upper right block; D is the lower right block corresponding to $\boldsymbol{v}$. In terms of estimating $\boldsymbol{\lambda}$, random effects associated with time-varying effects are independent of $\Sigma(\boldsymbol{\theta})$, solution to its score function at 0 is more straightforwards and expressed as:

$$\widehat{\lambda}_l^{(k)} = \frac{d-2}{\hat{\boldsymbol{a}}_l{}^{(k)'}\hat{\boldsymbol{a}}_l{}^{(k)} + tr\Big(K''^{-1}_{DPPL}(\hat{\boldsymbol{a}}_l{}^{(k)})\Big)}, \tag{5.20}$$

where $K''^{-1}_{DPPL}(\hat{\boldsymbol{a}}_l{}^{(k)})$ is the submatrix of K''_{DPPL} corresponding to $\boldsymbol{a}_l^{(k)}$. Variance is estimated by the inverse of observed information matrix:

$$\frac{1}{2}\Big[tr(\Sigma^{*-1}\frac{\partial\Sigma^*}{\partial\theta}\Sigma^{*-1}\frac{\partial\Sigma^*}{\partial\theta} + \Sigma^{*-1}\frac{\partial^2\Sigma^*}{\partial^2\theta}) + tr(K''^{-1}_{DPPL}\frac{\partial\Sigma^{*-1}}{\partial\theta}K''^{-1}_{DPPL}\frac{\partial\Sigma^{*-1}}{\partial\theta}) - tr(K''^{-1}_{DPPL}\frac{\partial^2\Sigma^{*-1}}{\partial^2\theta})\Big] \tag{5.21}$$

5.3.2 Computation

The maximization of DPPL is carried through two steps. First, an initial value for $\boldsymbol{\theta}$ is assigned. Solve (5.15) using the Newton-Raphson algorithm. Second, update variance parameters and smoothing parameters by (5.18) and (5.20) under current estimated $\widehat{\boldsymbol{\alpha}}^*$ and $\widehat{\boldsymbol{\gamma}}^*$. The two steps are iterated until convergence. Once convergence is achieved, the cumulative baseline hazard can be estimated using (5.16), which is the same as that of a Andersen and Gill [1] model with an offset of estimated random effects.

5.3.3 Statistical Test

The objective of this subsection is to elaborate tests that are commonly of interest. Test procedures are similar to that of 4.3.5. In addition to the test for each time-varying coefficient, equivalence of time-varying coefficients across various event types is examined. A similar test is applied on checking equivalence of time-fixed effects. General linear hypothesis test starts from:

$$H_0 : C(\boldsymbol{\alpha}^{*'}, \boldsymbol{\gamma}^{*'})' = \mathbf{0},\ H_a : C(\boldsymbol{\alpha}^{*'}, \boldsymbol{\gamma}^{*'})' \neq \mathbf{0}, \tag{5.22}$$

where C is a matrix with full row rank R. We continue to use Wald test statistic suggested by Gray [5]:

$$\Big[C(\boldsymbol{\alpha}^{*'}, \boldsymbol{\gamma}^{*'})'\Big]'(CI_p^{-1}C')^{-1}\Big[C(\boldsymbol{\alpha}^{*'}, \boldsymbol{\gamma}^{*'})'\Big], \tag{5.23}$$

And the generalized degrees of freedom of the test statistic is:

$$df = trace\Big[(CI_p^{-1}C')^{-1}CVC\Big], \tag{5.24}$$

where $V = -H_p^{-1}HH_P^{-1}$. Large test statistic which exceeds critical value signify evidence against null hypothesis.

If the objective is testing no effect or a constant effect on a single covariate on type k, C matrix is set up similarly to that of 4.3.5 Equivalence of time-varying coefficients across types can be set up similarly. For instance, consider a study with two types of event ($k = 1, 2$). To examine whether the effect of l-th covariate is the same, we can format the null hypothesis as $H_0 : \beta_l^{(1)}(t) = \beta_l^{(2)}(t)$. In this case C is a $d \times 2d$ matrix of the following form:

$$\begin{pmatrix} 1 & \cdots & 0 & -1 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & 0 & 0 & \cdots & -1 \end{pmatrix}. \tag{5.25}$$

We want to point out that for all the above four tests, it is essentially the same to conduct

a test on $(\beta_{0l}, \beta_{0l}, \boldsymbol{a}_l)$. The corresponding C matrix shall be multiplied by $\begin{pmatrix} \mathbf{1} & \boldsymbol{b}^{(k)} & A^{(k)} \end{pmatrix}$.

5.4 Application to 100-Car Naturalistic Driving Study

In this section, we apply the proposed model to the 100-Car Naturalistic Driving Study (NDS). The objective is to evaluate the pattern of crash influence on driving risk measured by NC and SCI simultaneously over time. Previous research has shown driving risk is reduced after crash, with amount of decrease varies over time and across gender. But it didn't take correlation between SCI and NC into consideration. Exploratory data analysis shows high NC rate coincides with high SCI rate. A numerical description of the association is desired. We are also interested in equality of crash influence on SCI and NC.

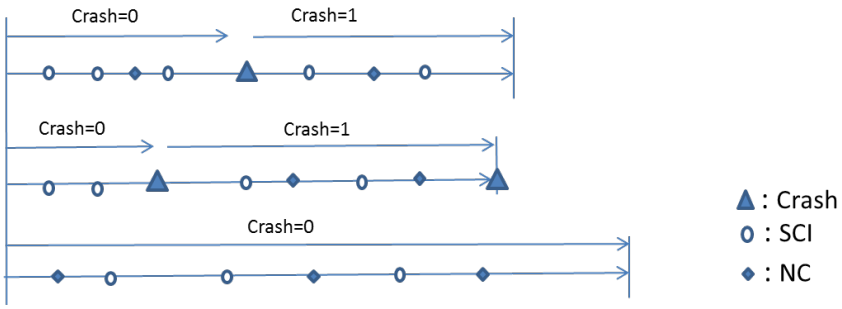


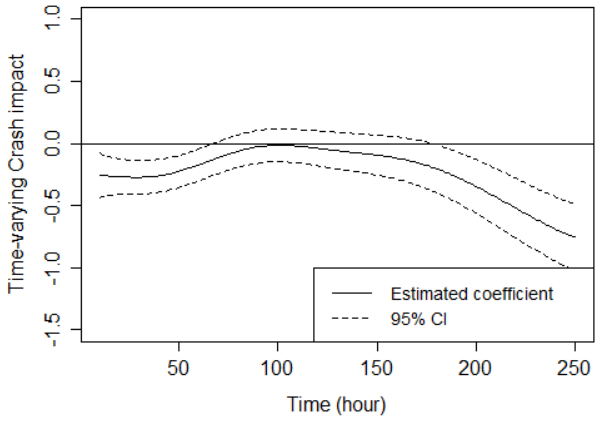
Figure 5.1: Data collection structure.

Data collection structure shown in Figure 5.1 incorporate crash-related incidents of all types. Each horizontal line represents record of one driver. SCIs and NCs aretreated as two types of processes that may be correlated. Each driving period was divided into two phases based on its relationship with crashes: before the first crash (coded as 0) and between the first and second crash (coded as 1). Since this was a one-year study, there were only 12 drivers had two crashes. After careful review, only 4 of them had NCs after second crash. Thus we considered to evaluate the first crash effect only. Observations after second crashes are treated as censored. Driving period was taken into account as a covariate, working as an external and independent factor on SCI/NC intensity. To account for potential confounding and interacting effects, gender and age of the driver when first enrolled in the study were

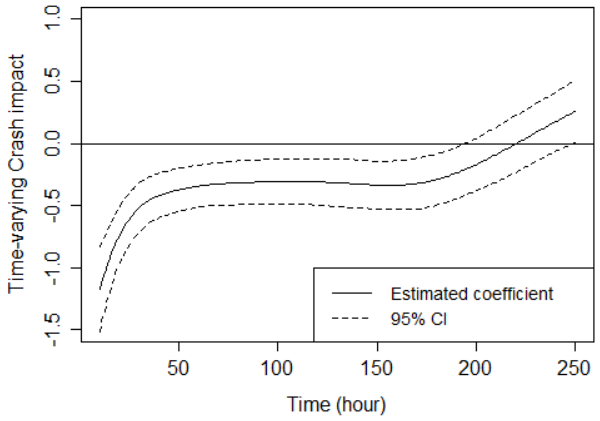
also evaluated as time-fixed covariates. For NC, SCI, and Crash, final model is given below:

$$\begin{aligned}
\lambda_i^{(1)}(t) =& \lambda_0^{(1)} \exp\left[\beta_1^{(1)} G_i + \beta_2^{(1)} Age_i + \beta_3^{(1)}(t-c_i) I_i(t) + \beta_4^{(1)}(t-c_i) G_i I_i(t) + u_i\right] \\
\lambda_i^{(2)}(t) =& \lambda_0^{(2)} \exp\left[\beta_1^{(2)} G_i + \beta_2^{(2)} Age_i + \beta_3^{(2)}(t-c_i) I_i(t) + \beta_4^{(2)}(t-c_i) G_i I_i(t) + v_i\right] \\
\lambda_i^{(3)}(t) =& \lambda_0^{(3)} \exp\left[\beta_1^{(2)} G_i + \beta_2^{(2)} Age_i\right].
\end{aligned} \tag{5.26}$$

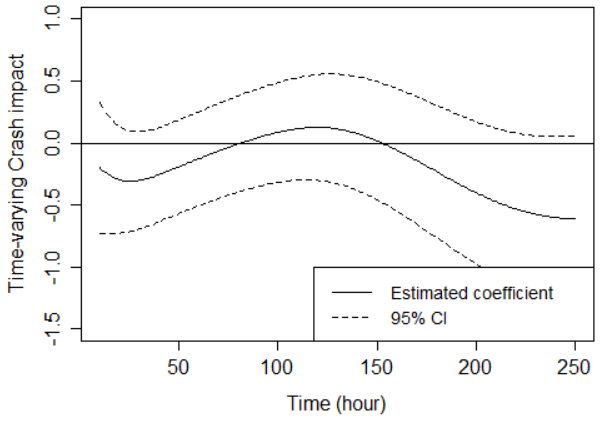
$\lambda_i^{(k)}(t)$, $k = 1, 2, 3$ denote intensity functions of NC, SCI, and crash respectively. $G_i = 1$ for male driver and 0 for female driver. $\beta_3^{(k)}(t - c_i)$ is crash influence on female driver; $\beta_3^{(k)}(t - c_i) + \beta_4^{(k)}(t - c_i)$ is crash influence on male driver. Influence of crash on its own intensity is not included since it is a rare event. Neither is there frailty terms associated with crash intensity. Number of participants experienced crash is 33. Introducing an extra set of frailty terms fro crash concerns the author regarding to estimation precision.



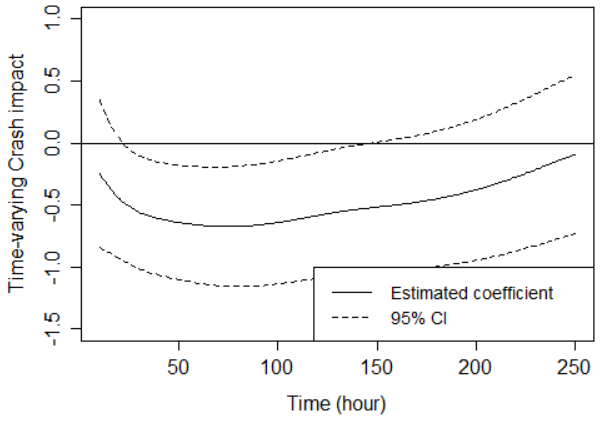
(a) Influence on SCI for female drivers



(b) Influence on SCI for male drivers



(a) Influence on NC for female drivers



(b) Influence on NC for male drivers

Figure 5.2: Crash influence on SCI and NC: $\widehat{\beta}(t)$, solid; 95% poitwise confidence interval, dashed.

Estimated crash influences on male drivers and female drivers are presented in Figure 5.2. As we can see from top two plots, SCI intensity drops after crash for male drivers and gradually increase after 200 hours. On contrary, SCI intensity increases after crash and then decrease after 200 hours. Both effects are consistent with results from constant coefficient model. In the bottom two figures, male driver experience lower driving risk after crash till 120 hours later. No significant change after that. For female drivers, there was no significant decrease in NC intensity. In summary, driving risk of male drivers tend to decrease after crash and then increase. But we did not find similar influence pattern for female drivers. Table 5.1 presents time-fixed effects estimation. On average, male driver is associated with

lower SCI/NC intensity, around 0.8 times of the rate of female drivers. As age increases, driving risk slightly decreases, with intensity rate ratio 0.99. Correlation ρ is 0.64, indicating a relatively large positive association between SCI and NC.

Table 5.1: Parametric coefficient estimates of 100-Car NDS

	Risk factor	Estimates	SE	Intensity rate ratio	p-value
NC	Gender	-0.23	0.102	0.79	0.025
	Age	-0.01	0.003	0.99	$< .001$
	θ_1	0.68	0.09		
SCI	Gender	-0.32	0.031	0.73	$< .001$
	Age	-0.02	0.001	0.98	$< .001$
	θ_2	1.26	0.14		
	ρ	0.64	0.07		

5.5 Simulation study

In this section, we conduct a simulation study to evaluate the performance of the proposed method. We simulate intensity models in multiple settings. In each setting, we examine both estimation precision and power of test for time-varying coefficients.

5.5.1 Simulation setup

The simulation procedure is designed to mimic the 100-Car Data. It consists generating censor time, crash time and events time for each subjects as described below:

1. Censor time is set as 4.

2. For each subject, generate a crash time C_i based on the following intensity function:

$$\lambda_i(t) = \frac{1}{2} \exp\left[\beta_1^{(3)} x_{1i} + \beta_2^{(3)} x_{2i}\right],\ t \leq 4. \tag{5.27}$$

   $\frac{1}{2}$ is selected based on the relationship between crash time and study period from 100-Car data. This intensity function assign distinct rates according to covariates x_1 and x_2. If one subject has C_i greater than 4, C_i shall be censored. We use a time-varying indicator function, $I_i(t) = (t > C_i)$, to denote the relationship between time t and crash time.

3. Recurrent events time are generated from the following intensity functions:

$$\begin{aligned}\lambda_i^{(1)}(t) &= c^{(1)} t^{k^{(1)}-1} \exp\left[\beta_1^{(1)} x_{1i} + \beta_2^{(1)} x_{2i} + \beta_3^{(1)}(t - C_i) I_i(t) + u_i\right] \\ \lambda_r^{(2)}(t) &= c^{(2)} t^{k^{(2)}-1} \exp\left[\beta_1^{(2)} x_{1i} + \beta_2^{(2)} x_{2i} + \beta_3^{(2)}(t - C_i) I_i(t) + v_i\right]\end{aligned} \tag{5.28}$$

where:

(a) Baseline intensity functions follow Weibull distribution, where two parameters c and r play a critical role. They may vary from stratum to stratum, as denoted by c_r and k_r. $k_r > 1$ indicates increasing rate over time, $k_r = 1$ refers to constant rate, and $k_r < 1$ denotes decreasing rate.

(b) C_i is the crash time generated in the previous step.

(c) x_{1i} is a binary covariate, with 50% probability to be 0/1.

(d) x_{2i} is a continuous covariate following $\mathcal{U}(-1, 1)$ distribution.

(e) I_i is considered a time-varying binary covariate. It takes a value of 1 when t is larger than the crash time.

(f) $\beta_1^{(k)}$ and $\beta_2^{(k)}$ are time-fixed effects.

(g) $\beta_3^{(k)}(t - c_i)$ is time-varying effect. Since c_i's are different across subjects, we focus evaluating those effects as functions of time after crash.

(h) u_i and v_i are two dependent frailty terms. They follow multivariate normal distribution with center $\mathbf{0}$ and covarance matrix $\begin{pmatrix} \theta & \rho\sqrt{\theta\eta} \\ \rho\sqrt{\theta\eta} & \eta \end{pmatrix}$.

In order to cover a certain range of parameter space, 6 different settings of baseline parameters and covariates' effects are explored. Within each setting, 200 realizations are generated and three models are implemented: a multi-type recurrent event model with time-varying coefficient, an independent stratified time-varying recurrent event model on each type, and a non-stratified independent model. Because of space limits, we only provide results from selected settings.

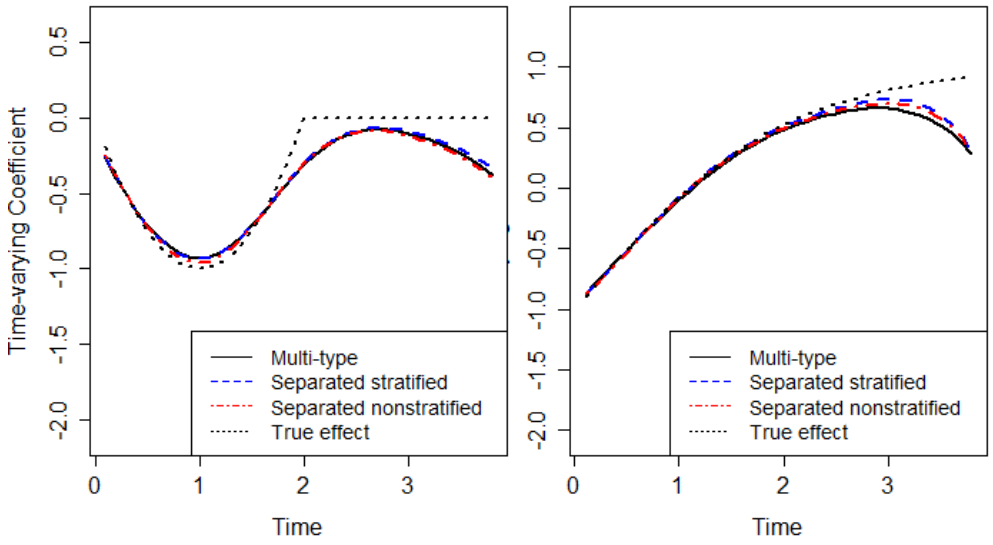
5.5.2 Simulation result

In the first setting we generate time-varying coefficient from:

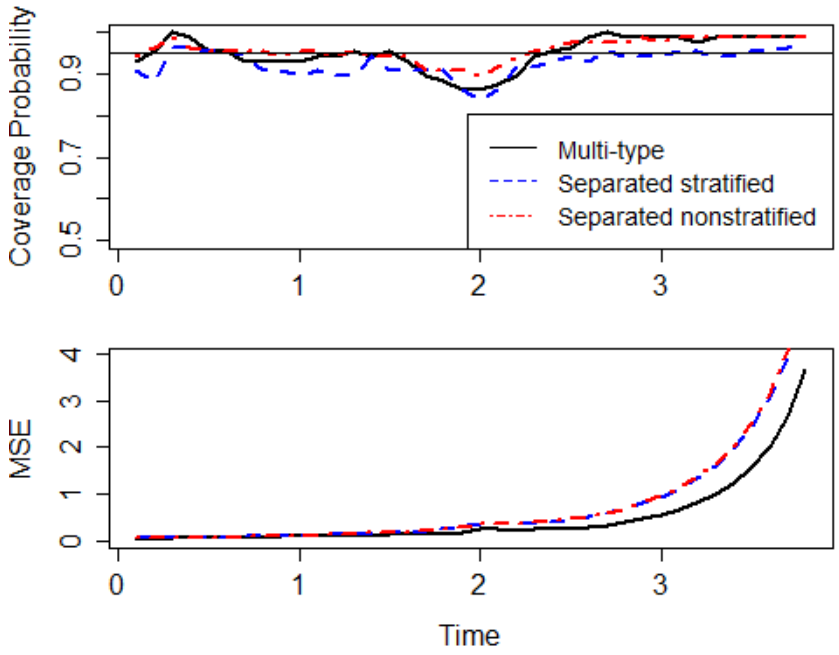
$$\begin{aligned}\beta_3^2(t) =& (t^2 - t) \times I(t < 2) + 0 \times I(t >= 2) \\ \beta_3^1(t) =& \frac{4}{1 + \exp(-t)} - 3,\end{aligned} \tag{5.29}$$

as shown in left panel of Figure 5.3. Scale and shape parameters of baseline functions are $c^1 = 1.5$, $c^2 = 3$, $k^1 = k^2 = 1$. It describes a constant type-I rate of 1.5 and type-II 3 events per time unit. Finite sample size is 50 subjects which yields about 15 events per subject in total. Correlation matrix between u_i and v_i is $\begin{pmatrix} 0.5 & 0.3 \\ 0.3 & 0.25 \end{pmatrix}$.

Figure 5.3 shows simulation results over 200 replications. Dashed black dots refer to true functions. The black solid curves are average estimating effects from multi-type recurrent event model. Blue and red dashed curves are estimation from separated stratified and non-stratified model by modeling two events independently. All of them provide estimation close to true values It is not too surprising to find lager bias close to the right end since there is only 2 time units after crash available on average. The pointwise empirical coverage probability (CP) are compared. Again, they perform similarly with CP slightly vary around 95%. Mean square error of time-varying coefficients estimation comparison favors multi-type model with smaller values.



(a) Varying coefficient estimation



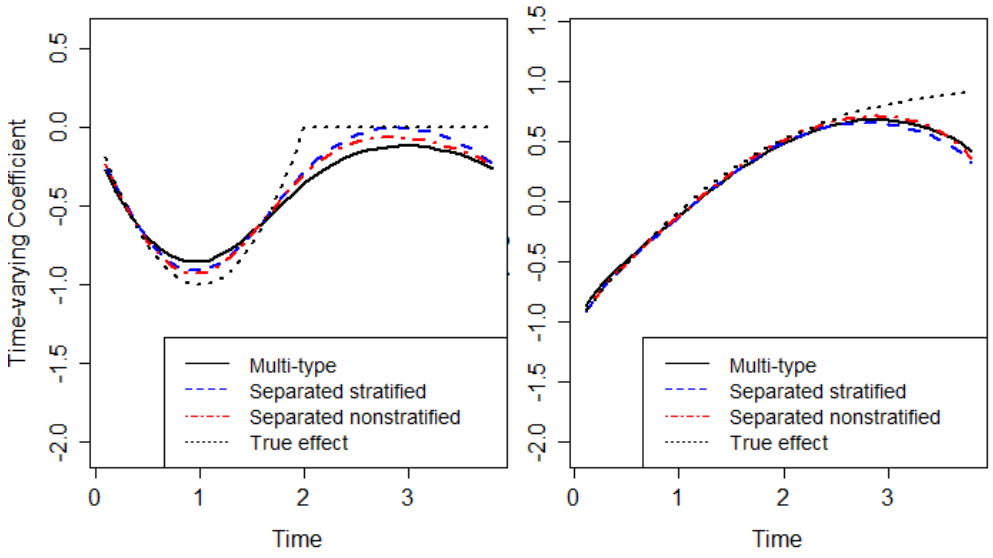
(b) Pointwise empirical CP and MSE

Figure 5.3: Simulation result: positive correlation.

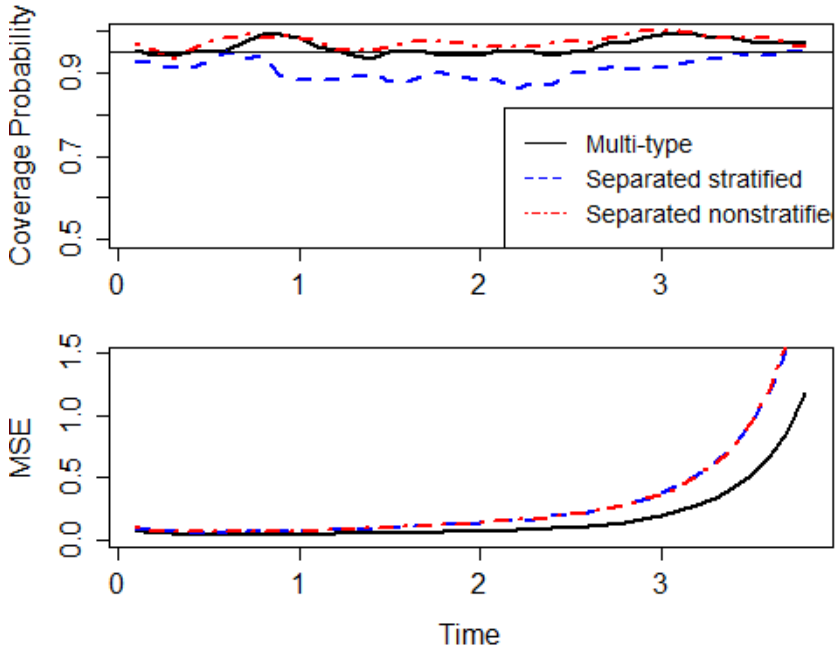
To examine the performance of the proposed method under negative correlation or stronger heterogeneity (e.g. larger variance of u_i and v_i), we further consider covariance matrix to $\begin{pmatrix} 0.5 & -0.3 \\ -0.3 & 0.25 \end{pmatrix}$ and $\begin{pmatrix} 1 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}$.

Figure 5.4 shows simulation results of negative correlation. Again, they perform similarly

with CP slightly vary around 95%. Mean square error of time-varying coefficients estimation comparison favors multi-type model with smaller values.



(a) Varying coefficient estimation



(b) Pointwise empirical CP and MSE

Figure 5.4: Simulation result: negative correlation.

Table 5.2 lists results of time-fixed coefficients, variance component, and smoothing parameter from various settings. In both settings, the bias of time-fixed coefficients are all

Table 5.2: Simulation results for parametric coefficients estimates in one time-varying effect model

Parameter	True value	Mean	Bias	SE $^{\%}$	SEM$^{\triangle}$	CP*
Setting I: positive correlation						
β_1^1	1	1.05	0.05	0.28	0.26	96%
β_2^2	-1	-0.98	0.02	0.22	0.23	95%
β_1^2	1	1.04	0.04	0.22	0.19	91%
β_2^2	-1	-0.97	0.03	0.15	0.17	96%
β_1^3	1	1.03	0.03	0.27	0.25	95%
β_2^3	-1	-1.04	-0.04	0.20	0.22	97%
θ_1	0.5	0.49	-0.01	0.12	0.07	74%
θ_2	0.25	0.27	0.02	0.07	0.04	71%
ρ	0.85	0.79	-0.06	0.09	0.13	98%
λ		(6, 33)				
Setting II: negative correlation						
β_1^1	1	1	0.00	0.26	0.26	92%
β_2^2	-1	-1.01	-0.01	0.23	0.23	96%
β_1^2	-0.5	-0.49	0.01	0.19	0.22	96%
β_2^2	0.5	0.49	-0.01	0.19	0.2	95%
β_1^3	1	1.03	0.03	0.27	0.25	94%
β_2^3	-1	-1.05	-0.05	0.22	0.22	95%
θ_1	0.5	0.44	-0.06	0.13	0.08	65%
θ_2	0.25	0.26	0.01	0.09	0.05	63%
ρ	-0.85	-0.7	0.15	0.19	0.21	94%
λ		(19, 65)				

%: Empirical standard error

$\triangle$: Mean of standard error

*: Coverage probability

small ($\leq 5\%$). The empirical coverage probabilites of 95% confidence intervals using the estimated standard error are around 95%. Bias for variance component θ is 8%. The estimated standard derivations of $\boldsymbol{\theta}$ (SEM) is smaller than empirical ones (SE).

In Table 5.3, there lists power/ type I error of proposed tests in subsection 5.3.3 under difference true functions. When true effect is not constant, power of test for no effect (85% 99%) is larger than that of constant effect (55% 99%). Test for equality of time-varying coefficients are reasonably well. In terms of test for fixed effect, under null hypothesis that all three sets of fixed effects are the same, type I error varies from 3% to 10%. The power to detect different effect of 1.5 is 100%.

Table 5.3: Empirical power/ type I error of tests for time-varying and time-fixed effects

True value	Null hypothesis		
	No effect	Constant	Equivalence
Setting I: Piecewise Polynomial; logistic	(91%, 100%)	(64%, 98%)	93%
Setting II: Piecewise Polynomial; logistic	(85, 99%)	(53, 99%)	81%
	$\boldsymbol{\beta}^1 = \boldsymbol{\beta}^2$	$\boldsymbol{\beta}^2 = \boldsymbol{\beta}^3$	$\boldsymbol{\beta}^1 = \boldsymbol{\beta}^3$
Setting I: $\boldsymbol{\beta}^1 = \boldsymbol{\beta}^2 = \boldsymbol{\beta}^3 = (1, -1)$	(8%, 8%)	(10%, 3%)	(7%, 4%)
Setting II: $\boldsymbol{\beta}^1 = \boldsymbol{\beta}^3 = (1, -1); \boldsymbol{\beta}^2 = (-0.5, 0.5)$	(100%, 100%)	(10%, 5%)	(100%, 100%)

5.6 Conclusion and discussion

In this Chapter, we propose a general platform for multi-type recurrent event data with both time-fixed and time-varying coefficients. Relationship among different types of events are assessed through correlated frailty term. It provides researchers insight of the relationship. Penalized B-spline was adopted to approximate time-varying coefficients. Writing extra penalties brought by variance components and smoothing parameters are estimated jointly by maximizing profile likelihood. Besides easy implementation, it allows us to make systematic inference on all components.

Application result indicates that male drivers tend to drive safer after crash with lower intensity rate of NC and SCI. The amount of decrease diminishes gradually over time. On contrary, female drivers do not show such pattern. Gender discrepancy is consistent between NC and SCI. Intensity rate of male drivers is about 0.8 times of that of female drivers. Older drivers more safe than younger drivers. Correlation confirms a strong positive association between two types of crash-related events.

Simulation result reveals good performance of the proposed model for both time-varying and time-fixed coefficients. Estimation almost captures true functions with small bias in the positive correlation. In the case of negative associate, we observe a larger bias. Comparison of estimation among three models do not imply outperformance of multi-type model. But MSE of multi-type models is uniformly smaller than single type which provide more precise estimation. Another interesting finding is that the smoothing parameter estimated in multi-type model is smaller than that from separated models. This agrees with larger bias and smaller variance.

One limitation of current study is computation. With larger number of subjects and/or more types of events in a study, computation load explodes dramatically. For a three-events model with 100 subjects, the size of hessian matrix is over 300×300. Estimation proce-

dure involves many times of inverse larger hessian matrix thus require more time. Future implementation of the proposed model will benefit very much from simplified algorithm.

Bibliography

[1] P. K. Andersen and R. D. Gill. Cox's Regression Model for Counting Processes: A Large Sample Study. *The Annals of Statistics*, 10(4):1100–1120, 1982.

[2] N. Breslow and D. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88:9–25, March 1993.

[3] J. Cai and D. E. Schaubel. Marginal means/rates models for multiple type recurrent event data. *Lifetime Data Anal.*, 10(2):121–138, 2004.

[4] R. J. Cook and J. F. Lawless. *The statistical analysis of recurrent events*. Statistics for biology and health. Springer, 2007.

[5] R. J. Gray. Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association*, 87(420): 942–951, 1992.

[6] X. Lin and D. Zhang. Inference in generalized additive mixed modelsby using smoothing splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61 (2):381–400, 1999.

[7] Y. Mazroui, S. Mathoulin-Plissier, G. MacGrogan, V. Brouste, and V. Rondeau. Multivariate frailty models for two types of recurrent events with a dependent terminal event: Application to breast cancer data. *Biometrical Journal*, 55(6):866–884, 2013. ISSN 1521-4036.

[8] F. O'Sullivan. Nonparametric estimation of relative risk using splines and cross-validation. *SIAM Journal on Scientific and Statistical Computing*, 9(3):531–542, 1988.

[9] S. Ripatti and J. Palmgren. Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics*, 56(4):1016–1022, 2000.

[10] P. G. Sankaran and P. Anisha. Shared frailty model for recurrent event data with multiple causes. *Journal of Applied Statistics*, 38(12):2859–2868, 2011.

[11] L. Sun, L. Zhu, and J. Sun. Regression analysis of multivariate recurrent event data with time-varying covariate effects. *Journal of Multivariate Analysis*, 100(10):2214 – 2223, 2009.

[12] M.-C. Wang, J. Qin, and C.-T. Chiang. Analyzing recurrent event data with informative censoring. *Journal of the American Statistical Association*, 96(455):1057–1065, 2001.

[13] Z. Yu, L. Liu, D. M. Bravata, L. S. Williams, and R. S. Tepper. A semiparametric recurrent events model with time-varying coefficients. *Statistics in Medicine*, 32(6): 1016–1026, 2013.

[14] L. Zhu, J. Sun, X. Tong, and D. Srivastava. Regression analysis of multivariate recurrent event data with a dependent terminal event. *Lifetime Data Analysis*, 16(4):478–490, 2010.

Chapter 6 General Conclusions

This chapter summarizes the major conclusions and contributions of this dissertation and suggests some possible future research direction.

6.1 Conclusion and Contribution

In this dissertation, different methodologies have been proposed in the content of recurrent events and applied to 100-Car NDS. Although the research was motivated by a transportation safety research question, the methodologies developed are general and can be applied to a wide spectrum of research fields.

In Chapter 2, the influence of crash was evaluated on drivers' distraction behavior. Distraction behavior was measured by secondary driving tasks. Crash influence on driving behavior was evaluated with a count-based approach using a mixed binomial regression model. The results indicate that drivers' engagement in moderate and complex secondary tasks tends to be lower after crashes, especially within a 15-hour driving time window. This decreasing effect tends to diminish over time.

Chapter 3 provides the first systematic study of crash influence on driving risk in the context of recurrent event model. Four semi-parametric recurrent event models were implemented and compared. Cox-Snell residual plots suggested that stratified frailty model fitted the data best. The results suggest that crashes have a positive effect on driver behavior with lower SCI intensity after crashes. Drivers might either learn from the crashes experience or be more attentive while driving.

The objective of Chapter 4 is to estimate time-varying coefficient, which was approximated by penalized B-spline. Variance components and smoothing parameters were estimated jointly by maximizing profile likelihood. We proposed to link time-varying coefficient model to a regular frailty model. This created an easy access to new approach and statistical inference for smoothing parameters. In addition, we studied the asymptotic distribution and conducted statistical tests. Application results showed driving risk of male drivers tend to decrease after crash first and then increase. Similar pattern was not found for female drivers. These findings provide crucial information for understanding drivers' response to dramatic driving events and can be critical for development safety education programs and safety counter measures.

In Chapter 5 I proposed a general platform of multi-type recurrent event models with

time-varying coefficients. The chapter focuses on two types of recurrent events, each in a shared frailty model. Relationship among different types of events was assessed from frailty terms. In terms of time-varying crash influence, results from multi-type model were similar to that in Chapter 4. A positive correlation was found between two types of crash-related events.

6.2 Future Work

Since there was no closed form for marginal likelihood given normally distributed frailty terms, Laplace approximation was adopted to achieve a close estimation. Another widely-used technique handling unobserved random effects is E-M algorithm. In the future, it is of interest to implement the EM algorithm and compare the results.

Regarding to time-varying coefficient model, model selection is worth future research. In a study with many potentially time-varying coefficients, quickly and correctly identifying them is essential. Large degrees of freedom is often in demand to estimate considerable time-varying coefficients. It could lead to identifiability issue. How to solve this problem is an appealing question.

Comparing time-varying coefficient in the context of recurrent event and frequency is another future topic. Similar to the study in Chapter 2, an alternative of examining time-varying coefficient is to divide study period into several intervals and model the occurrence in each period. It would be interesting to know which performs better in certain circumstances.

Computational issue affects future implementation of multi-type recurrent event model. With larger number of subjects and/or more types of events in a study, computation load increase substantially. Estimation procedure involves many times of optimization of a large amount of parameters thus require substantial computing resources. Future implementation of the proposed model for large data will benefit from simplified algorithm.

Lastly, application of this dissertation is based on a relative small number of crashes with mild crash severity. With larger NDS data sets becoming available, such as the Second Strategic Highway Research Program (SHRP 2) Naturalistic Driving Study, more concrete evidence will be available on the influence of crashes on driver behavior and potentially the influence of crashes by severity.