The Self-Controlled Case Series Design: A Simulation Study

by

©Kaida Cai

A Thesis submitted to the School of Graduate Studies in partial fulfillment of the requirements for the degree of

Master of Science

Department of Mathematics and Statistics

Memorial University of Newfoundland

07/2015

St. John's Newfoundland

Abstract

The self-controlled case series (SCCS) design is an outcome dependent sampling design developed to investigate the association between time-varying exposures and outcome events. This design automatically adjusts for all fixed covariates acting multiplicatively on the intensity function of a subject. It is based only on cases, and ignores controls. Since only cases are included, it is economically and computationally efficient compared with a cohort design. This property of the SCCS design also helps protecting data privacy. Because of these reasons, the SCCS design is an important alternative to the cohort design especially when the outcome of interest is a rare event, and has been used in many studies in medicine, epidemiology and pharmacoepidemiology. Therefore, the main objective of this thesis is to investigate the SCCS design through simulations. We considered parametric, semiparametric and weakly parametric SCCS models, and compared them with well-known models based on the classical cohort design. We also illustarted the methods with a real life data set from medicine.

Acknowledgments

I would like to expresss my sincere appreciation to Dr. Candemir Cigsar for his guidance and help of graduate thesis. I also would like to express my sincere apprecation to Dr. Zhaozhi Fan for his guidance and support for the learning and living during the master study period. I am gratefull to Dr. Alwell Oyet, Dr. Asokan Variyath, Dr. J Concepción Loredo-Osti and Dr. Yildiz Yilmaz for giving me such interesting and wonderful courses. Also, thanks to all the faculties and staff of Department of Mathematics and Statistics for provding a comfortable study environment. I would like to say thank you to my fanmily members and friends. Thanks for their trust and support during the master study period. Finally, thanks to this beautiful city, St.John's, where I lived and studied for the past two years.

Table of Contents

Abstract ii

Acknowledgments iii

Table of Contents vi

List of Tables viii

List of Figures ix

1 Introduction 1

1.1 The Self-Controlled Case Series Design 1

1.1.1 Types of Data 3

1.1.2 Example: Bleeding Disorders 5

1.1.3 Literature Review 5

1.2 Notation and Terminology 7

1.3 Fundamental Models 10

1.3.1 Poisson Processes 10

1.3.2 Renewal Processes 13

1.4 Simulation of Recurrent Event Processes 14

1.5 Outline of Thesis 17

2 Likelihood Based Estimation Methods for Recurrent Event Processes 18

2.1 Likelihood for Recurrent Event Data under the Cohort Design 19

2.1.1 Parametric Estimation 21

2.1.2 Semiparametric Estimation 23

2.2 Likelihood for the SCCS design 25

2.2.1 The Parametric SCCS Method 27

2.2.1.1 Simulation Study: Comparison of Parametric SCCS Model with Parametric Cohort Model 30

2.2.2 The Semiparametric SCCS Method 33

2.3 Analysis of Recurrent Event Data Using Piecewise-Constant Rate Functions 36

2.3.1 Piecewise-Constant Rate Models for Poisson Processes 36

2.3.2 Piecewise-Constant Rate Models for the SCCS Design 39

2.4 Simulation Study 42

3 A Simulation Study for the Comparison of Semiparametric SCCS model with SCCS model with Piecewise-Constant Baseline Rate Functions 47

3.1 Design of the Simulation Study 48

3.2 Simulation Results 52

4 Application: Measles, Mumps and Rubella Vaccination and Idiopathic Thrombocytopaenic Purpura 65

4.1 Background 66

4.2 Data Analysis 67

4.3 Conclusion 73

5 Summary and Conclusion **75**

Bibliography **77**

Appendix **84**

List of Tables

2.1 Averages of the estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ of β based on the cohort and SCCS models, respectively, are given under various combinations of $(N, \Delta, E\{N_i(\tau)\}, E\{N_i(\Delta)\})$. 32

2.2 Simulation results of Models 1, 2, 3 and 4. $\bar{\hat{\beta}}$ is the average estimates of β. $\bar{S}(\hat{\beta})$ is the average standard error. $\bar{Bias}(\hat{\beta})$ is the average bias. $\bar{RB}(\hat{\beta})$ and $\bar{MSE}(\hat{\beta})$ are relative bias and mean square error. 45

3.1 Simulation results for setting A with scenarios $m = 25$ $\Delta = 50$. 55

3.2 Simulation results for setting A with scenarios $m = 25$ $\Delta = 25$. 56

3.3 Simulation results for setting A with scenarios $m = 25$ $\Delta = 10$. 57

3.4 Simulation results for setting B with scenarios $m = 25$ $\Delta = 50$. 58

3.5 Simulation results for setting B with scenarios $m = 25$ $\Delta = 25$. 59

3.6 Simulation results for setting B with scenarios $m = 25$ $\Delta = 10$. 60

3.7 Simulation results for setting C with scenarios $m = 25$ $\Delta = 50$. 61

3.8 Simulation results for setting C with scenarios $m = 25$ $\Delta = 25$. 62

3.9 Simulation results for setting C with scenarios $m = 25$ $\Delta = 10$. 63

3.10 Simulation results for setting C with scenarios $m = 25$ $\Delta = 10$. 64

4.1 Estimation results for Model (4.1) are given. Δ denotes risk period in days. $SE(\hat{\gamma})$ and $SE(\hat{\beta})$ denote the standard errors of the maximum likelihood estimates $\hat{\gamma}$ and $\hat{\beta}$, respectively. $-\ell_c^{\max}$ is the negative of the log of $L_c(\hat{\gamma}, \hat{\beta})$ given in (4.2). 71

4.2 Estimates of β for Model 1, Model 2 and Model 3. Δ is the length value of risk period. $Obs(\Delta)$ and $Exp(\Delta)$ are the observed number of events in exposure time period and expected number of events in exposure time period when $\beta = 0$. $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$ are the estimate of β under Model 1, Model 2 and Model 3, respectively, and SE denotes their standard errors. 73

List of Figures

1.1 Multistate diagram of a recurrent event process. 3

3.1 Age effect in Setting 1. 51

3.2 Age effect in Setting 2. 51

3.3 Age effect in Setting 3. 52

4.1 The Nelson-Aalen plot of the cumulative mean function $\mu(t)$, where time t denotes days. 69

4.2 Event history plot of a subject with its corresponding path of the covariate $x_i(t)$, where the exposure (risk) period Δ is taken as 42 days. 70

Chapter 1

Introduction

1.1 The Self-Controlled Case Series Design

There is an increasing interest in outcome dependent sampling designs such as case cohort, case-control and nested case-control designs (Prentice, 1986; Borgan et al., 1995; Aalen et al., 2008; and Keogh and Cox, 2014). These designs are more appealing than the classical cohort design, in particular, in studies where a small fraction of individuals experience the outcome of interest, to which we will henceforth refer as an *event*. Although the analysis of regression models under these designs is well established in time to event settings, they are not considered in detail when events are allowed to occur more than once over time.

A relatively new outcome dependent sampling design, called self-controlled case series (SCCS), was proposed by Farrington (1995). The main goal of a SCCS design is to investigate the effect of time-varying exposures or conditions on a specified event. The SCCS methodology is based on the conditional Poisson cohort. As a result, the SCCS design can be used in settings where individuals can possibly experience an event of interest more than once over their lifetime; that is, in recurrent event settings.

In case-control studies, an individual who experiences an event is called a *case* and who is at risk of having an event is called a *control*. Typically, cases are first observed and their exposure and other related infromation are retrospectively collected. Then, a number of controls are sampled from individuals who are at risk of having an event in the cohort. Therefore, case-control studies are usually retrospective and observational studies (Keogh and Cox, 2014).

The SCCS design is a special type of case-control design which uses data only on cases, and ignores controls. In other words, individuals who have not experienced the event are not included in the study. The number of events experienced by an individual over the exposure periods is then compared with the number of events experienced by the same individual over his or her nonexposure periods; that is, a case serves as its own control. Therefore, this design is reasonably called self-controlled. Under certain conditions, which we will discuss in Chapters 2 and 5, a SCCS design provides consistent estimates of relative incidence of events. This design automatically adjusts for all fixed covariates acting multiplicatively on the intensity function of an individual. Since only cases are included, it is economically and computationally efficient compared with a cohort design, and helps protect data privacy. Because of these properties, SCCS design has received considerable recent attention, especially in pharmacoepidemiology studies based on large administrative data bases (Xu et al., 2013; Simpson, 2013).

There are important research questions related to estimation and statistical efficiency of the SCCS design. One particular research question which is of practical importance is the computational efficiency of the SCCS design. Because of its advantages, this design has been recently applied to the settings in which the outcome of interest is not a rare event, and computational efficiency is especially important in such settings. In this thesis, we discuss this issue under different estimation methods. We compare the

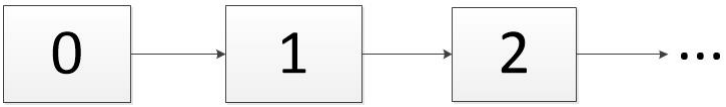

Figure 1.1: Multistate diagram of a recurrent event process.

SCCS method with the cohort design with parametric and semiparametric procedures. Semiparametric procedures provide greater flexibility in model fitting. However, they are computationally demanding when the data sets are large. Alternatively, parametric Poisson process models with piecewise-constant baseline intensities (Lawless, 1987; Hu and Lawless, 1996) can fit the SCCS design as well, and provide enough flexiblity to model the baseline intensity functions. Therefore, in this thesis we discuss parametric and semiparametric models for the cohort design and the SCCS design.

The remaining of this introduction chapter is organized as follows. In the remaining part of Section 1.1, we discuss the types of data in SCCS studies with an illustrative example. Section 1.2 introduces the notation and defintions frequently used in the thesis. Section 1.3 gives the fundamental models that are useful for modeling recurrent events. Simulation procedures are explained in Section 1.4. We give the outline of the thesis in the last section.

1.1.1 Types of Data

The initial intention of the SCCS method was to investigate the associations between vaccination and acute adverse events, which are potentially recurrent. Afterwards, it has been applied in other settings in epidemiology and pharmacoepidemiology as well. The processes generating recurrent events over time are called recurrent event processes and data sets obtained from such processes are called recurrent event data (Cook and Lawless, 2007). The multistate diagram of a recurrent event process is given in Figure 1.1, where each state is represented by the cumulative number of

events. Transitions between states are defined by the probabilistic characteristics of a recurrent event process.

Recurrent event data typically include event occurrence times or gap times (i.e., times between successive events), and a censoring time for each process under observation. The choice of a time scale usually depends on the goals of a study. The calendar time scale for time variable is often preferable if the event counts are of interest and processes involve humans or animals as subjects of a study. The gap times are useful when the interest is in modeling the duration between state transition. There are other measures that can be used as time scales such as operating time of a machine or milage scale for cars. In this thesis, the time scale is defined as calendar time, unless otherwise stated.

In many studies, data sets also include information about one or more covariates. Covariates can be classified as either internal or external. An external covariate is a variable whose values do not depend on the process. Othwerwise, it is called an internal covariate. For example, air pollution can be considered as an external covariate. External covariates can be time-varying or time-fixed.

Cook and Lawless (2007) consider the statistical analysis of recurrent event data in various settings. The statistical methods of different observation schemes of recurrent event processes such as intermittent observation or interval censoring are given in the literature. However, for convenience, we assume continuous observation of subjects over the followup period. We briefly discuss generaliztion of observation scheme of a process in Section 1.2. In a typical setting, where the association between a single point exposure and a recurrent event is of interest, the SCCS method requires information on start and end-of-followup times, event times and exposure time of subjects in a calendar time scale.

1.1.2 Example: Bleeding Disorders

In this section, we briefly explain the data set used in Chapter 4 to illustrate the methodology discussed in this thesis.

Farrington and Whitaker (2006) give the data to investigate the association between measles, mumps and rubella (MMR) vaccine and idiopathic thrombocytopenic pupura (ITP), which is a rare, potentially recurrent bleeding disorder. They hypothesise that the MMR vaccination may be associated with an increase in the relative incidence of ITP in children. The data set includes times of start and end of followup, administration of MMR vaccination and development of ITP in 35 children aged between 366 and 730 days. The data set is presented in the Appendix. The details of the data can be found in Chapter 4.

1.1.3 Literature Review

Farrington (1995) proposed the SCCS method to investigate associations between acute outcomes and transient exposures, using only data on cases. Since the SCCS design uses only cases, it is computationally efficient. Mainly because of this propoerty, it has received considerable recent attention. Therefore, technical issues and applications of the SCCS method have been later developed in specific studies.

Some properties of the SCCS method including its origins, assumptions and limitations are described in a tutorial manuscript with examples by Whitaker et al. (2006). Becker et al. (2006) applied the SCCS method to analyze a potential trigger of an acute illness. Douglas and Smeeth (2008) used the SCCS method to study the association between exposure to antipsychotics and the risk of stroke. Grosso et al. (2011) considered the SCCS method to address the issues in drug safety assessment. The review studies of application of the SCCS method in investigating potential associa-

tions between vaccines and adverse events are given by Weldeselassie et al. (2011). Some important issues in the development of the SCCS model have been discussed by Whitaker and Farrington (2006). Whitaker et al. (2009) presented both parametric and semiparametric SCCS models, and used them on the MMR vaccine and bleeding disorders examples.

More recently, the SCCS method has undergone extensive development. Vines and Farrington (2001) explained the SCCS method in case-crossover studies to address within subject exposure dependency. Hocine et al. (2005) tested the independence between two Poisson-generated multinomial variables in a SCCS desing. Li and Huang (2006) gave the existence and uniqueness of relative incidence estimates in the SCCS method. Musonda et al. (2006) discussed the sample size calculations for the SCCS design. Musonda et al. (2008) considered the small sample performances of the SCCS method.

An important development of the extension of the SCCS model was given by Farrington and Whitaker (2006). They introduced the semiparametric analysis of the SCCS model with applications to various data sets from epidemiology. Whitaker et al. (2007) gave an extension of the SCCS method for environmental time series data. Musonda (2006) and Simpson (2013) discussed the performance and design of the SCCS method in studies of vaccine safety.

Issues with modeling the SCCS design have been also discussed in the literature. Musonda et al. (2008) adapted SCCS method for routine surveillance of vaccine safety using cumulative sum charts. Farrington et al. (2009) discussed the case series analysis for censored, perturbed or curtailed post-event exposures. Within-individual dependence in the SCCS models for recurrent events are proposed by Farrington and Hocine (2010). A new adaption of the SCCS method, in which observation periods are truncated according to the vaccination schedule, are considered by Kuhnert et al.

(2011). Xu et al. (2011) identified optimal risk windows for SCCS studies of vaccine safety. Farrington et al. (2011) considered the issues about the SCCS analysis with event-dependent observation periods. Keogh and Cox (2014) mentioned about the connection between the analysis of the SCCS design with the analysis of systems representable with stochastic processes.

1.2 Notation and Terminology

In this section, we introduce the notation frequently used in the remaining parts of the thesis. The concepts of counting processes and intensity functions are useful in analyzing recurrent event data. For notational simplicity, we consider a single process in this section. However, we introduce the necessary subscripts later whenever a setup for multiple individuals is needed.

Suppose that a single process is under observation and that $T_0 < T_1 < T_2 < \cdots$ denote its event times, where T_k is the time of the kth event and $T_0 = 0$. Let $W_j = T_j - T_{j-1}$, $j = 1, 2, \ldots$, which is called the waiting time or gap time between the $(j-1)$st and jth events. The counting process $\{N(t); t \geq 0\}$ is a stochastic process which records the number of cumulative events occurring in the interval $[0, t]$. Let $I(E)$ be the indicator random variable of event E; that is, $I(E)$ is equal to 1 if event E occurs, and 0 otherwise. Then, $N(t) = \sum_{k=1}^{\infty} I(T_k \leq t)$ is the number of events occurring in the time interval $[0, t]$. Let $N(s, t)$ denote the number of events occurring in the time interval $(s, t]$; that is, $N(s, t) = N(t) - N(s)$. The mean and rate functions of $\{N(t); t \geq 0\}$ is defined as $\mu(t) = E\{N(t)\}$ and $\rho(t) = \mu'(t)$, respectively. The history of a counting process at time t is denoted by $\mathcal{H}(t) = \{N(s); 0 \leq s < t\}$, which includes all information about the counting process $\{N(t); t \geq 0\}$ up to time t, but not at time t. We use t^- and t^+ to denote the time period that is infinitesimally

smaller and larger, respectiely, than t. We let $\Delta N(t) = N((t+\Delta t)^-) - N(t^-)$ denote the number of events occurring in a small interval $[t, t+\Delta t)$.

We can now define an important concept in modeling recurrent event processes, the intensity function of a counting process. Let $\{N(t); t \geq 0\}$ be a counting process. Its intensity function is defined by

$$\lambda(t \,|\, \mathcal{H}(t)) = \lim_{\Delta t \downarrow 0} \frac{\Pr\{\Delta N(t) = 1 \,|\, \mathcal{H}(t)\}}{\Delta t}, \qquad t \geq 0. \tag{1.1}$$

The intensity function of a counting process gives the instantaneous probability of an event occurring at time t, conditional on the history of the process at time t. It completely specifies a counting process when t is continuous (Cook and Lawless, 2007). The intensity function (1.1) can be expanded by including fixed or time-varying covariates in the history of the process $\mathcal{H}(t)$. A mathematical model for a counting process can be then defined via its intensity function to include covariates so that effects of covariates on the event occurrences can be investigated. We discuss how to incorporate covariates into the intensity function of a Poisson process in the following section.

Another useful concept, especially in the analysis of gap times, is the hazard function. To define it, we let W be a continuous, nonnegative random variable. The cumulative distribution function (c.d.f.) and the probability density function (p.d.f.) of W is defined as $F(w) = \Pr\{W \leq w\}$ and $f(w) = dF(w)/dw$, respectively. The survival function of W is given by $S(w) = 1 - F(w)$. The hazard function of W is then given by

$$h(w) = \lim_{\Delta w \downarrow 0} \frac{\Pr\{w \leq W < w + \Delta w \,|\, W \geq w\}}{\Delta w}, \qquad w \geq 0. \tag{1.2}$$

It can be shown that $h(w) = f(w)/S(w)$. This result and other important properties of the hazard function can be found in Lawless (2003).

In this thesis, unless otherwise stated, we assume that subjects are continuously under obersevation. Under some assumptions, the methods can be generalized to different observation schemes as well. This generalization can be done via the at-risk indicator $Y(t)$. It is possibly a random indicator function which is equal to 1 when a subject is under observation and at-risk of experiencing the event of interest. It is equal to 0 otherwise. For instance, let a subject be under observation over a time interval $[\tau_0, \tau]$, which is called the observation window with the starting time τ_0 and end-of-followup time τ. Then, the at-risk indicator is $Y(t) = I(\tau_0 \leq t \leq \tau)$, which means that the subject is under observation and at risk of having an event over $[\tau_0, \tau]$. In this case, the intensity function of an observable counting process $\{\bar{N}(t) = \int_0^t Y(s)\, dN(s);\, t \geq 0\}$ can be defined as

$$\bar{\lambda}(t \,|\, \bar{\mathcal{H}}(t)) = \lim_{\Delta t \downarrow 0} \frac{\Pr\{\Delta \bar{N}(t) = 1 \,|\, \bar{\mathcal{H}}(t)\}}{\Delta t}, \tag{1.3}$$

where $\bar{\mathcal{H}}(t) = \{\bar{N}(s), Y(s); 0 \leq s < t\}$ is the history of the observable counting process at time t.

The advantages of using the at-risk function $Y(t)$ was discussed by Cook and Lawless (2007, Section 2.6). For example, methods can be generalized to intermittent observation schemes or observation schemes with random τ_0 and τ under certain conditions. These conditions basically postulate either the indepedence of $N(t)$ and $Y(t)$ or conditionally independence of $N(t)$ and $Y(t)$ given the history. In such cases, the intensity function of the observable process is proportional to the intensity function of the underlying process (Cook and Lawless, 2007, p. 49). That is,

$$\bar{\lambda}(t \,|\, \bar{\mathcal{H}}(t)) = Y(t)\, \lambda(t \,|\, \mathcal{H}(t)). \tag{1.4}$$

In this study, we use a continuous observation scheme over a prespecified (i.e., non-random) observation window. Therefore, we do not discuss these conditions further

here, but we will discuss difficulties with using the intensity function (1.4) in SCCS designs in the last chapter.

1.3 Fundamental Models

Poisson processes and renewal processes are two important families of models for recurrent event processes. Poisson processes are canonical models for the event counts over specified time intervals or space. Covariates can be incorporated into the Poisson process models. Parametric, semiparametric and nonparametric methods of model fitting and regression analysis for Possion processes are available. Renewal processes are useful when the analysis of gap times are of interest. Similar to the Poisson processes, models based on renewal processes can be extended to include covariates. In this section, we briefly introduce these two fundamental families of models. Since the SCCS design is based on a conditional Poisson process cohort, our focus is on the Poisson processes. Many of the results in this section can be found in point process textbooks such as Kingman (1993), Grandell (1997) and Daley and Vere-Jones (2003).

1.3.1 Poisson Processes

Suppose that $\{N(t); t \geq 0\}$ is a counting process with the intensity function $\lambda(t \,|\, \mathcal{H}(t))$. Also, suppose that the mean function $\mu(t) = E\{N(t)\}$ of the process is continuous and finite. Then, $\{N(t); t \geq 0\}$ is a Poisson process if its intensity function is given by

$$\lambda(t \,|\, \mathcal{H}(t)) = \rho(t), \qquad t > 0, \tag{1.5}$$

where $\rho(t) = \mu'(t)$ is the rate function of the process (Cook and Lawless, 2007). Since the intensity function (1.5) does not depend on the history $\mathcal{H}(t)$, Poisson processes are Markovian. A Poisson process is called a homogeneous Poisson process when the rate

function $\rho(t)$, $t > 0$, is constant. Otherwise, it is called a nonhomogeneous Poisson process.

Another characterization of a Poisson process can be given as follows. Suppose that events are occuring randomly over time, and the random variable $N(t)$ gives the number of events by time t. Then, the process $\{N(t); t \geq 0\}$ is called a Poisson process with intensity function $\rho(t)$, $0 \leq t < \infty$, if

1. $\Pr\{N(0) = 0\} = 1$, and

2. The number of events occurring over non-overlapping time intervals are independent, and

3. Let $N(s,t) = N(t) - N(s)$ and $\mu(s,t) = \mu(t) - \mu(s)$, where $\mu(t) = \int_0^t \rho(u)\, du$ and $0 \leq s < t$. Then,

$$\Pr\{N(s,t) = n\} = \frac{e^{-\mu(s,t)}\, \mu^n(s,t)}{n!}, \qquad n = 0, 1, 2, \ldots. \tag{1.6}$$

The second condition in the above characterization is sometimes called the independent increment property of the Poisson process, and can be rephrased as follows. For any $0 \leq s_1 < t_1 \leq s_2 < t_2$, the random variables $N(s_1, t_1)$ and $N(s_2, t_2)$ are independent. This property also shows that Poisson processes are Markovian. The third condition states that $N(s,t)$ is a Poisson random variable with mean $\mu(s,t)$ for any $0 \leq s < t$.

The following well-known proposition is useful in simulations to generate a realization of a homogeneous Poisson process with rate function ρ. Its proof can be found in Rigdon and Basu (2000, pp. 45–49).

Proposition 1.3.1. *Let $\{N(t); t \geq 0\}$ be a counting process with the intensity function $\lambda(t \,|\, \mathcal{H}(t))$. The process $\{N(t); t \geq 0\}$ is a homogeneous Poisson process where*

$\lambda(t \mid \mathcal{H}(t)) = \rho$, $0 \leq t < \infty$, if and only if the gap times W_j, $j = 1, 2, \ldots$, are independent and identically (i.i.d.) distributed exponential random variables with mean ρ^{-1}. In this case, the p.d.f. of the W_j is given by $f(w) = \rho\, e^{-\rho w}$, $0 < w < \infty$.

The following proposition is taken from Cook and Lawless (2007, p. 33), and can be used to generate event times of a nonhomogeneous Poisson process.

Proposition 1.3.2. *Let $\{N(t); t \geq 0\}$ be a nonhomogeneous Poisson process with rate function $\rho(t)$ and mean function $\mu(t) = \int_0^t \rho(u)\, du$. Let $s = \mu(t)$ be a new time scale and $\{N^*(s); s \geq 0\}$ be a process, where $N^*(s) = N(\mu^{-1}(s))$, $0 < s$. Then, $\{N^*(s); s \geq 0\}$ is a homogeneous Poisson process with rate function $\rho^*(s) = 1$.*

A Poisson process can be easily extended to include covariates. Let $x(t) = (x_1(t), \ldots, x_p(t))'$ be a $p \times 1$ vector of covariates. Then, the intensity function of a Poisson process is given by

$$\lambda(t \mid \mathcal{H}(t)) = \rho_0(t)\, g(x(t); \beta), \qquad t \geq 0, \tag{1.7}$$

where $\rho_0(t)$ is a baseline intensity or baseline rate function, $g(x(t); \beta)$ is a positive-valued function of covariates $x(t)$ and β which is a $p \times 1$ vector of unknown parameters. A convenient choice of the function $g(x(t); \beta)$ is given by

$$g(x(t); \beta) = \exp(x'(t)\, \beta), \tag{1.8}$$

which guarantees the positiveness of $g(x(t); \beta)$.

The model (1.7) is called the multiplicative or proportional intensity model. There are also additive models with intensity function $\lambda(t \mid \mathcal{H}(t)) = \rho_0(t) + g(x(t); \beta)$ (Aalen et al., 2008). However, the methods in this thesis are based on the multiplicative models. Also, methods and models are termed *parametric* if the baseline intensity

function $\rho_0(t)$ is specified with a vector of parameters and *semiparametric* if it is left unspecified.

1.3.2 Renewal Processes

Models based on renewal processes are widely used to analyze the gap times between successive events. These models are especially useful if there is a type of renewal occurs at each event time. For example, in reliability analysis if the event of interest is a failure of a component of a machinery and if this component is replaced at each failure, a renewal process is suitable for modeling purposes.

Suppose that $0 = T_0 < T_1 < T_2 < \cdots$ are the event times of a process $\{N(t); t \geq 0\}$, and at most one event can occur at any time t. The process $\{N(t); t \geq 0\}$ is a renewal process if the gap times $W_j = T_j - T_{j-1}$ are i.i.d. with a c.d.f. $F(w)$. In renewal process settings, the event times T_j are sometimes called renewal times to imply the occurrence of a renewal at each event time. Similarly, the gap times W_j are sometimes called the inter-renewal times. It should be noted that if the inter-renewal times of a renewal process $\{N(t); t \geq 0\}$ have an exponential distribution with mean ρ^{-1}, then $\{N(t); t \geq 0\}$ is equivalent of a Poisson process with the intensity function ρ. In that respect, renewal processes can be considered as a generalization of homogeneous Poisson processes where the inter-arrival times are allowed to have a distribution other than the exponential distribution.

Renewal processes can be classified by their intensity functions as well. The intensity function of a renewal process $\{N(t); t \geq 0\}$ is of the form

$$\lambda(t \,|\, \mathcal{H}(t)) = h(B(t)), \qquad t > 0, \tag{1.9}$$

where $B(t)$ is the backward recurrence time (the elapsed time since the most recent

event before t), and $h(w)$ is the hazard function for the gap times W_j defined in (1.2). Covariates can be included in a renewal process by expanding the intensity function (1.9). Following the setup given in Section 1.3.1, a very convenient and flexible model including covariates is the proportional hazards model in which the hazard function is given by $h(w|x) = h_0(w)\exp(x'\beta)$, $w > 0$, where x is a $p \times 1$ vector of fixed covariates and $h_0(w)$ is the baseline hazard function. When $x(t)$ includes funtions of the time variable t or the history $\mathcal{H}$(t), a useful model is given by the intensity function

$$\lambda(t \,|\, \mathcal{H}(t)) = h(B(t))\, e^{x'(t)\,\beta}, \qquad t > 0, \tag{1.10}$$

which is called the modulated renewal process. Parametric and semiparametric methods for modulated renewal processes are available (Oakes and Cui, 1993; and Cook and Lawless, 2007).

1.4 Simulation of Recurrent Event Processes

The goal of this section is to introduce a general procedure to generate a recurrent event process with a given intensity function. We first present an algorithm to generate event times of a homogeneous Poisson process. Next, we discuss how to generalize this algorithm to generate event times of a nonhomogeneous Poisson process. Finally, we give an algorithm that can be used to generate event times of a general recurrent event process. It should be noted that the second algorithm covers the first one as a special case, but since Poisson processes are used extensively in this thesis, we prefer to present separate algorithm for them.

In the following discussion, we assume that the process is continuously observed over the observation window $[0, \tau]$, where the end-of-follow up time τ is a fixed positive real number. Therefore, we safely drop the at-risk function $Y(t) = I(0 \le t \le \tau)$.

Let $\{N(t); t \geq 0\}$ be a counting process and $\lambda(t \,|\, \mathcal{H}(t))$ be its associated intensity function. We first consider the simulation of a homogeneuous Poisson process with the intensity function ρ, and give a computer algorithm to generate the event times T_j in $[0, \tau]$. Let U be a random variable having the p.d.f. $f(x) = 1$, if $0 < x < 1$, and $f(x) = 0$, otherwise; which is a standard uniform distribution denoted by $U \sim \text{Unif}(0, 1)$. By using the result given in Proposition 1.3.1, the steps of a computer algorithm in this simple setting are given as follows:

1. Set $t = 0$ and $j = 0$.

2. Generate a random number U from a standard uniform distribution.

3. Let $t = t - \rho^{-1} \log(U)$ and if $t > \tau$ stop.

4. If $t \leq \tau$, advance j by 1 and let $T_j = t$.

5. Go to step 2.

At the end of the above algorithm, the final value of J gives the total number of events over the observation window $[0, \tau]$ and the T_j are the J event times in an increasing order.

The above algorithm can be used to generate the event times of a nonhomogeneous Poisson process as follows. Let $\{N^*(t); t \geq 0\}$ be a homogeneous Poisson process with rate function $\rho^* = 1$ and mean function $\mu^*(t) = t$. In this case, the third step in the above algorithm gives $\mu^*(T_j) = \mu^*(t_{j-1}) - \log(U)$. Let $\{N(t); t \geq 0\}$ be a nonhomogeneous Poisson process with rate function $\rho(t)$ and mean function $\mu(t)$. Then, by Proposition 1.3.1 the inverse transformation $T_j = \mu^{-1}(\mu^*(t_{j-1}) - \log(U))$ gives the jth event time of the nonhomogeneous Poisson process (see, Lewis and Shedler, 1976).

We give now an algorithm to generate event times of a recurrent event process with a general intensity function $\lambda(t \,|\, \mathcal{H}(t))$ over the observation window $[0, \tau]$, but we first need the following result. Its proof can be found in Cook and Lawless (2007, p. 30).

Proposition 1.4.1. *Let* $\{N(t); t \geq 0\}$ *be a counting process with the intensity function* $\lambda(t \,|\, \mathcal{H}(t))$. *Then,*

$$\Pr\{W_j > w \,|\, T_{j-1} = t_{j-1}, \mathcal{H}(t_{j-1})\} = e^{-\int_{t_{j-1}}^{t_{j-1}+w} \lambda(s \,|\, \mathcal{H}(s))\, ds}, \tag{1.11}$$

where $W_j = T_j - T_{j-1}$, $j = 1, 2, \ldots,$ *and* $T_0 = 0$.

We now define the random variable

$$E_j = \int_{t_{j-1}}^{t_{j-1}+W_j} \lambda(t \,|\, \mathcal{H}(t))\, dt, \qquad j = 1, 2, ..., \tag{1.12}$$

where the W_j are the gap times generated by the process $\{N(t); t \geq 0\}$ with intensity $\lambda(t|H(t))$ and $t_0 = 0$. From the result of Proposition 1.4.1, it is easy to see that, given t_{j-1} and $\mathcal{H}(t_{j-1})$, each random variable E_j has an exponential distribution with mean 1. Therefore, $U = \exp(-E_j)$ has a standard uniform distribution. For the purpose of generating event times of a recurrent event process with a general intensity function, the algorithm steps of a computer simulation procedure are given as follows.

1. Set $j = 1$ and $t_0 = 0$.

2. Generate U_j from a standard uniform distribution.

3. Use the transformation $E_j = -\log(U_j)$.

4. Calculate the jth event time T_j by solving $E_j = \int_{t_{j-1}}^{T_j} \lambda(t \,|\, \mathcal{H}(t))\, dt$ for T_j, where $T_j = t_{j-1} + W_j$.

5. If $T_j < \tau$, advance j by 1 and let $t_{j-1} = T_{j-1}$. Then, return to the second step. Otherwise, stop the loop and the recurrent event times observed over $[0, \tau]$ are given by t_1 ,. . ., t_n, where $n = j - 1$.

It should be noted that in the above algorithm we need to solve the equation (1.12) to find the gap times W_j. In some settings, the W_j can be obtained analytically. Otherwise, numerical procedures can be used. Also, in the simulation procedure, the model and the history $\mathcal{H}(t)$ may include external covariates. There are other algorithms to generate event times of an intenstiy based model (for example, see, Daley and Vere-Jones, 2003). However, for the purpose of this study, the above algorithms are useful.

1.5 Outline of Thesis

The remaining parts of the thesis are organized as follows. In Chapter 2, we set up framework for likelihood procedures under parametric and semiparametric models for the cohort and SCCS designs. We also give the results of simulation studies to investigate the bias in estimates of parameters in the models based on the cohort and SCCS designs. Furthermore, we describe the piecewise-constant rate functions methodology for Poisson processes and the SCCS models. In Chapter 3, we describe a simulation study for the investigation of the bias in estimates of parameters in the semiparametric SCCS model and the SCCS model with piecewise-constant rate functions. The results of this extensive simulation study is also given in Chapter 3. In Chapter 4, we analyze a data set from medicine by applying the methods explained in Chapter 2. This study is briefly explained in Section 1.1.2. Finally, in Chapter 5, we give the summary and conclusion, as well as some future research topics.

Chapter 2

Likelihood Based Estimation Methods for Recurrent Event Processes

In this chapter, we focus on parametric and semiparametric estimation methods based on likelihood function developed under cohort and the self-controlled case series (SCCS) designs. A rigorous treatment of the development of likelihood function for counting processes in cohort settings is given by Andersen et al. (1993). A very detailed treatment of parametric and semiparametric methods for the analysis of recurrent event processes in the cohort design is given by Cook and Lawless (2007). The parametric SCCS model is introduced by Farrington (1995). The semiparametric analysis of the SCCS model is discussed by Farrington and Whitaker (2006). Their discussion includes large sample properties of the estimators based on SCCS design. Our goals in this chapter are to summarize important results and present the likelihood functions in different settings. There are four main settings:

1. Parametric estimation under the cohort design,

2. Semiparametric estimation under the cohort design,

3. Parametric estimation under the SCCS design, and

4. Semiparametric estimation under the SCCS design.

We consider the estimation of the relative incidence rate under all these settings. Furthermore, we discuss the flexible parametric models for the cohort and SCCS designs based on piecewise-constant baseline rate functions. Without loss of generality, throughout this chapter, we assume that processes are observed continously over prespecified observation windows.

This chapter is outlined as follows. In the next section, we will introduce the likelihood procedures for the cohort design. This includes a discussion about the development of the likelihood function for recurrent event processes, and then, parametric and semiparametric methods for the estimation of parameters. In Section 2.2, we explain the development of the likelihood function under the SCCS design. We next discuss the parametric and semiparametric estimation based on the SCCS model. This section also includes the results of a simulation study for comparison of the bias in estimation of the relative incidence rate under the parametric SCCS model with that under the parametric cohort design. Section 2.3 gives the set-up for the models with piecewise-constant baseline functions under cohort and SCCS designs. In the last section, we presented the results of a simulation study.

2.1 Likelihood for Recurrent Event Data under the Cohort Design

We first consider the likelihood-based parametric estimation procedures. To develop estimation methods, we need to write down the likelihood function for data observed

over an observation window.

Suppose that there are N independent subjects in a cohort. The followup of the ith subject, $i = 1, \ldots, N$, starts at time τ_{0i} and stops at time τ_i. Also, suppose that the event ocurrences of the counting process $\{N_i(t); t \geq 0\}$, $i = 1, \ldots, N$, are governed by the intensity function

$$\lambda_i(t \,|\, \mathcal{H}_i(t)) = \lim_{\Delta t \downarrow 0} \frac{\Pr\{\Delta N_i(t) = 1 \,|\, \mathcal{H}_i(t)\}}{\Delta t}, \tag{2.1}$$

where $\mathcal{H}_i(t) = \{N_i(t); t \geq 0\}$ is the history of the process.

From the intensity function (2.1) and under the assumption that two or more events cannot occur at the same time, the jump probabilities of the process $\{N_i(t); t \geq 0\}$ in a small interval $[t, t + \Delta t)$ are given by

$$\Pr\{\Delta N_i(t) = 0 \,|\, \mathcal{H}_i(t)\} = 1 - \lambda_i(t \,|\, \mathcal{H}_i)\, \Delta t + o(\Delta t), \tag{2.2}$$

$$\Pr\{\Delta N_i(t) = 1 \,|\, \mathcal{H}_i(t)\} = \lambda_i(t \,|\, \mathcal{H}_i)\, \Delta t + o(\Delta t), \tag{2.3}$$

and

$$\Pr\{\Delta N_i(t) > 1 \,|\, \mathcal{H}_i(t)\} = o(\Delta t), \tag{2.4}$$

where $o(t)$ represents a function $g(t)$ with $g(t)/t \to 0$ as $t \to 0$. These jump probabilities and the concept of product integration (Andersen et al., 1993, Section II.6) can be used to develop the likelihood function for recurrent event processes after considering a partition of the observation window $[\tau_{0i}, \tau_i]$ and then taking the limit. Here, we present the following result without a proof (see Cook and Lawless (2007, pp. 28–30) for a skecth proof and Andersen et al. (1993, Section II.7) for a more comprehensive discussion).

In this setup, conditional on $\mathcal{H}_i(\tau_{0i})$, the probability of the event {exactly n_i events

occur at time $t_{i1} < \cdots < t_{in_i}$ over the observation window $[\tau_{0i}, \tau_i]\}$ is given by

$$L_i = \left[\prod_{j=1}^{n_i} \lambda(t_{ij} \,|\, \mathcal{H}_i(t_{ij}))\right] \exp\left\{-\int_{\tau_{0i}}^{\tau_i} \lambda_i(s \,|\, \mathcal{H}_i(s))\, ds\right\} \tag{2.5}$$

Therefore, the likelihood function for N independent subjects is then given by

$$L = \prod_{i=1}^{N} L_i. \tag{2.6}$$

The validity of the likelihood function (2.6) in more complicated observation schemes is discussed by Cook and Lawless (2007, Section 2.6).

2.1.1 Parametric Estimation

We now discuss the parametric maximum likelihood estimation method for Poisson processes. We, therefore, specifiy the intensity function (2.1) fully parametrically as follows.

Let $\theta = (\alpha', \beta')'$ be a $q \times 1$ vector of parameters, where $\alpha = (\alpha_1, \ldots, \alpha_r)'$ is an r-dimensional vector of parameters, $\beta = (\beta_1, \ldots, \beta_p)'$ is a p-dimensional vector of parameters, and $q = r + p$. Suppose that the intensity function is specified by θ as follows. For $i = 1, \ldots, N$,

$$\lambda_i(t \,|\, \mathcal{H}_i(t); \theta) = \rho_i(t; \theta) = \rho_0(t; \alpha)\, e^{x_i'(t)\beta}, \tag{2.7}$$

where $\rho_0(t; \alpha)$ is the baseline rate function parametrically specified and $x_i(t) = (x_{i1}(t), \ldots, x_{ip}(t))'$ is a $p \times 1$ vector of covariates for the ith subject. From the function (2.5), the contribution of the ith subject to the likelihood function (2.6) is

$$L_i(\theta) = \left[\prod_{j=1}^{n_i} \rho_i(t_{ij}; \theta)\right] \, e^{-\int_{\tau_{0i}}^{\tau_i} \rho_i(s;\theta)\, ds}. \tag{2.8}$$

The log likelihood function is then given by $\ell(\theta) = \sum_{i=1}^{N} \ell_i(\theta)$, where

$$\ell_i(\theta) = \sum_{j=1}^{n_i} [\log \rho_0(t_{ij}; \alpha) + x_i'(t_{ij})\, \beta] - \int_{\tau_{0i}}^{\tau_i} \rho_0(s; \alpha)\, e^{x_i'(s)\beta}\, ds. \tag{2.9}$$

Under regularity conditions, we can obtain the score vector $U(\theta) = (U_\alpha'(\theta), U_\beta'(\theta))'$ where

$$U_\alpha(\theta) = \left(\frac{\partial \ell(\theta)}{\partial \alpha_1}, \ldots, \frac{\partial \ell(\theta)}{\partial \alpha_r} \right)', \tag{2.10}$$

and

$$U_\beta(\theta) = \left(\frac{\partial \ell(\theta)}{\partial \beta_1}, \ldots, \frac{\partial \ell(\theta)}{\partial \beta_p} \right)', \tag{2.11}$$

with components

$$\frac{\partial \ell(\theta)}{\partial \alpha_l} = \sum_{i=1}^{N} \left\{ \sum_{j=1}^{n_i} \frac{\partial \log \rho_0(t_{ij}; \alpha)}{\partial \alpha_l} - \int_{\tau_{0i}}^{\tau_i} \frac{\partial \rho_0(s; \alpha)}{\partial \alpha_l} e^{x_i'(s)\,\beta}\, ds \right\}, \qquad l = 1, \ldots, r. \tag{2.12}$$

and

$$\frac{\partial \ell(\theta)}{\partial \beta_k} = \sum_{i=1}^{N} \left\{ \sum_{j=1}^{n_i} x_{ik}(t_{ij}) - \int_{\tau_{0i}}^{\tau_i} \rho_0(s; \alpha)\, x_{ik}(s)\, e^{x_i'(s)\,\beta}\, ds \right\}, \qquad k = 1, \ldots, p, \tag{2.13}$$

respectively. The maximum likelihood estimate $\hat{\theta} = (\hat{\alpha}', \hat{\beta}')'$ of θ can be found by solving $U(\theta) = 0$, where 0 is a q-dimensional vector of zeros.

Let $I(\theta)$ be $q \times q$ observed information matrix partitioned as follows.

$$I(\theta) = \begin{pmatrix} I_{\alpha\alpha}(\theta) & I_{\alpha\beta}(\theta) \\ I_{\beta\alpha}(\theta) & I_{\beta\beta}(\theta) \end{pmatrix}, \tag{2.14}$$

where the components of $I(\theta)$ are $I_{\alpha\alpha}(\theta) = (-\partial/\partial\alpha')\, U_\alpha(\theta)$, $I_{\alpha\beta}(\theta) = (-\partial/\partial\beta')\, U_\alpha(\theta)$, $I_{\beta\alpha}(\theta) = (-\partial/\partial\alpha')\, U_\beta(\theta)$, and $I_{\beta\beta}(\theta) = (-\partial/\partial\beta')\, U_\beta(\theta)$. The maximum likelihood estimates $\hat{\theta}$ are usually obtained through an optimization software with the maxi-

mization of the log likelihood function $\ell(\theta)$. We use the `nlm` function in R software for this purpose. It also produces the Hessian matrix $-I(\theta)$ evaluated at $\hat{\theta}$ so that we also obtain the estimate of the asymptotic covariance matrix for $\hat{\theta}$ through the `nlm` function.

2.1.2 Semiparametric Estimation

In many settings, it is desirable to leave the baseline rate function $\rho_0(t)$ in (2.7) parametrically unspecified. A useful semiparametric specification of the regression model for Poisson processes is given below.

Let $\beta = (\beta_1, \ldots, \beta_p)'$ be a $p \times 1$ vector of parameters and, by convenience, let $\tau_{0i} = 0$ for $i = 1, \ldots, N$. The intensity function of the ith subject, $i = 1, \ldots, N$, is given by

$$\lambda_i(t \,|\, \mathcal{H}_i(t); \beta) = \rho_i(t; \beta) = \rho_0(t)\, e^{x_i'(t)\beta}, \tag{2.15}$$

where no parametric form is specified for the baseline rate function $\rho_0(t)$, and $x_i(t) = (x_{i1}(t), \ldots, x_{ip}(t))'$ is a $p \times 1$ vector of covariates of the ith subject.

The semiparametric multiplicative intensity model (2.15) is proposed by Andersen and Gill (1982), and called Andersen-Gill model. They extended the Cox's regression model for the survival data (Cox, 1972) to recurrent event settings, and obtained maximum likelihood estimates of parameters β and showed their large sample properties by applying the statistical theory for counting processes. The method of estimation can be based either on profile likelihood or partial likelihood function (see, Andersen et. al, 1993, Section VII.2). Here, we consider the profile likelihood approach. Both procedures lead to exactly the same inference on β.

Let $Y_i(t)$, $i = 1, \ldots, N$, be the at-risk indicator of the ith subject as explained in the previous chapter, and $dN_i(t)$, $i = 1, \ldots, N$, denote a small increment of $N_i(t)$ over

the time interval $[t, t+dt)$. Then, the log likelihood function ℓ_i, $i = 1, \ldots, N$, in (2.9) with the semiparametric model (2.15) can be rewritten as follows.

$$\ell_i(\beta) = \int_0^\infty Y_i(s) \left[\log \rho_i(s;\theta)\, dN_i(s) - e^{x_i'(s)\beta}\, d\mu_0(s) \right], \tag{2.16}$$

where $\mu_0(s) = \int_0^s \rho_0(u)\, du$ and $y_i(s)$ is at-risk indicator. The full log likelihood function is then given by

$$\ell(\beta) = \sum_{i=1}^{N} \ell_i(\beta). \tag{2.17}$$

Under the assumption that two or more events cannot occur simultaneously and by considering β as fixed, the likelihood function (2.17) can be maximized with respect to $d\mu_0(s) = \rho_0(s)\, ds$ by solving the Poisson profile score equation (Andersen et al., 1993, Section VII.2.1; also, see Cook and Lawless, 2007, Section 3.4.2)

$$\sum_{i=1}^{N} Y_i(s) \left\{ dN_i(s) - e^{x_i'(s)\beta} d\mu_0(s) \right\} = 0, \qquad 0 \le s, \tag{2.18}$$

which gives the quantity

$$d\mu_0^*(s) = \frac{\sum_{i=1}^N Y_i(s)\, dN_i(s)}{\sum_{i=1}^N Y_i(s)\, e^{x_i'(s)\beta}}. \tag{2.19}$$

If we let $U(\beta) = (\partial/\partial\beta)\ell(\beta)$, we obtain $p \times 1$ vector of score equations

$$U(\beta) = \sum_{i=1}^{N} \int_0^\infty Y_i(s)\, x_i(s) \left[dN_i(s) - e^{x_i'(s)\beta}\, d\mu_0(s) \right]. \tag{2.20}$$

An estimator $\hat{\beta}$ of β can be obtained by solving $U(\beta) = 0$, where 0 is a p-dimensional vector of zeros, after plugging the quantity $d\mu_0^*(s)$ into (2.20). The asymptotic covariance matrix and large sample properties of $\hat{\beta}$ is given by Andersen and Gill (1982). An important remark is that the components of the score vector (2.20) is exactly

the same with the partial likelihood score functions obtained by maximizing the Cox partial likelihood function for survival models with respect to β (see, Andersen et. al, Section VII.2.1). This fact allows us to adapt the Cox model software for survival models to Andersen-Gill model for recurrent event processes. Therefore, we can obtain $\hat{\beta}$ and estimate of its covariance matrix through a Cox model software. In this thesis, we used the `coxph` function in R to obtain $\hat{\beta}$ and its estimated covariance matrix.

It worths mentioning that the Andersen-Gill model (2.15) requires covariate infromation from all subjects in the cohort to obtain $\hat{\beta}$. Because of this, the Andersen-Gill model may not be cost-effective in some settings, in particular, when the cohort size is large or expensive covariates are of interest. In that respect, outcome dependent sampling designs are important alternatives to cohort designs. In the next section, we consider parametric and semiparametric estimation methods in SCCS design, which is an outcome dependent sampling design.

2.2 Likelihood for the SCCS design

The SCCS model was first developed to estimate the relative incidence of acute events following transient exposures (Farrington, 1995). It provides an alternative method for investigating the association between outcome events and time varying exposures. The SCCS design is a specific type outcome dependent sampling design, in which only cases are sampled. It is self-controlled because the estimate of the relative incidence is obtained by comparing the number of events occurred over the exposure times against the number of events occurred over the non-exposure intervals of the same subject. It is, thus, appropriately named as "self-controlled."

In this section, we introduce a setup for the SCCS method. We use this setup first to explain parametric estimation and then semiparametric estimation method. In the

following development, the key point is that the SCCS methodology should address the fact that the design is based on sampling only subjects with at least one event, i.e. cases, and ignoring all subjects without event, i.e. controls.

We will consider a single type of recurrent event and a single exposure period, which is called the risk period. This simple setup allows us to make a better comparison of the SCCS method with other methods. However, the SCCS method can be extended to deal with multitype events and more complex exposure schemes.

We consider a cohort of N independent subjects who are at risk of being exposed to an external condition for a time period called risk period and denoted by Δ. The external condition can be a point exposure which can possibly increase the risk of experiencing an event of interest for a short time period. Subjects are observed over the observation window $(\tau_{0i}, \tau_i]$ and each of them have n_i events, $i = 1, \ldots, N$. We let m denote the number of cases. It is possible that some of the cases may not be exposed. The intensity function of the SCCS model is

$$\begin{aligned} \lambda_i(t \,|\, \mathcal{H}_i(t)) = \rho_i(t) &= \rho_0(t)\, \exp\{\gamma_i + x_i(t)\beta\}, \qquad i = 1, \ldots, N, \\ &= \eta\, \psi(t)\, \exp\{\gamma_i + x_i(t)\beta\}, \qquad i = 1, \ldots, N, \end{aligned} \tag{2.21}$$

where $\rho_0(t)$ is the age-specific baseline rate function, η is the age effect at the start of the followup, $\psi(t)$ is the age-specific relative incidence, γ_i represents all fixed covariates and random effects, and the external covariate $x_i(t)$ denotes the time-varying exposure that is experienced at age t (Farrington and Whitaker, 2006).

In the model (2.21), the covariate $x_i(t)$ is an indicator function for the risk period. For example, let e_i denote the time of occurrence of the external condition such as administration of a vaccine for the ith subject, $i = 1, \ldots, N$. Suppose that the risk period starts immidiately at the occurrence of an external condition. Then, the

covariate $x_i(t)$ takes the value of 1 over the time interval $(e_i, e_i + \Delta]$, i.e. over the risk period. Otherwise, it is equal to 0.

The main goal of a SCCS study is to investigate whether the external condition is associated with an increased risk of experiencing an event over the risk period. Therefore, the main objective of a SCCS design is to make inference about the parameter β.

2.2.1 The Parametric SCCS Method

In the parametric SCCS method, the age-specific baseline rate function $\rho_0(t)$ is parametrically specified. In this section, we consider a simple setting, where a constant form for $\rho_0(t)$ is specified for all subjects. We will consider an extension of the parametric SCCS in Section 2.3.2, where age-specific baseline rate function is allowed to have various constant rates over certain age intervals. We also consider only one time varying exposure $x_i(t)$ during the followup.

In this simple setting, we let $\psi(t) = 1$ in (2.21) so that the parameter η specifies a constant underlying incidence rate for all subjects over their followup periods. The SCCS design is based on the fact that subjects are sampled bacause they experienced the event of interest at least one time during the followup. The SCCS method is based on a conditional likelihood function to reflect this fact. In this setting, the conditional probability of the outcome that events observed at times $t_{i1} < \cdots < t_{in_i}$ in the time interval $[\tau_{0i}, \tau_i]$, given that $N_i(\tau_{0i}, \tau_i) = n_i$, $i = 1, \ldots, N$, (with a little abuse of notation) is

$$\Pr\{t_{i1}, \ldots, t_{in_i}, [\tau_{0i}, \tau_i] \mid N_i(\tau_{0i}, \tau_i) = n_i\} = \frac{\Pr\{t_{i1}, \ldots, t_{in_i}, [\tau_{0i}, \tau_i], N_i(\tau_{0i}, \tau_i) = n_i\}}{\Pr\{N_i(\tau_{0i}, \tau_i) = n_i\}}. \tag{2.22}$$

Notice that the numerator in (2.22) is the outcome of a recurrent event process con-

sidered in the likelihood function (2.5). Therefore, from the likelihood function (2.5) with the intensity function (2.21) and $\psi(t) = 1$, we can write the probability in the numerator in (2.22) as

$$L_i(\theta_i^*) = \left[\prod_{j=1}^{n_i} \eta\, e^{\gamma_i + x_i(t_{ij})\beta}\right] \exp\left\{-\int_{\tau_{0i}}^{\tau_i} \eta\, e^{\gamma_i + x_i(s)\beta}\, ds\right\}, \tag{2.23}$$

where $\theta_i^* = (\eta, \gamma_i, \beta)'$, $i = 1, \ldots, N$. The SCCS model (2.21) with $\psi(t) = 1$ is a Poisson process model. Thus, for $i = 1, \ldots, N$,

$$\Pr\{N_i(\tau_{0i}, \tau_i) = n_i\} = [\mu_i(\tau_{0i}, \tau_i)]^{n_i}\, \frac{e^{-\mu_i(\tau_{0i}, \tau_i)}}{n_i!}, \qquad n_i = 0, 1, 2, \ldots, \tag{2.24}$$

where $\mu_i(\tau_{0i}, \tau_i) = \int_{\tau_{0i}}^{\tau_i} \eta\, e^{\gamma_i + x_i(s)\,\beta}\, ds$. By replacing (2.23) and (2.24) with the numerator and denominator in (2.22), respectively, the conditional likelihood for the ith subject, $i = 1, \ldots, N$, can be written as

$$L_{ic}(\beta) = \left[\prod_{j=1}^{n_i} \exp[\, x_i(t_{ij})\beta]\right] \left\{\int_{\tau_{0i}}^{\tau_i} \exp[\, x_i(s)\beta]\, ds\right\}^{-n_i}. \tag{2.25}$$

There are two important remarks to be made about the conditional likelihood function (2.25). First, there is no fixed covariates (i.e., neither η nor γ_i) left to be estimated. This fact shows that there is no need to collect values of fixed covariates such as gender, income level and genetic information on subjects in the SCCS design. Second, if a subject did not experience the event of interest over the observation window (i.e., when $n_i = 0$), the conditional likelihood (2.25) becomes one so that there is no contribution of controls to the likelihood of the SCCS design. This means that there is no need to sample controls, and thus, the SCCS design is solely based on the cases. Therefore, the likelihood function of the SCCS can be given only by considering cases, instead of all subject in the cohort, as follows.

Suppose that, among N subjects, only m of them experience the event of interest over their followup and $N - m$ of them do not experience the event of interest. In this case, we say there are m cases and $N - m$ controls in the cohort. Therefore, for m cases, the conditional likelihood of the outcome that events observed at times $t_{i1} < \cdots < t_{in_i}$ in the time interval $[\tau_{0i}, \tau_i]$, given that $N_i(\tau_{0i}, \tau_i) = n_i$, $i = 1, \ldots, m$, is

$$L_c(\beta) = \prod_{i=1}^{m} L_{ic}(\beta), \tag{2.26}$$

where $L_{ic}(\beta)$ is defined in (2.25). The log likelihood function $\ell_c(\beta) = \log L_c(\beta)$ is

$$\ell_c(\beta) = \sum_{i=1}^{m} \left\{ \beta \sum_{j=1}^{n_i} x_i(t_{ij}) - n_i \log \left(\int_{\tau_{0i}}^{\tau_i} \exp\left[x_i(s)\,\beta\right]\, ds \right) \right\}. \tag{2.27}$$

Let $U_c(\beta) = (\partial/\partial\beta)\ell_c(\beta)$ be the score function. Then, under mild regularity conditions,

$$U_c(\beta) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} x_i(t_{ij}) - \sum_{i=1}^{m} n_i \frac{\int_{\tau_{0i}}^{\tau_i} x_i(s)\, \exp[x_i(s)\beta]\, ds}{\int_{\tau_{0i}}^{\tau_i} \exp[x_i(s)\beta]\, ds}. \tag{2.28}$$

The observed information function $I_c(\beta) = -(\partial^2/\partial\beta^2)\ell_c(\beta) = -(\partial/\partial\beta)U_c(\beta)$ is given by

$$I_c(\beta) = \sum_{i=1}^{m} n_i \left\{ \frac{\int_{\tau_{0i}}^{\tau_i} x_i(s)\, e^{x_i(s)\beta}\, ds \int_{\tau_{0i}}^{\tau_i} e^{x_i(s)\beta}\, ds - \left[\int_{\tau_{0i}}^{\tau_i} x_i(s)\, e^{x_i(s)\,\beta}\, ds\right]^2}{\left[\int_{\tau_{0i}}^{\tau_i} e^{x_i(s)\,\beta}\, ds\right]^2} \right\}. \tag{2.29}$$

The maximum likelihood estimator $\hat{\beta}$ of β can be obtained by setting $U_c(\beta)$ in (2.28) to 0 and solving for β. Notice that since $x_i(t)$ takes values of 0 and 1 over $[\tau_{0i}, \tau_i]$, we have

$$\int_{\tau_{0i}}^{\tau_i} e^{x_i(s)\,\beta}\, ds \geq \int_{\tau_{0i}}^{\tau_i} x_i(s)\, e^{x_i(s)\,\beta}\, ds. \tag{2.30}$$

It is, therefore, easy to see that the observed information function $I_c(\beta)$ is negative, and thus the conditional log likelihood function $\ell_c(\beta)$ in (2.27) is convex and $\hat{\beta}$ is the

unique maximizer of it.

Optimization software can be used to maximize $U_c(\beta)$ to obtain the maximum likelihood estimator $\hat{\beta}$. To this end, we used in this thesis the `nlm` function in R, which also produces the value of $-I_c(\beta)$ evaluated at $\hat{\beta}$. As the number of cases increases, i.e. as $m \to \infty$, it can be shown that $(\hat{\beta} - \beta)$ converges in distribution to a normal distribution with mean 0 and variance $I_c^{-1}(\hat{\beta})$ (Cook and Lawless, 2007, Section 3.2).

2.2.1.1 Simulation Study: Comparison of Parametric SCCS Model with Parametric Cohort Model

In order to investigate the estimation of β under parametric models, we conducted a Monte Carlo simulation study. We considered parametric cohort model with the intensity function $\eta_1 \exp(x_i(t)\,\beta_1)$ and the parametric SCCS model with the intensity function $\eta_2 \exp(x_i(t)\,\beta_2)$.

We generated N realizations of a nonhomogeneous Poisson process with the intensity function $\alpha \exp(x_i(t)\,\beta)$ for 1,000 simulation runs. Furthermore, we considered a fixed observation window $[0, \tau = 1000]$ for all N processes. For a given iteration of the Monte Carlo simulation, we simulated the event times t_{ij} and the start times of exposure periods e_i. The event times t_{ij}, $i = 1, \ldots, N$ and $j = 1, \ldots, n_i$, are generated by the algorithm given in the previous chapter. To generate the start times of exposure periods e_i, $i = 1, \ldots, N$, we first decided whether a subject was exposed to the external condition from a Bernoulli distribution with exposure probability $p = 0.8$. Therefore, there was a high rate of exposed subjects in the population. For a subject with exposure, we generated his start of the exposure period e_i from a Uniform$(0, \tau)$ distribution.

Factors of the Monte Carlo simulations were the cohort size N, the risk period Δ, $E\{N_i(\tau)\}$ when $\beta = 0$, and $E\{N_i(\Delta)\}$. Notice that, for $i = 1, \ldots, N$, $E\{N_i(\tau)\} =$

$\alpha\,\tau$ which gives the expected number of events for subjects who were not exposed over $[0, \tau]$. Also, for $i = 1, \ldots, N$, $E\{N_i(e_i, e_i + \Delta)\} = \alpha\, e^{\beta}\Delta$ which gives the expected number of events within the risk period Δ. By notational convenience, we let $E\{N_i(\Delta)\} = E\{N_i(e_i, e_i+\Delta)\}$. Therefore, the scenarios included the combinations of $(N, \Delta, E\{N_i(\tau)\}, E\{N_i(\Delta)\})$, where $N = 100$, 500, 1000, $\Delta = 10$, 20 and 50 days, $E\{N_i(\tau)\} = 1$, 2 and 5, and $E\{N_i(\Delta)\} = 1$, 2, and 5. It should be noted that the expected number of events for an exposed subject is given by $\alpha\left(\Delta e^{\beta} + \tau - \Delta\right)$.

For each Monte Carlo iteration b, $b = 1, \ldots, B = 1000$, and under each simulation scenario, we generated the data and obtained the estimates $\hat{\beta}_{b1}$ and $\hat{\beta}_{b2}$ of β_1 and β_2, respectively. We reported $\bar{\hat{\beta}}_k = \frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_{bk}$, $k = 1, 2$. Since the sample sizes of the cohort design were significantly larger than the sample sizes of the SCCS design, the standard errors of $\hat{\beta}_1$ were significantly smaller than those of $\hat{\beta}_2$. Therefore, we only reported the estimates of the parameters β_1 and β_2.

The results are presented in Table 2.1. The true value of β can be calculated from the equations $E\{N_i(\tau)\} = \alpha\tau$ and $E\{N_i(e_i, e_i + \Delta)\} = \alpha e^{\beta}\Delta$, and is given for each simulation scenario in Table 2.1. Overall, the estimates of β are close under both designs, which indicates that most of the information about β is obtained from the cases. This result encourages the use of the SCCS model as it depends only on the cases. When Δ is large and $E\{N_i(\tau)\}$ are small, there is a small bias in $\bar{\hat{\beta}}_1$. For example, when $(N, \Delta, E\{N_i(\tau)\}, E\{N_i(\Delta)\}) = (100, 50, 1, 1)$, $\bar{\hat{\beta}}_1 = 2.668$ while $\beta = 2.995$. This bias remains even when N gets larger, but decreases when $E\{N_i(\tau)\}$ or $E\{N_i(\Delta)\}$ increases. The estimate based on the SCCS model does not have a significant bias in all simulation scenarios considered in Table 2.1.

Table 2.1: Averages of the estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ of β based on the cohort and SCCS models, respectively, are given under various combinations of $(N, \Delta, E\{N_i(\tau)\}, E\{N_i(\Delta)\})$.

				$N = 100$		$N = 500$		$N = 1000$	
Δ	$E\{N_i(\tau)\}$	$E\{N_i(\Delta)\}$	β	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
10	1	1	4.605	4.602	4.609	4.603	4.607	4.599	4.604
10	1	2	5.298	5.297	5.299	5.295	5.296	5.288	5.300
10	1	5	6.214	6.208	6.219	6.213	6.214	6.214	6.213
10	2	1	3.912	3.895	3.912	3.906	3.912	3.896	3.911
10	2	2	4.605	4.601	4.605	4.600	4.605	4.600	4.605
10	2	5	5.521	5.524	5.523	5.520	5.522	5.521	5.521
10	5	1	2.995	2.996	2.985	3.000	2.996	2.998	2.996
10	5	2	3.688	3.686	3.684	3.688	3.688	3.689	3.689
10	5	5	4.605	4.610	4.608	4.603	4.606	4.604	4.606
20	1	1	3.912	3.882	3.918	3.892	3.913	3.894	3.913
20	1	2	4.605	4.591	4.603	4.591	4.603	4.595	4.606
20	1	5	5.521	5.519	5.522	5.519	5.522	5.506	5.520
20	2	1	3.218	3.189	3.218	3.189	3.218	3.187	3.218
20	2	2	3.912	3.895	3.912	3.895	3.912	3.897	3.911
20	2	5	4.828	4.819	4.829	4.819	4.829	4.827	4.827
20	5	1	2.302	2.339	2.301	2.339	2.301	2.317	2.304
20	5	2	2.995	2.977	2.995	2.977	2.995	2.985	2.994
20	5	5	3.912	3.896	3.910	3.896	3.910	3.904	3.911
50	1	1	2.995	2.668	2.996	2.668	2.996	2.644	2.994
50	1	2	3.688	3.509	3.686	3.509	3.686	3.493	3.687
50	1	5	4.605	4.604	4.607	4.604	4.607	4.607	4.603
50	2	1	2.302	2.558	2.304	2.558	2.304	2.599	2.300
50	2	2	2.995	2.838	2.995	2.838	2.995	2.775	2.994
50	2	5	3.912	3.910	3.909	3.910	3.909	3.912	3.909
50	5	1	1.386	1.384	1.383	1.384	1.383	1.385	1.383
50	5	2	2.079	2.078	2.077	2.078	2.077	2.078	2.075
50	5	5	2.995	2.994	2.995	2.994	2.995	2.995	2.993

2.2.2 The Semiparametric SCCS Method

In some studies, baseline rate function $\rho_0(t)$ in the intensity function (2.21) may depend on age. In such studies, the parametric SCCS model considered in the previous section could be restrictive. In the next section, we discuss how to parametrically generalize the model explained in the previous section to deal with this issue. Another approach to inference is to develop a semiparametric model, which left the baseline rate function parametrically unspecified. To this end, Farrington and Whitaker (2006) introduced the semiparametric analysis of the SCCS design, which we discuss in this section.

Similar to the Andersen-Gill model discussed in Section (2.1.2), Farrington and Whitaker (2006) consider an unspecified baseline rate function. However, since the semiparametric method is developed for the SCCS design, it should be based on the conditional likelihood function. Following the setup given in the previous section, the likelihood function in the SCCS design for m cases with intensity function (2.21) is given by

$$L_c = \prod_{i=1}^{m} L_{ic} = \prod_{i=1}^{m} \prod_{j=1}^{n_i} \frac{\psi(t_{ij})\, e^{x_i(t_{ij})\,\beta}}{\int_{\tau_{0i}}^{\tau_i} e^{x_i(s)\,\beta}\, d\Psi(s)}, \tag{2.31}$$

where $\Psi(s) = \int_{\tau_0}^{s} \psi(t)\, dt$ and $\tau_0 = \min(\tau_{01}, \ldots, \tau_{0m})$. Since the interest is focused on the parameter β (or equivalently on e^{β}, the relative incidence associated with exposure), the age-specific relative incidence $\psi(t)$ can be considered as a nuisance parameter in (2.31).

Farrington and Whitaker (2006) introduce a setup where $\psi(t)$ is left parametrically unspecified. Let ζ be the set of all distinct event times t_{ij} ($i = 1, \ldots, m$; $j = 1, \ldots, n_i$) for all m cases. Suppose that there are M distinct event times denoted by $s_1, \ldots, s_M$ in an incerasing order. The nonparametric maximum likelihood estimator of $\Psi(t) = \int_{\tau_0}^{t} \psi(s)\, ds$ should be a non-decreasing and positive-valued step function.

Let $\Psi(t)$ have jumps with heights $\Delta\Psi(t)$, $t \in \zeta$. Notice that $\prod_{j=1}^{n_i} \psi(t_{ij}) \exp[x_i(t_{ij})\beta]$ is equal to $\prod_{j=1}^{n_i} \Delta\Psi(t_{ij}) \exp[x_i(t_{ij})\beta]$, $i = 1, \ldots, m$. Therefore, from the likelihood function (2.31), the semiparametric likelihood function of the SCCS design is given by

$$L_{SP}(\beta, \Psi(t)) = \prod_{i=1}^{m} \prod_{j=1}^{n_i} \frac{\Delta\Psi(t_{ij})\, e^{x_i(t_{ij})\beta}}{\int_{\tau_{0i}}^{\tau_i} e^{x_i(s)\beta}\, d\Psi(s)}. \tag{2.32}$$

The maximum likelihood estimators $\hat{\beta}$ and $\hat{\Psi}(t)$ can be obtained by maximizing the semiparametric likelihood function (2.32). For this purpose, Farrington and Whitaker (2006) define the jump heights at distinct times s_r as $\Delta\Psi(s_1) = 1$ and $\Delta\Psi(s_r) = \exp(\alpha_r)$, $r = 2, \ldots, M$. Also, let $w_{ir} = I(\tau_{0i} < s_r \leq \tau_i)$ for each individual i and each $s_r \in \zeta$. Then, the semiparametric likelihood function (2.32) can be rewritten as

$$L_{SP}(\beta, \Psi(t)) = \prod_{i=1}^{m} \prod_{j=1}^{n_i} \frac{e^{\alpha_{ij} + x_i(t_{ij})\beta}}{\sum_{r=1}^{M} w_{ir}\, e^{\alpha_r + x_i(s_r)\beta}}, \tag{2.33}$$

where $\alpha_{ij} = \sum_{r=1}^{M} I(s_r = t_{ij})\, \alpha_r$; that is, the value of α_r corresponding to t_{ij}. From the likelihood function (2.33), the log likelihood function $\ell_{SP}(\beta, \Psi(t)) = \log L_{SP}(\beta, \Psi(t))$ is given by

$$\ell_{SP}(\beta, \Psi(t)) = \alpha_{\cdot} + \beta\, x_{\cdot} - \sum_{i=1}^{m} \sum_{i=1}^{n_i} \log \left[\sum_{r=1}^{M} w_{ir}\, e^{\alpha_r + x_i(s_r)\beta} \right], \tag{2.34}$$

where $\alpha_{\cdot} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \alpha_{ij}$ and $x_{\cdot} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} x_i(t_{ij})$. Let $U_\alpha = (U_{\alpha_2}, \ldots, U_{\alpha_M})'$ be an $(M-1)$-dimensional score vector with components $U_{\alpha_r} = (\partial/\partial\alpha_r)\ell_{SP}(\beta, \Psi(t))$, $r = 2$, $\ldots$, M, which gives

$$U_{\alpha_r} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} I(s_r = t_{ij}) - \sum_{i=1}^{m} \sum_{j=1}^{n_i} \frac{w_{ir}\, e^{\alpha_r + x_i(s_r)\beta}}{\sum_{r=1}^{M} w_{ir}\, e^{\alpha_r + x_i(s_r)\beta}}. \tag{2.35}$$

Let $U_\beta = (\partial/\partial\beta)\ell_{SP}(\beta, \Psi(t))$ be the score function for β, which gives

$$U_\beta = x_\cdot - \frac{\sum_{r=1}^{M} w_{ir}\, x_i(s_r)\, e^{\alpha_r + x_i(s_r)\beta}}{\sum_{r=1}^{M} w_{ir}\, e^{\alpha_r + x_i(s_r)\beta}}. \tag{2.36}$$

The components of the $M \times M$ observed information matrix $I(\beta, \Psi)$ can be obtained by taking the negative of the derivatives of U_α and U_β with respect to α_r, $r = 2, \ldots, M$, and β. The maximum likelihood estimators $\hat{\beta}$ and $\hat{\Psi}$ can be obtained by solving $U_\alpha = 0$, where 0 is an $(M-1)$-dimensional vector of zeros, and $U_\beta = 0$ for $\alpha_2, \ldots, \alpha_M$ and β, simultaneously. This can be done with an optimization software. In this thesis, we used the `nlm` function in R for this purposes. The `nlm` function produces the values of the Hessian matrix $-I(\beta, \Psi)$ evaluated at $\hat{\beta}$ and $\hat{\Psi}$ by numerical differentiation. We therefore obtained the $M \times M$ matrix $I(\hat{\beta}, \hat{\Psi})$ via the `nlm` function in R. The primary interest in the SCCS method is to make inference on β. Under some regularity conditions, Farrington and Whitaker (2006) showed that the maximum likelihood estimators $\hat{\beta}$ and $\hat{\Psi}$ are consistent estimators of β and $\hat{\Psi}$, respectively. Also, as $m \to \infty$, $\sqrt{m}(\hat{\beta} - \beta)$ converges to a normal distribution with mean 0 and efficient variance.

Computational efficiency can be an important issue with fitting the semiparametric SCCS model when the number of cases m is large. In particular, there are M parameters to be estimated in the seiparametric SCCS model. Therefore, if the parameter vector to be estimated is high dimensional, the nonparametric maximum likelihood procedure explained in this section can be computationally demanding. This issue is also discussed by Farrington and Whitaker (2006). An alternative method of model fitting is to use weakly parametric SCCS models. This method is fully parametric in theory. However, as the number of parameters increases, the estimate of β becomes close to $\hat{\beta}$ obtained from the semiparametric SCCS model. We discuss this method

in the next section.

2.3 Analysis of Recurrent Event Data Using Piecewise-Constant Rate Functions

Models based on Poisson processes can be made more flexible by replacing the baseline rate function $\rho_0(t;\alpha)$ in (2.7) with piecewise-constant rate functions. Computational efficiency is an important advantage of such models comparing with the semiparametric models. In particular, if the event of interest is not rare, the semiparametric models discussed in the previous sections become computationally demanding. In such settings, piecewise-contant rate models are important alternatives to the semiparametric models. In this section, we therefore discuss modeling Poisson processes as well as the SCCS model with piecewise-constant rate functions. At the end, we also present the results of a simulation study. There is a vast literature in piecewise modeling, but our approach in this section is based on Farrington (1995) and Cook and Lawless (2007, Section 3.3).

2.3.1 Piecewise-Constant Rate Models for Poisson Processes

Piecewise models for Poisson processes employs constant rate functions over prespecified time intervals to specify the baseline rate functions of Possion models. Following the setting given in Section 2.1.1, we consider a counting process $\{N_i(t); t \geq 0\}$, $i = 1$, ..., N, observed over $(\tau_{0i}, \tau_i]$ continuously with the intensity function

$$\rho_0(t;\alpha)\, e^{x_i'(t)\beta}, \tag{2.37}$$

where $\rho_0(t;\alpha)$ is the baseline rate function parametrically specified as explained below, $\beta = (\beta_1, \ldots, \beta_p)'$ is a vecor of parameters and $x_i(t) = (x_{i1}(t), \ldots, x_{ip}(t))'$ is a $p \times 1$ vector of covariates.

The baseline rate function $\rho_0(t;\alpha)$ in (2.37) can be specified as follows. Suppose we partition the time interval $(\tau_0, \tau]$, where $\tau_0 = \min(\tau_{01}, \ldots, \tau_{0N})$ and $\tau = \max(\tau_1, \ldots, \tau_N)$, into K pieces with $\tau_0 = a_0 < a_1 < \cdots < a_K = \tau$. denote K cutpoints such that $a_0 = 0$ and $a_K = \tau$. The baseline rate function is then given by

$$\rho_0(t;\alpha) = \begin{cases} \alpha_k, & \text{if } a_{k-1} < t \leq a_k, \\ 0, & \text{otherwise,} \end{cases} \tag{2.38}$$

where $\alpha = (\alpha_1, \ldots, \alpha_K)'$ is a K-dimensional vector of parameters which specifies the baseline rate function in the intensity function (2.37). It should be noted that the model (2.37) can be made more flexible by including more parameters in the baseline rate function (2.38). However, the more flexibility in the model, the more parameters to be estimated. Therefore, specification of the number of pieceses K is an important issue in piecewise modeling.

The likelihood function (2.6) can be used for inference purposes. Let the indicator function $w_k(t) = I(a_{k-1} < t \leq a_k)$ indicate whether $t \in (a_{k-1}, a_k]$, $k = 1, \ldots, K$. Suppose that $t_{i1} < t_{i2} < \cdots < t_{in_i}$ are the event times over $(\tau_{0i}, \tau_i]$ for the ith subject, $i = 1, \ldots, N$. Also, let $n_{ik} = \sum_{j=1}^{n_i} w_k(t_{ij})$ denote the number of events experienced by the ith subject experienced in time interval $(a_{k-1}, a_k]$, and $n_{\cdot k} = \sum_{i=1}^{N} n_{ik}$ denote the total number of events experienced by all subjects over $(a_{k-1}, a_k]$, $k = 1, \ldots, K$. Then, the likelihood function (2.6) with the intensity function (2.37) can be rewritten

in terms of events over time intervals $(a_{k-1}, a_k]$, $k = 1, \ldots, K$, as

$$L(\theta) = \prod_{k=1}^{K} \left\{ \left[\prod_{i=1}^{N} \alpha_k^{n_{ik}} e^{\sum_{j=1}^{n_i} x_i'(t_{ij})\,\beta} \right] \exp\left(-\alpha_k \sum_{i=1}^{N} \int_{a_{k-1}}^{a_k} Y_i(s)\, e^{(x_i'(s)\beta}\, ds \right) \right\}, \qquad (2.39)$$

where $\theta = (\alpha', \beta')'$ is a $(K + p) \times 1$ vector of parameters, and $Y_i(s)$ is the at risk indicator of the ith subject, $i = 1, \ldots, N$. The log likelihood function $\ell(\theta) = \log L(\theta)$ is given by

$$\ell(\theta) = \sum_{k=1}^{K} n_{.k}\, \log \alpha_k + \sum_{i=1}^{N} \sum_{j=1}^{n_i} x_i'(t_{ij})\, \beta - \sum_{i=1}^{N} \sum_{k=1}^{K} \alpha_k\, S_{ik}(\beta), \qquad (2.40)$$

where $S_{ik}(\beta) = \int_{a_{k-1}}^{a_k} Y_i(s) e^{x_i'(s)\,\beta}\, ds$. The score function $U_{\alpha_k}(\theta) = (\partial/\partial\alpha_k)\ell(\theta)$, $k = 1$, $\ldots$, K, is given by

$$U_{\alpha_k}(\theta) = \frac{n_{.k}}{\alpha_k} - \sum_{i=1}^{N} S_{ik}(\beta). \qquad (2.41)$$

Letting $U_{\alpha_k}(\theta) = 0$ gives the estimator $\tilde{\alpha}_k(\beta)$ of α_k for a fixed β as follows.

$$\tilde{\alpha}_k(\beta) = \frac{n_{.k}}{\sum_{i=1}^{N} S_{ik}(\beta)}, \qquad k = 1, \ldots, K. \qquad (2.42)$$

The $p \times 1$ score vector $U_\beta(\theta) = (U_{\beta_1}(\theta), \ldots, U_{\beta_p}(\theta))'$ with components $U_{\beta_r}(\theta) = (\partial/\partial\beta_r)\ell(\theta)$, $r = 1$, $\ldots$, p, is given by

$$U_{\beta_r}(\theta) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} x_{ir}(t_{ij}) - \sum_{i=1}^{N} \sum_{k=1}^{K} \alpha_k \int_{a_{k-1}}^{a_k} Y_i(s)\, x_{ir}(s)\, e^{x_i'(s)\,\beta}\, ds. \qquad (2.43)$$

We can insert $\tilde{\alpha}_k(\beta)$ given in (2.42) into the score function (2.43) in place of α_k and obtain the profile score function for β_r. In this context, the function $L(\tilde{\alpha}(\beta), \beta)$, where $\tilde{\alpha}(\beta) = (\tilde{\alpha}_1(\beta), \ldots, \tilde{\alpha}_K(\beta))'$, is called the profile likelihood function for β and corresponding profile log likelihood function for β is given by $\ell(\tilde{\alpha}(\beta), \beta) = \log L(\tilde{\alpha}(\beta), \beta)$. In our setting, it is possible to obtain the expilicit formula of the profile log likelihood

function for β, which is given by Cook and Lawless (2007). Notice that, by replacing the $\tilde{\alpha}_k(\beta)$ with α_k in (2.40), we can write

$$\ell(\tilde{\alpha}(\beta), \beta) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \left\{ x_i'(t_{ij})\, \beta - \sum_{k=1}^{K} w_k(t_{ij}) \sum_{l=1}^{N} \int_{a_{k-1}}^{a_k} Y_l(s)\, e^{x_l'(s)\,\beta}\, ds \right\}. \tag{2.44}$$

The $p \times 1$ profile score vector for β is then given by (Cook and Lawless, 2007)

$$\frac{\partial \ell(\tilde{\alpha}(\theta), \beta)}{\beta} = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \left\{ x_i(t_{ij}) - \frac{\sum_{k=1}^{K} w_k(t_{ij}) \int_{a_{k-1}}^{a_k} \sum_{l=1}^{N} Y_l(s)\, e^{x_l'(s)\,\beta}\, x_l(s)\, ds}{\sum_{k=1}^{K} w_k(t_{ij}) \int_{a_{k-1}}^{a_k} \sum_{l=1}^{m} Y_l(s)\, e^{x_l'(s)\,\beta}\, ds} \right\}. \tag{2.45}$$

The maximum likelihood estimator $\hat{\beta}$ can be obtained by maximizing the profile score function (2.45). The maximum likelihood estimator $\hat{\alpha}_k$, $k = 1, \ldots, K$, is then given by inserting $\hat{\beta}$ into $\tilde{\alpha}_k(\beta)$ in (2.42). The observed information matrix $I(\hat{\theta})$ can be inverted to obtain the estimates of the variance of $\hat{\theta} = (\hat{\alpha}', \hat{\beta}')'$. In this study, we used the `nlm` function in R to obtain the estimates $\hat{\theta}$ and their standard errors. For this purpose, the log likelihood function (2.40) was maximized with the `nlm` function, which also produced the Hessian matrix.

2.3.2 Piecewise-Constant Rate Models for the SCCS Design

The parametric self-controlled case series (SCCS) model introduced in Section 2.2.1 can be extended by applying the piecewise-constant baseline approach. In this setting, varying age effects can be included in the baseline rate function by specifying constant baseline functions for various age groups.

Similar to the setup given in Section 2.2.1, we consider a cohort of N independent subjects who are at risk of being exposed to an external condition for a time period called risk period and denoted by Δ. Among N subjects, we let m denote the numbers of subjects with at least one event(i.e. cases). Subjects are observed over the

observation window $(\tau_{0i}, \tau_i]$ and each of them have n_i events, $i = 1, \ldots, N$. From (2.21), the SCCS model is given by

$$\eta\, \psi(t)\, \exp\{\gamma_i + x_i(t)\beta\}, \qquad i = 1, \ldots, N, \tag{2.46}$$

where η is the age effect at the start of the followup, $\psi(t)$ is the age-specific relative incidence, γ_i represents all fixed covariates and random effects, and the external covariate $x_i(t)$ denotes the time-varying exposure that is experienced at age t.

In this section, we specify the age-specific relative incidence function $\psi(t)$ parametrically as follows. Suppose that there are K age intervals. Let $\psi(t) = \exp\{g(t;\alpha)\}$ where $g(t;\alpha)$ is a linear step function defined with parameters $\alpha = (\alpha_1, \ldots, \alpha_K)'$ and the number of age intervals K (Farrington, 1995). That is, for $k = 1, \ldots, K$,

$$\psi(t;\alpha) = \begin{cases} e^{\alpha_k} & \text{if } t \in (t_{k-1}, t_k], \\ 0 & \text{otherwise.} \end{cases} \tag{2.47}$$

The numerator of the conditional probability in (2.22) with the SCCS model (2.46) and (2.47) is given by

$$L_i(\theta_i) = \left[\prod_{j=1}^{n_i} \eta\, \psi(t_{ij};\alpha)\, e^{\gamma_i + x_i(t_{ij})\beta}\right] \exp\left\{-\int_{\tau_{0i}}^{\tau_i} \eta\, \psi(s;\alpha)\, e^{\gamma_i + x_i(s)\beta}\, ds\right\}, \tag{2.48}$$

where $\theta_i = (\eta, \alpha', \beta)'$. Also, the denominator of (2.22) is given by

$$\Pr\{N_i(\tau_{0i}, \tau_i) = n_i\} = [\mu_i(\tau_{0i}, \tau_i)]^{n_i}\, \frac{e^{-\mu_i(\tau_{0i}, \tau_i)}}{n_i!}, \qquad n_i = 0, 1, 2, \ldots, \tag{2.49}$$

where $\mu_i(\tau_{0i}, \tau_i) = \int_{\tau_{0i}}^{\tau_i} \eta\, \psi(s;\alpha)\, e^{\gamma_i + x_i(s)\beta}\, ds$.

Once again, by replacing (2.48) and (2.49) with the numerator and denominator in

(2.22), respectively, we obtain the conditional likelihood for the ith subject, $i = 1$, ..., N, which is given by

$$L_{ic}(\theta) = \left[\prod_{j=1}^{n_i} \psi(t_{ij};\alpha)\, \exp[\, x_i(t_{ij})\beta]\right] \left\{\int_{\tau_{0i}}^{\tau_i} \psi(s;\alpha)\, \exp[\, x_i(s)\beta]\, ds\right\}^{-n_i}, \qquad (2.50)$$

where $\theta = (\alpha', \beta)'$. Similar to the result showed in Section 2.2.1, when $n_i = 0$, the conditional likelihood (2.50) is equal to one so that there is no contribution of controls to the likelihood of the SCCS design. Therefore, the SCCS design is solely based on the cases, and once again, the likelihood function of the SCCS can be written only with cases.

In this case, the conditional likelihood of the outcome that events observed at times $t_{i1} < \cdots < t_{in_i}$ in the time interval $[\tau_{0i}, \tau_i]$, given that $N_i(\tau_{0i}, \tau_i) = n_i$, $i = 1, \ldots, m$, is

$$L_c(\theta) = \prod_{i=1}^{m} L_{ic}(\theta), \qquad (2.51)$$

where

$$L_{ic}(\theta) = \prod_{j=1}^{n_i} \frac{\psi(t_{ij};\alpha)\, e^{x_i(t_{ij})\,\beta}}{\int_{\tau_{0i}}^{\tau_i} \psi(s;\alpha)\, e^{x_i(s)\,\beta}\, ds}. \qquad (2.52)$$

The log likelihood function $\ell_c(\theta) = \log L_c(\theta)$ is

$$\ell_c(\theta) = \sum_{i=1}^{m} \left\{\sum_{j=1}^{n_i} \log \psi(t_{ij};\alpha) + \beta \sum_{j=1}^{n_i} x_i(t_{ij}) - n_i\, \log\left(\int_{\tau_{0i}}^{\tau_i} \psi(s;\alpha)\, e^{x_i(s)\,\beta}\, ds\right)\right\}. \qquad (2.53)$$

Let $U_{\alpha_k}(\theta) = (\partial/\partial\alpha_k)\ell_c(\theta)$, $k = 1$, ..., K, and $U_\beta(\beta) = (\partial/\partial\alpha)\ell_c(\theta)$, we have $(K + 1) \times 1$ score vector $U(\theta) = (U_{\alpha_1}, \ldots, U_{\alpha_K}, U_\beta)'$. The maximum likelihood estimator $\hat{\theta} = (\hat{\alpha}', \hat{\beta})'$ can be obtained by solving $U(\theta) = 0$ for θ, where 0 is a $(K+1)$-dimensional vector of zeros. In this study, we used the `nlm` function in R to obtain $\hat{\theta}$ by maximizing the conditional log likelihood function (2.53). We also obtained the estimates of the

variance of $\hat{\beta}$ by obtaining the Hessian matrix evaluated at $\hat{\alpha}$ and $\hat{\beta}$ via `nlm`. As discussed by Farrington (1995), the asymptotic properties of $\hat{\beta}$ given in Section 2.2.1 hold in this setting as well.

2.4 Simulation Study

In this section, we present the result of our second Monte Carlo simulation study. Our goal is to investigate the bias and precision in the estimation of β (or similarly, the relative incidence rate) under semiparametric models and piecewise-constant baseline models.

We compared the following models:

M1: The Andersen-Gill model, introduced in Section 2.1.2, with the intensity function

$$\text{Model 1:} \qquad \rho_0(t)\, \exp(x_i(t)\, \beta). \tag{2.54}$$

M2: The semiparametric SCCS model, introduced in Section 2.2.2, with the intensity function

$$\text{Model 2:} \qquad \psi(t)\, \exp(x_i(t)\, \beta). \tag{2.55}$$

M3: The Poisson process with piecewise-constant baseline rate functions, introduced in Section 2.3.1, with the intensity function

$$\text{Model 3:} \qquad \rho_0(t; \alpha)\, \exp(x_i(t)\, \beta), \tag{2.56}$$

where $\alpha = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)'$ and

$$\rho_0(t;\alpha) = \begin{cases} \alpha_1, & \text{if } t \in (0, 125]; \\ \alpha_2, & \text{if } t \in (125, 250]; \\ \alpha_3, & \text{if } t \in (250, 375]; \\ \alpha_4, & \text{if } t \in (375, 500]; \end{cases} \tag{2.57}$$

M4: The SCCS model process with piecewise-constant baseline rate functions, introduced in Section 2.3.2, with the intensity function

$$\text{Model 4:} \qquad \psi(t;\alpha)\, \exp(x_i(t)\, \beta), \tag{2.58}$$

where $\alpha = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)'$ and

$$\psi(t;\alpha) = \begin{cases} \alpha_1, & \text{if } t \in (0, 125]; \\ \alpha_2, & \text{if } t \in (125, 250]; \\ \alpha_3, & \text{if } t \in (250, 375]; \\ \alpha_4, & \text{if } t \in (375, 500]; \end{cases} \tag{2.59}$$

In this study, we simulated a rare event setting with stochastic processes. The difficulties related to rare event simulation were discussed by Asmussen and Glynn (2007, Chapter VI). One of the difficulties we encountered was the computational efficiency. When we tried to increase the number of events experienced by subjects, Model 2 (the semiparametric SCCS model) was computationally demanding. Another issue was about the estimation of β with Model 3 and Model 4. In these cases, when we tried to increase the number of pieces in the baseline functions, we encountered

numerical issues because many pieces had zero counts. Because of these issues, we conducted a simulation study under limited scenarios. However, as discussed in the last chapter, we will investigate better simulation algorithms to extend the scenarios in this setting as a future work.

We generated $N = 5000$ realizations of a counting process $\{N_i(t); t \geq 0\}$ with the intensity function $\alpha \exp(x_i(t)\,\beta)$, $i = 1, \ldots, N$, over the observation window $[0, \tau = 500]$. Notice that, different than the simulation study in Section 2.2.2, we take $\tau = 500$ units in this section. The reason why we changed the value of τ is that Farrington and Whikter (2006) conducted a simulation study only for the Model 2 and their choice of τ was 500 units. To compare our results with theirs, we selected $\tau = 500$ as well. We fixed Δ at 50 days and the number of cases m at 50. For 1000 simulation runs, we generated event times t_{ij} by using the algorithm given in the previous chapter. We decided whether a subject was exposed or not by a Bernoulli distribution with success probability $p = 0.5$. If a subject was exposed, its start time of exposure periods were generated from a Uniform$(0, 500)$ distribution.

The relative incidence rate e^{β} was the factor of the Monte Carlo simulations. We considered $e^{\beta} = 6$, 8 and 10. For each Monte Carlo iteration b, $b = 1, \ldots, B = 1000$, we generated the data and obtained the estimate $\hat{\beta}_b$ and its standard error under Models 1, 2, 3 and 4. We reported $\bar{\hat{\beta}} = \frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_b$, the average of the standard error of $\hat{\beta}$, average bias $\bar{Bias}(\hat{\beta}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\beta}_b - \beta)$, relative bias $\bar{RB}(\hat{\beta}) = 100 \times \left|\frac{\bar{\hat{\beta}}-\beta}{\beta}\right|$ and the mean squared error $\bar{MSE}(\hat{\beta}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\beta}_b - \beta)^2$.

The results are presented in Table 2.2. In our simulation study, Model 1 (Andersen-Gill model) performed the best in terms of bias and precision. The bias however increases as β increases. The standard error of $\hat{\beta}$ is getting smaller on the average as β increases. The average bias induced by Model 2 (the semiparametric SCCS model) is the largest among all models, and the average bias increases as β increases. Also,

Table 2.2: Simulation results of Models 1, 2, 3 and 4. $\bar{\hat{\beta}}$ is the average estimates of β. $\bar{S}(\hat{\beta})$ is the average standard error. $\bar{Bias}(\hat{\beta})$ is the average bias. $\bar{RB}(\hat{\beta})$ and $\bar{MSE}(\hat{\beta})$ are relative bias and mean square error.

	e^{β}	β	$\bar{\hat{\beta}}$	$\bar{S}(\hat{\beta})$	$\bar{Bias}(\hat{\beta})$	$\bar{RB}(\hat{\beta})$	$\bar{MSE}(\hat{\beta})$
Model 1	6	1.791	1.697	0.590	-0.094	-5.269	0.306
	8	2.079	1.962	0.450	-0.117	-5.647	0.244
	10	2.302	2.162	0.396	-0.139	-6.075	0.178
Model 2	6	1.791	2.669	0.462	0.877	48.96	1.009
	8	2.079	3.003	0.438	0.923	44.43	1.043
	10	2.302	3.249	0.422	0.947	41.14	1.069
Model 3	6	1.791	2.521	0.325	0.730	40.75	2.298
	8	2.079	2.715	0.296	0.635	30.57	1.542
	10	2.302	2.904	0.272	0.601	26.12	1.070
Model 4	6	1.791	2.428	0.434	0.636	35.53	0.596
	8	2.079	2.675	0.397	0.595	28.65	0.508
	10	2.302	2.855	0.372	0.552	23.99	0.425

it produces the second biggest mean square error. It should be noted that Model 1 and Model 2 were the most computationally demanding models. Model 3 produces smaller average bias than Model 2. Also, Model 3 gives the smallest average standard error among the models, and the average standard error decreases as β increases. Model 4 shows smaller average bias than Model 2 and Model 3. However, the average standard error $\bar{S}(\hat{\beta})$ is larger than that of Model 3.

To sum up, in this limited simulation study, we showed that the Andersen-Gill model, which is based on the full cohort performed well in terms of bias and precision in the estimation of the relative incidence rate. However, it is computationally not efficient. The semiparametric SCCS model did not performed better than models based on piecewise-constant rate functions. However, it should be noted that, in this simulation study, we only consider a constant baseline rate function to generate the data. In the next chapter, we compare two computatonally effcient models; the semiparametric SCCS model (Model 2) and the SCCS model with piecewise-constant baseline rate

functions (Model 4) under various settings.

Chapter 3

A Simulation Study for the Comparison of Semiparametric SCCS model with SCCS model with Piecewise-Constant Baseline Rate Functions

In this chapter, we present the results of a simulation study to compare the estimate of the relative incidence rate based on semiparametric self-controlled case series (SCCS) model with that of SCCS model with piecewise-contant baseline rate functions. Our main objective is to investigate the performance of the flexible SCCS models when there is age effect in the baseline rate functions. For this purpose, we conduct an extensive simulation study, where we consider three different baseline functions with age dependency.

The outline of this chapter is as follows. In Section 3.1, we explain the data generating

process, and give the factors of the simulations. In Section 3.2, we present the results of the simulation study and give the conclusion of the simulation study.

3.1 Design of the Simulation Study

We introduc the design of the Monte Carlo simulations in this section. We consider two models under various scenarios. The models considered are

M1: The semiparametric SCCS model, introduced in Section 2.2.2, with the intensity function

$$\text{Model 1:} \qquad \psi_1(t)\, \exp(x_i(t)\, \beta_1). \tag{3.1}$$

M2: The SCCS model with piecewise-constant baseline rate functions, introduced in Section 2.3.2, with the intensity function

$$\text{Model 2:} \qquad \psi_2(t; \alpha)\, \exp(x_i(t)\, \beta_2), \tag{3.2}$$

where $\alpha = (\alpha_1, \ldots, \alpha_K)'$ and for $k = 1, \ldots, K$,

$$\psi_2(t; \alpha) = \alpha_k, \quad \text{if } t \in (a_{k-1}, a_k]. \tag{3.3}$$

It should be noted that both Model 1 and Model 2 are models for the SCCS design. As discussed in Section 2.2, we only need to sample the cases to fit them, and we can ignore the controls. In our setup, we considered $m = 25$ independent processes (i.e., subjects), where m denotes the number of cases. Each process is observed over the same observation window $[0, \tau = 500]$. Each subject has at least one event over $[0, 500]$; that is $N_i(500) = n_i \geq 1$, $i = 1, \ldots, m$.

For a given iteration of the Monte Carlo simulation, we simulate event times t_{ij}, $i = 1$, ..., m; $j = 1$, ..., n_i and start times of the exposure periods e_i, $i = 1$, ..., m. We assume that all subjects are exposed, and their start times of exposure (i.e., risk) periods e_i, $i = 1$, ..., m, are generated from a Uniform$(0, 500)$ distribution. It should be noted that Model 1 and Model 2 above are based on the SCCS design. Since the SCCS design is defined by conditioning on the observed values of n_i, $i = 1$, ..., N, the appropriate simulation procedure should be based on the conditionally on the values of $N_i(500)$ and $x_i(t)$, $i = 1$, ..., m. It can be shown that, for $i = 1$, ..., m, conditional on the values of $N_i(500) = n_i$, the event times T_{i1}, ..., T_{in_i} are distributed as the order statistics of a random smaple of size n_i from the truncated distribution with c.d.f. (Cox and Lewis, 1966)

$$F_i(t) = \frac{\int_0^t \lambda_i(u \,|\, n_i)\, du}{\int_0^{500} \lambda_i(u \,|\, n_i)\, du}, \tag{3.4}$$

where $\lambda_i(u \,|\, n_i)$ is the intensity function of the process.

According to the above discussion, we first generated a set of e_i, $i = 1$, ..., m, values as explained above. We kept the values of e_i fixed in all simualtion runs b, $b = 1$, ..., $B = 1000$. Then we decide the values of n_i. To keep the underlying rate of event small, we considere small values for n_i. To this end, we use the Bernoulli distribution with $p = 0.5$ to decide whether the subject i has one or two events; that is, $n_i = 1$ or 2, $i = 1$, ..., m. We fixe the values of n_i in all simulation runs. Then, we randomly allocate generated the event times t_{ij} by using the c.d.f. given in (3.4). A similar data-generating process is also applied by Farrington and Whitaker (2006). After the data generation, we fit Model 1 and Model 2 given above, and obtain the estimates $\hat{\beta}_{b1}$ and $\hat{\beta}_{b2}$, $b = 1$, ..., B, of β_1 in Model 1 and β_2 in Model 2, respectively. We repeated this procedure for $B = 1000$ times. As for the statistical analysis, we report

the average of the estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ and the average of their standard errors. We also report the average bias define by $\bar{Bias}(\hat{\beta}_j) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\beta}_{bj} - \beta)$, $j = 1, 2$.
We considere the above setup in three settings; setting A, setting B and setting C. In each seeting, we generated the data from the model

$$\rho_0(t;\alpha)\, e^{\beta\, x_i(t)}, \tag{3.5}$$

where $\rho_0(t)$ represents the age effect. The shape of the $\rho_0(t;\alpha)$ defined the settings. In these settings, we considered the following age effect

1. Setting A – Bell-shaped age effect group:

$$\begin{aligned}
\rho_0(t) = 1.00 \quad & if\ 0 \le t < 50 \quad && = 2.00 \quad if\ 250 \le t < 300 \\
= 1.25 \quad & if\ 50 \le t < 100 \quad && = 1.75 \quad if\ 300 \le t < 350 \\
= 1.50 \quad & if\ 100 \le t < 150 \quad && = 1.50 \quad if\ 350 \le t < 400 \\
= 1.75 \quad & if\ 150 \le t < 200 \quad && = 1.25 \quad if\ 400 \le t < 450 \\
= 2.00 \quad & if\ 200 \le t < 250 \quad && = 1.00 \quad if\ 450 \le t < 500
\end{aligned} \tag{3.6}$$

2. Setting B – Bathtup-shaped age effect group:

$$\begin{aligned}
\rho_0(t) = 2.00 \quad & if\ 0 \le t < 50 \quad && = 1.00 \quad if\ 250 \le t < 300 \\
= 1.75 \quad & if\ 50 \le t < 100 \quad && = 1.25 \quad if\ 300 \le t < 350 \\
= 1.50 \quad & if\ 100 \le t < 150 \quad && = 1.50 \quad if\ 350 \le t < 400 \\
= 1.25 \quad & if\ 150 \le t < 200 \quad && = 1.75 \quad if\ 400 \le t < 450 \\
= 1.00 \quad & if\ 200 \le t < 250 \quad && = 2.00 \quad if\ 450 \le t < 500
\end{aligned} \tag{3.7}$$

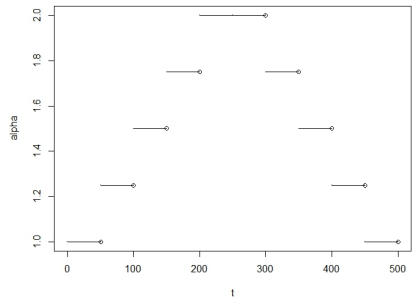

Figure 3.1: Age effect in Setting 1.

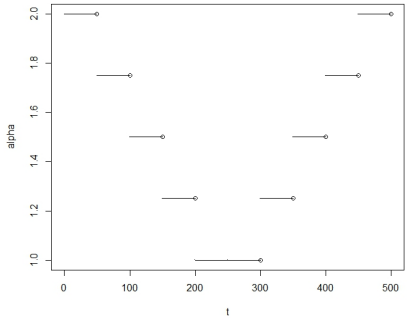

Figure 3.2: Age effect in Setting 2.

3. Setting C – Monotonically increasing age effect group:

$$\begin{aligned}
\rho_0(t) = 1.00 \quad & if\ 0 \le t < 50 \quad && = 3.50 \quad if\ 250 \le t < 300 \\
= 1.50 \quad & if\ 50 \le t < 100 \quad && = 4.00 \quad if\ 300 \le t < 350 \\
= 2.00 \quad & if\ 100 \le t < 150 \quad && = 4.50 \quad if\ 350 \le t < 400 \\
= 2.50 \quad & if\ 150 \le t < 200 \quad && = 5.00 \quad if\ 400 \le t < 450 \\
= 3.00 \quad & if\ 200 \le t < 250 \quad && = 5.50 \quad if\ 450 \le t < 500
\end{aligned} \tag{3.8}$$

The age effect is moderate in Setting A and Setting B, and strong in Setting C. The distribution of these age groups are given in Figures 3.1, 3.2 and 3.3, respectively.

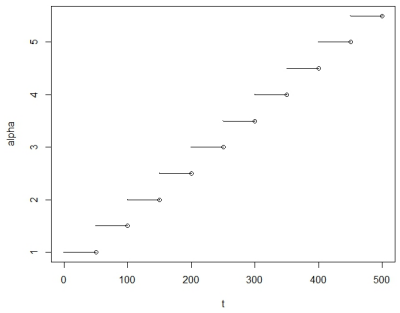

Figure 3.3: Age effect in Setting 3.

For each setting, the factors of the Monte Carlo simulations are the length of exposure periods Δ, the number of pieces K in Model 2 and the relative incidence rate e^{β}. We took $\Delta = 10,\ 25,$ and $50,\ K = 2,\ 4,\ 6,\ 8$ and$10,$ and $e^{\beta} = 1, 2, 4, 6, 8 and 10$. The combination of these factors gave 90 scenarios for each setting.

3.2 Simulation Results

In this section, we presente the results of our simulation study. The results for Setting A with $\Delta = 10, 25 and 50$ are given in Table 3.1, Table 3.2 and Table 3.3, respectively. The results for Setting B with $\Delta = 10, 25 and 50$ are given in Table 3.4, Table 3.5 and Table 3.6, respectively. Finally, the results for Setting C with $\Delta = 10, 25 and 50$ are given in Table 3.7, Table 3.8 and Table 3.9, respectively. In each table, we present the results with $K = 2, 4, 6, 8 and 10$ and $e^{\beta} = 1, 2, 4, 6, 8 and 10$. For each setting and scenario, we calculate $\bar{\hat{\beta}}$, $\bar{SE}(\hat{\beta})$ and $\bar{Bias}(\hat{\beta})$. They are respectively the average of estimates of β, average of standard error and average of bias. Table 10 shows the simulation results $\bar{SE}(\hat{\beta})$, $SD(\hat{\beta})$ and $MSE(\hat{\beta})$ of third age effect group with $\Delta = 10$, $K = 2, 4, 6, 8 and 10$ and $e^{\beta} = 1, 2, 4, 6, 8 and 10$. $\bar{SE}(\hat{\beta})$, $SD(\hat{\beta})$ and $MSE(\hat{\beta})$ are the average standard error, standard deviation and mean squared error. To keep the

simulation study on a reasonable size, we do not conduct simulation runs under Setting A and Setting B and considere only one value of Δ. However, we will investigate the effect of Δ and age groups as a future work.

First, we consider Setting A. For each k in Table 3.1, the average bias $\bar{Bias}(\hat{\beta}_1)$ and $\bar{Bias}(\hat{\beta}_2)$ decrease first and increase later when e^{β} range from 1 to 10. After all, the average bias, $\bar{Bias}(\hat{\beta}_1)$ and $\bar{Bias}(\hat{\beta}_2)$, are small when $K = 2, 4, 6, 8 and 10$. That is the estimates of β for model $1(\hat{\beta}_1)$ and model $2(\hat{\beta}_2)$ are close to β. The model 1 and model 2 work well for the first age effect group with $\Delta = 50$. Also, the value in the columns of $\bar{SE}(\hat{\beta}_1)$ and $\bar{SE}(\hat{\beta}_2)$ show that the average standard error are small for both model 1 and model 2 with each K and e^{β}. That is the estimates we obtain from this simulation studies for model 1 and model 2 are reliable estimates. When e^{β} range from 2 to 10, the value of $\bar{SE}(\hat{\beta}_1)$ and $\bar{SE}(\hat{\beta}_2)$ becomes smaller. This shows that the estimate results is more reliable when e^{β} becomes larger for each K.

Table 3.2 and Table 3.3 are the simulation results from setting A with $\Delta = 25$ and $\Delta = 10$. By comparing Table 3.1, Table 3.2 and Table 3.3, the value of $\bar{Bias}(\hat{\beta}_1)$ and $\bar{Bias}(\hat{\beta}_2)$ increase when Δ becomes smaller. Relatively speaking, the estimate of β becomes worse when Δ becomes smaller. But the values of $\bar{Bias}(\hat{\beta}_1)$ and $\bar{Bias}(\hat{\beta}_2)$ in Table 3.2 and Table 3.3 are small which means the estimates in Table 3.2 and table 3.3 for model 1 and model 2 are close to β in this simulation studies. As shown by Table 3.2 and table 3.4, $\bar{SE}(\hat{\beta}_1)$ and $\bar{SE}(\hat{\beta}_2)$ are small. It shows that the simulation results in table 2 and table 4 are reliable. By observing Tables 3.1-3.9, we find that the simulation results are similar between different age effect group. This tells us that all these simulation work well for both model 1 and model 2.

As mentioned in previous paragraph, the number of individual with events is $m = 25$ and the number of events for each individuals is $n_i = 1$ or 2. Then the number of events for all individuals in our simulation study could be between 25 and 50. Obviously,

this number of events are small. Normally, we use model 1 to deal the case with small number of events. But, by comparing the simulation results for model 1 and model 2 in all Tables 3.1-3.9, the values of $\hat{\beta}$ and $\bar{Bias}(\hat{\beta})$ all shows that there is no obvious difference between the estimate of model 1 and model 2. In these tables, $\bar{SE}(\hat{\beta}_2)$ are also small and show that the simulation results of model 2 are reliable. This means that model 2 does not lose much information and works well comparing with model 1 in this kind of case. Also, we find that when the number of events becomes larger, the time consuming of model 1 is much larger than the one of model 2. For instance, time cosuming of model 1 and model 2 are 13615 and 499 with $m = 25$, 147381 and 13615 with $m = 50$, 712544 and 3253 with $m = 80$. These time consuming are recorded by R program `pro.time()` under scenario with setting C, $e^{\beta} = 8$, $K = 4$ and $\Delta = 50$. Thus, model 2 is a better choice when number of events becomes larger. At last, by observing table 10, we could find the values of standard error $\bar{SE}(\hat{\beta})$ and standard deviation $SD(\hat{\beta})$ are close with each other. This means that model 1 and model 2 all work well in this simulation studies, that is the conclusions obtained from these simulation results are convicing.

Table 3.1: Simulation results for setting A with scenarios $m = 25$ $\Delta = 50$.

			Model 1			Model 2		
K	e^{β}	β	$\hat{\beta}_1$	$\bar{SE}(\hat{\beta}_1)$	$\bar{Bias}(\hat{\beta}_1)$	$\hat{\beta}_2$	$\bar{SE}(\hat{\beta}_2)$	$\bar{Bias}(\hat{\beta}_2)$
2	1	0.000	-0.148	0.630	-0.148	-0.091	0.642	-0.091
	2	0.693	0.572	0.511	-0.120	0.562	0.498	-0.130
	4	1.386	1.443	0.391	0.057	1.489	0.360	0.103
	6	1.791	1.867	0.390	0.076	1.728	0.365	-0.063
	8	2.079	2.191	0.375	0.111	2.126	0.343	0.047
	10	2.302	2.469	0.421	0.167	2.174	0.345	-0.128
4	1	0.000	-0.103	0.654	-0.103	-0.030	0.743	-0.030
	2	0.693	0.598	0.473	-0.094	0.520	0.567	-0.172
	4	1.386	1.437	0.400	0.051	1.251	0.416	-0.134
	6	1.791	1.881	0.378	0.089	1.726	0.386	-0.065
	8	2.079	2.159	0.431	0.080	1.975	0.422	-0.103
	10	2.302	2.483	0.412	0.180	2.250	0.362	-0.052
6	1	0.000	-0.164	0.649	-0.164	-0.195	0.705	-0.195
	2	0.693	0.625	0.518	-0.067	0.591	0.533	-0.101
	4	1.386	1.418	0.405	0.032	1.321	0.412	-0.064
	6	1.791	1.884	0.456	0.092	1.558	0.445	-0.232
	8	2.079	2.078	0.426	-0.001	1.810	0.387	-0.268
	10	2.302	2.404	0.395	0.101	2.040	0.354	-0.261
8	1	0.000	-0.077	0.618	-0.077	0.022	0.775	0.022
	2	0.693	0.635	0.513	-0.0575	0.594	0.657	-0.098
	4	1.386	1.387	0.389	0.001	1.276	0.497	-0.109
	6	1.791	1.781	0.391	-0.010	1.534	0.454	-0.257
	8	2.079	2.191	0.394	0.111	1.884	0.430	-0.194
	10	2.302	2.4706	0.434	0.168	1.903	0.494	-0.398
10	1	0.000	-0.071	0.655	-0.071	0.088	0.818	0.088
	2	0.693	0.674	0.432	-0.018	0.586	0.613	-0.016
	4	1.386	1.429	0.439	0.043	1.210	0.515	-0.175
	6	1.791	1.901	0.402	0.109	1.566	0.474	-0.224
	8	2.079	2.137	0.452	0.057	1.528	0.504	-0.550
	10	2.302	2.574	0.447	0.272	1.913	0.474	-0.389

Table 3.2: Simulation results for setting A with scenarios $m = 25$ $\Delta = 25$.

			Model 1			Model 2		
K	e^β	β	$\hat{\beta}_1$	$\bar{SE}(\hat{\beta}_1)$	$\bar{Bias}(\hat{\beta}_1)$	$\hat{\beta}_2$	$\bar{SE}(\hat{\beta}_2)$	$\bar{Bias}(\hat{\beta}_2)$
2	1	0.000	0.262	0.785	0.262	0.289	0.755	0.289
	2	0.693	0.989	0.580	0.296	0.946	0.543	0.253
	4	1.386	1.671	0.452	0.284	1.572	0.398	0.185
	6	1.791	2.087	0.449	0.296	1.998	0.394	0.206
	8	2.079	2.287	0.477	0.208	2.120	0.402	0.041
	10	2.302	2.596	0.433	0.293	2.390	0.350	0.088
4	1	0.000	0.386	0.754	0.386	0.394	0.745	0.394
	2	0.693	1.019	0.543	0.326	0.958	0.531	0.265
	4	1.386	1.666	0.539	0.280	1.479	0.500	0.093
	6	1.791	1.971	0.450	0.180	1.864	0.417	0.072
	8	2.079	2.460	0.422	0.380	2.184	0.367	0.105
	10	2.302	2.537	0.409	0.235	2.325	0.371	0.023
6	1	0.000	0.362	0.718	0.362	0.335	0.713	0.335
	2	0.693	1.019	0.598	0.326	0.925	0.580	0.232
	4	1.386	1.808	0.474	0.422	1.653	0.459	0.267
	6	1.791	2.064	0.408	0.272	1.910	0.392	0.118
	8	2.079	2.455	0.388	0.375	2.243	0.257	0.164
	10	2.302	2.738	0.476	0.435	2.337	0.402	0.035
8	1	0.000	0.186	0.726	0.186	0.275	0.759	0.275
	2	0.693	0.997	0.553	0.304	0.997	0.562	0.209
	4	1.386	1.668	0.509	0.281	1.539	0.543	0.153
	6	1.791	2.123	0.437	0.332	1.830	0.473	0.038
	8	2.079	2.331	0.460	0.251	2.105	0.431	0.026
	10	2.302	2.589	0.415	0.286	2.411	0.379	0.108
10	1	0.000	0.343	0.742	0.343	0.440	0.757	0.440
	2	0.693	1.006	0.628	0.313	1.007	0.658	0.314
	4	1.386	1.691	0.433	0.304	1.637	0.459	0.251
	6	1.791	2.137	0.445	0.345	1.982	0.435	0.191
	8	2.079	2.332	0.419	0.253	2.102	0.435	0.023
	10	2.302	2.622	0.421	0.319	2.376	0.428	0.074

Table 3.3: Simulation results for setting A with scenarios $m = 25$ $\Delta = 10$.

			Model 1			Model 2		
K	e^{β}	β	$\hat{\beta}_1$	$\bar{SE}(\hat{\beta}_1)$	$\bar{Bias}(\hat{\beta}_1)$	$\hat{\beta}_2$	$\bar{SE}(\hat{\beta}_2)$	$\bar{Bias}(\hat{\beta}_2)$
2	1	0.000	0.871	0.897	0.871	0.826	0.834	0.826
	2	0.693	1.200	0.811	0.507	1.130	0.736	0.437
	4	1.386	1.742	0.690	0.356	1.632	0.603	0.246
	6	1.791	2.044	0.599	0.252	1.854	0.516	0.062
	8	2.079	2.443	0.626	0.364	2.177	0.480	0.097
	10	2.302	2.500	0.621	0.197	2.282	0.488	-0.020
4	1	0.000	0.659	0.887	0.659	0.609	0.849	0.609
	2	0.693	1.151	0.821	0.458	1.099	0.759	0.406
	4	1.386	1.707	0.694	0.321	1.578	0.618	0.191
	6	1.791	2.135	0.590	0.344	1.946	0.501	0.154
	8	2.079	2.437	0.569	0.357	2.155	0.473	0.075
	10	2.302	2.652	0.570	0.349	2.283	0.454	-0.019
6	1	0.000	0.861	0.902	0.861	0.777	0.841	0.777
	2	0.693	1.237	0.810	0.544	1.140	0.734	0.447
	4	1.386	1.785	0.670	0.399	1.607	0.588	0.221
	6	1.791	2.068	0.615	0.276	1.858	0.535	0.066
	8	2.079	2.340	0.651	0.260	2.084	0.528	0.004
	10	2.302	2.587	0.613	0.284	2.212	0.500	-0.090
8	1	0.000	0.671	0.868	0.671	0.666	0.835	0.666
	2	0.693	1.146	0.833	0.452	1.035	0.778	0.342
	4	1.386	1.747	0.700	0.360	1.606	0.636	0.220
	6	1.791	2.106	0.591	0.314	1.886	0.540	0.094
	8	2.079	2.462	0.586	0.383	2.169	0.477	0.089
	10	2.302	2.573	0.630	0.270	2.202	0.523	-0.099
10	1	0.000	0.769	0.946	0.769	0.691	0.903	0.691
	2	0.693	1.229	0.764	0.536	1.167	0.724	0.474
	4	1.386	1.695	0.688	0.309	1.574	0.642	0.187
	6	1.791	2.169	0.680	0.377	1.898	0.606	0.107
	8	2.079	2.511	0.551	0.431	2.184	0.466	0.104
	10	2.302	2.613	0.603	0.310	2.218	0.489	-0.084

Table 3.4: Simulation results for setting B with scenarios $m = 25$ $\Delta = 50$.

			Model 1			Model 2		
K	e^{β}	β	$\hat{\beta}_1$	$\bar{SE}(\hat{\beta}_1)$	$\bar{Bias}(\hat{\beta}_1)$	$\hat{\beta}_2$	$\bar{SE}(\hat{\beta}_2)$	$\bar{Bias}(\hat{\beta}_2)$
2	1	0.000	-0.125	0.678	-0.125	-0.188	0.686	-0.188
	2	0.693	0.660	0.510	-0.032	0.581	0.508	-0.111
	4	1.386	1.204	0.436	-0.181	1.057	0.414	-0.329
	6	1.791	1.828	0.420	0.036	1.708	0.363	-0.083
	8	2.079	2.098	0.408	0.018	1.860	0.359	-0.219
	10	2.302	2.493	0.418	0.191	2.289	0.366	-0.013
4	1	0.000	-0.175	0.589	-0.175	-0.194	0.625	-0.194
	2	0.693	0.653	0.545	-0.039	0.618	0.519	-0.074
	4	1.386	1.255	0.414	-0.130	1.172	0.431	-0.214
	6	1.791	1.787	0.386	-0.003	1.667	0.378	-0.124
	8	2.079	2.0848	0.377	0.005	1.973	0.360	-0.106
	10	2.302	2.447	0.420	0.145	2.223	0.367	-0.078
6	1	0.000	-0.155	0.739	-0.155	-0.116	0.782	-0.116
	2	0.693	0.474	0.593	-0.218	0.388	0.648	-0.304
	4	1.386	1.322	0.440	-0.063	1.150	0.513	-0.235
	6	1.791	1.842	0.390	0.050	1.669	0.406	-0.122
	8	2.079	2.051	0.427	-0.027	1.709	0.425	-0.370
	10	2.302	2.343	0.429	0.041	2.001	0.394	-0.301
8	1	0.000	-0.118	0.679	-0.118	0.018	0.818	0.018
	2	0.693	0.585	0.481	-0.107	0.489	0.635	-0.203
	4	1.386	1.381	0.410	-0.004	1.205	0.489	-0.181
	6	1.791	1.814	0.409	0.022	1.506	0.468	-0.285
	8	2.079	2.217	0.381	0.138	1.829	0.428	-0.249
	10	2.302	2.330	0.413	0.028	1.929	0.430	-0.373
10	1	0.000	-0.228	0.636	-0.228	-0.171	0.840	-0.171
	2	0.693	0.663	0.453	-0.030	0.582	0.655	-0.110
	4	1.386	1.351	0.464	-0.034	1.039	0.614	-0.347
	6	1.791	1.869	0.423	0.077	1.401	0.525	-0.390
	8	2.079	2.066	0.454	-0.012	1.602	0.546	-0.477
	10	2.302	2.356	0.388	0.053	1.713	0.431	-0.588

Table 3.5: Simulation results for setting B with scenarios $m = 25$ $\Delta = 25$.

			Model 1			Model 2		
K	e^β	β	$\hat{\beta}_1$	$\bar{SE}(\hat{\beta}_1)$	$\bar{Bias}(\hat{\beta}_1)$	$\hat{\beta}_2$	$\bar{SE}(\hat{\beta}_2)$	$\bar{Bias}(\hat{\beta}_2)$
2	1	0.000	0.260	0.685	0.260	0.302	0.665	0.302
	2	0.693	0.948	0.603	0.255	0.863	0.548	0.170
	4	1.386	1.674	0.474	0.288	1.501	0.432	0.115
	6	1.791	2.127	0.454	0.335	2.047	0.380	0.225
	8	2.079	2.389	0.424	0.310	2.151	0.356	0.072
	10	2.302	2.330	0.464	0.027	2.253	0.375	-0.048
4	1	0.000	0.402	0.733	0.402	0.383	0.716	0.383
	2	0.693	1.025	0.614	0.332	0.897	0.615	0.203
	4	1.386	1.698	0.499	0.312	1.561	0.447	0.175
	6	1.791	2.058	0.509	0.266	1.826	0.439	0.034
	8	2.079	2.305	0.445	0.226	2.116	0.386	0.036
	10	2.302	2.627	0.440	0.324	2.405	0.374	0.102
6	1	0.000	0.334	0.698	0.334	0.314	0.719	0.314
	2	0.693	0.996	0.578	0.303	0.939	0.569	0.246
	4	1.386	1.690	0.534	0.304	1.528	0.489	0.142
	6	1.791	2.047	0.459	0.255	1.903	0.425	0.111
	8	2.079	2.360	0.413	0.280	2.251	0.382	0.171
	10	2.302	2.571	0.416	0.269	2.322	0.372	0.019
8	1	0.000	0.240	0.776	0.240	0.240	0.819	0.240
	2	0.693	1.090	0.571	0.397	1.049	0.617	0.356
	4	1.386	1.673	0.494	0.287	1.545	0.528	0.158
	6	1.791	2.148	0.415	0.357	1.996	0.411	0.205
	8	2.079	2.288	0.429	0.209	2.131	0.385	0.052
	10	2.302	2.544	0.457	0.241	2.300	0.419	-0.002
10	1	0.000	0.287	0.698	0.287	0.279	0.729	0.279
	2	0.693	0.997	0.574	0.303	1.008	0.622	0.315
	4	1.386	1.744	0.469	0.358	1.654	0.477	0.267
	6	1.791	2.099	0.491	0.307	1.932	0.493	0.140
	8	2.079	2.175	0.467	0.095	1.880	0.524	-0.199
	10	2.302	2.585	0.457	0.282	2.343	0.447	0.040

Table 3.6: Simulation results for setting B with scenarios $m = 25$ $\Delta = 10$.

			Model 1			Model 2		
K	e^β	β	$\hat{\beta}_1$	$\bar{SE}(\hat{\beta}_1)$	$\bar{Bias}(\hat{\beta}_1)$	$\hat{\beta}_2$	$\bar{SE}(\hat{\beta}_2)$	$\bar{Bias}(\hat{\beta}_2)$
2	1	0.000	0.859	0.900	0.859	0.742	0.842	0.742
	2	0.693	1.256	0.833	0.563	1.103	0.759	0.410
	4	1.386	1.815	0.690	0.429	1.575	0.591	0.189
	6	1.791	2.173	0.593	0.382	1.983	0.494	0.192
	8	2.079	2.341	0.600	0.262	2.088	0.490	0.009
	10	2.302	2.562	0.622	0.259	2.208	0.482	-0.094
4	1	0.000	0.795	0.884	0.795	0.728	0.835	0.728
	2	0.693	1.253	0.799	0.560	1.134	0.729	0.440
	4	1.386	1.838	0.699	0.451	1.615	0.600	0.229
	6	1.791	2.162	0.595	0.371	1.953	0.503	0.161
	8	2.079	2.327	0.583	0.247	2.051	0.491	-0.027
	10	2.302	2.523	0.551	0.221	2.290	0.450	-0.012
6	1	0.000	0.689	0.843	0.689	0.678	0.821	0.678
	2	0.693	1.181	0.836	0.488	1.142	0.784	0.449
	4	1.386	1.627	0.646	0.240	1.520	0.591	0.134
	6	1.791	2.086	0.670	0.294	1.840	0.574	0.048
	8	2.079	2.271	0.602	0.191	2.033	0.508	-0.046
	10	2.302	2.649	0.545	0.347	2.377	0.441	0.075
8	1	0.000	0.784	0.902	0.784	0.682	0.856	0.682
	2	0.693	1.140	0.830	0.447	1.054	0.791	0.361
	4	1.386	1.762	0.724	0.376	1.568	0.646	0.182
	6	1.791	2.129	0.608	0.337	1.935	0.528	0.143
	8	2.079	2.330	0.610	0.251	2.104	0.513	0.024
	10	2.302	2.537	0.561	0.234	2.265	0.466	-0.036
10	1	0.000	0.779	0.864	0.779	0.666	0.838	0.827
	2	0.693	1.276	0.811	0.583	1.161	0.754	0.468
	4	1.386	1.830	0.705	0.444	1.615	0.640	0.229
	6	1.791	2.179	0.615	0.387	1.907	0.542	0.115
	8	2.079	2.440	0.621	0.360	2.067	0.506	-0.011
	10	2.302	2.640	0.628	0.338	2.227	0.511	-0.075

Table 3.7: Simulation results for setting C with scenarios $m = 25$ $\Delta = 50$.

			Model 1			Model 2		
K	e^{β}	β	$\hat{\beta}_1$	$\bar{SE}(\hat{\beta}_1)$	$\bar{Bias}(\hat{\beta}_1)$	$\hat{\beta}_2$	$\bar{SE}(\hat{\beta}_2)$	$\bar{Bias}(\hat{\beta}_2)$
2	1	0.000	-0.256	0.721	-0.256	-0.231	0.729	-0.231
	2	0.693	0.524	0.457	-0.168	0.550	0.435	-0.142
	4	1.386	1.298	0.373	-0.088	1.250	0.362	-0.135
	6	1.791	1.715	0.398	-0.076	1.567	0.362	-0.233
	8	2.079	2.307	0.431	0.228	2.098	0.388	0.019
	10	2.302	2.238	0.391	-0.064	2.089	0.337	-0.213
4	1	0.000	-0.228	0.681	-0.228	-0.177	0.717	-0.177
	2	0.693	0.496	0.520	-0.196	0.455	0.568	-0.237
	4	1.386	1.357	0.476	-0.028	1.247	0.492	-0.138
	6	1.791	1.914	0.405	0.122	1.792	0.379	0.001
	8	2.079	2.142	0.440	0.062	1.810	0.413	-0.269
	10	2.302	2.502	0.460	0.200	2.270	0.424	-0.032
6	1	0.000	-0.192	0.634	-0.192	-0.165	0.707	-0.165
	2	0.693	0.578	0.494	-0.114	0.550	0.553	-0.142
	4	1.386	1.256	0.404	-0.129	1.206	0.429	-0.179
	6	1.791	1.740	0.403	-0.051	1.595	0.413	-0.196
	8	2.079	2.033	0.386	-0.046	1.888	0.402	-0.190
	10	2.302	2.275	0.396	-0.026	1.964	0.428	-0.338
8	1	0.000	-0.284	0.675	-0.284	-0.291	0.820	-0.291
	2	0.693	0.435	0.504	-0.257	0.351	0.586	-0.341
	4	1.386	1.352	0.433	-0.034	1.185	0.512	-0.200
	6	1.791	1.715	0.443	-0.075	1.421	0.479	-0.370
	8	2.079	2.014	0.407	-0.065	1.617	0.442	-0.461
	10	2.302	2.371	0.429	0.068	2.013	0.500	-0.288
10	1	0.000	-0.191	0.625	-0.191	-0.117	0.776	-0.117
	2	0.693	0.450	0.534	-0.242	0.298	0.764	-0.394
	4	1.386	1.057	0.433	-0.328	0.763	0.550	-0.623
	6	1.791	1.859	0.420	0.067	1.463	0.531	-0.327
	8	2.079	2.223	0.423	0.144	1.862	0.472	-0.217
	10	2.302	2.472	0.473	0.170	1.552	0.564	-0.750

Table 3.8: Simulation results for setting C with scenarios $m = 25$ $\Delta = 25$.

			Model 1			Model 2		
K	e^{β}	β	$\hat{\beta}_1$	$\bar{SE}(\hat{\beta}_1)$	$\bar{Bias}(\hat{\beta}_1)$	$\hat{\beta}_2$	$\bar{SE}(\hat{\beta}_2)$	$\bar{Bias}(\hat{\beta}_2)$
2	1	0.000	0.259	0.659	0.259	0.283	0.642	0.283
	2	0.693	0.929	0.611	0.236	0.858	0.573	0.165
	4	1.386	1.631	0.476	0.244	1.532	0.426	0.145
	6	1.791	2.123	0.483	0.331	2.004	0.400	0.212
	8	2.079	2.320	0.473	0.240	2.323	0.399	0.243
	10	2.302	2.537	0.507	0.234	2.223	0.407	-0.079
4	1	0.000	0.245	0.760	0.245	0.293	0.792	0.293
	2	0.693	0.946	0.585	0.253	0.885	0.562	0.192
	4	1.386	1.863	0.534	0.477	1.809	0.467	0.422
	6	1.791	1.991	0.445	0.199	1.939	0.386	0.147
	8	2.079	2.339	0.422	0.260	2.164	0.377	0.084
	10	2.302	2.493	0.443	0.191	2.277	0.379	-0.024
6	1	0.000	0.318	0.801	0.318	0.399	0.802	0.399
	2	0.693	0.953	0.538	0.259	0.917	0.533	0.223
	4	1.386	1.386	0.531	0.000	1.268	0.514	-0.117
	6	1.791	2.107	0.417	0.316	1.962	0.392	0.171
	8	2.079	2.34	0.421	0.269	2.148	0.410	0.069
	10	2.302	2.517	0.470	0.214	2.352	0.424	0.050
8	1	0.000	0.274	0.651	0.274	0.273	0.687	0.273
	2	0.693	1.004	0.549	0.311	0.899	0.600	0.206
	4	1.386	1.615	0.477	0.228	1.509	0.474	0.123
	6	1.791	2.059	0.450	0.267	1.850	0.481	0.058
	8	2.079	2.298	0.437	0.218	2.132	0.431	0.053
	10	2.302	2.685	0.482	0.382	2.401	0.452	0.098
10	1	0.000	0.201	0.793	0.201	0.219	0.855	0.219
	2	0.693	1.101	0.580	0.408	1.112	0.593	0.419
	4	1.386	1.560	0.550	0.174	1.515	0.614	0.129
	6	1.791	2.046	0.443	0.254	1.893	0.444	0.101
	8	2.079	2.332	0.446	0.253	2.129	0.466	0.0495
	10	2.302	2.488	0.451	0.186	2.198	0.429	-0.103

Table 3.9: Simulation results for setting C with scenarios $m = 25$ $\Delta = 10$.

			Model 1			Model 2		
K	e^β	β	$\hat{\beta}_1$	$\bar{SE}(\hat{\beta}_1)$	$\bar{Bias}(\hat{\beta}_1)$	$\hat{\beta}_2$	$\bar{SE}(\hat{\beta}_2)$	$\bar{Bias}(\hat{\beta}_2)$
2	1	0.000	0.777	0.892	0.777	0.781	0.843	0.781
	2	0.693	1.148	0.847	0.455	0.989	0.773	0.295
	4	1.386	1.769	0.750	0.383	1.562	0.656	0.175
	6	1.791	2.131	0.667	0.339	1.891	0.557	0.100
	8	2.079	2.380	0.600	0.300	2.174	0.487	0.095
	10	2.302	2.531	0.552	0.228	2.270	0.451	-0.032
4	1	0.000	0.877	0.928	0.877	0.823	0.865	0.823
	2	0.693	1.205	0.780	0.512	1.091	0.718	0.397
	4	1.386	1.637	0.658	0.251	1.561	0.595	0.175
	6	1.791	2.027	0.633	0.235	1.875	0.555	0.084
	8	2.079	2.238	0.593	0.158	2.048	0.509	-0.030
	10	2.302	2.538	0.610	0.236	2.173	0.498	-0.128
6	1	0.000	0.743	0.908	0.743	0.702	0.860	0.702
	2	0.693	1.029	0.773	0.336	0.968	0.737	0.275
	4	1.386	1.757	0.659	0.370	1.612	0.586	0.226
	6	1.791	2.074	0.624	0.283	1.863	0.536	0.071
	8	2.079	2.310	0.621	0.231	2.112	0.522	0.033
	10	2.302	2.425	0.599	0.123	2.178	0.483	-0.124
8	1	0.000	0.794	0.889	0.794	0.781	0.860	0.781
	2	0.693	1.167	0.812	0.474	1.150	0.777	0.456
	4	1.386	1.688	0.755	0.302	1.538	0.710	0.152
	6	1.791	2.037	0.591	0.245	1.860	0.536	0.068
	8	2.079	2.299	0.566	0.220	2.029	0.502	-0.049
	10	2.302	2.456	0.583	0.153	2.185	0.498	-0.117
10	1	0.000	0.784	0.921	0.784	0.679	0.885	0.679
	2	0.693	1.123	0.721	0.430	1.027	0.694	0.334
	4	1.386	1.687	0.663	0.301	1.555	0.625	0.168
	6	1.791	2.028	0.637	0.236	1.821	0.568	0.029
	8	2.079	2.349	0.595	0.270	2.050	0.538	-0.029
	10	2.302	2.500	0.573	0.197	2.165	0.486	-0.137

Table 3.10: Simulation results for setting C with scenarios $m = 25$ $\Delta = 10$.

		Model 1			Model 2		
K	e^{β}	$\bar{SE}(\hat{\beta}_1)$	$SD(\hat{\beta}_1)$	$MSE(\hat{\beta}_1)$	$\bar{SE}(\hat{\beta}_2)$	$SD(\hat{\beta}_2)$	$MSE(\hat{\beta}_2)$
2	1	0.892	0.594	0.957	0.843	0.506	0.867
	2	0.847	0.684	0.674	0.773	0.586	0.431
	4	0.750	0.740	0.695	0.656	0.597	0.387
	6	0.667	0.713	0.623	0.557	0.567	0.331
	8	0.600	0.669	0.537	0.487	0.505	0.264
	10	0.552	0.610	0.424	0.451	0.480	0.231
4	1	0.928	0.627	1.162	0.865	0.511	0.938
	2	0.780	0.700	0.752	0.718	0.518	0.600
	4	0.658	0.666	0.506	0.595	0.578	0.364
	6	0.633	0.662	0.493	0.555	0.554	0.313
	8	0.593	0.659	0.459	0.509	0.529	0.280
	10	0.610	0.674	0.509	0.498	0.553	0.322
6	1	0.908	0.578	0.886	0.860	0.489	0.732
	2	0.773	0.657	0.544	0.737	0.576	0.407
	4	0.659	0.676	0.595	0.586	0.579	0.386
	6	0.624	0.679	0.541	0.536	0.556	0.314
	8	0.621	0.676	0.510	0.522	0.541	0.294
	10	0.599	0.671	0.465	0.483	0.527	0.293
8	1	0.889	0.614	1.007	0.860	0.524	0.885
	2	0.812	0.709	0.727	0.777	0.605	0.574
	4	0.755	0.739	0.637	0.710	0.646	0.441
	6	0.591	0.641	0.471	0.536	0.547	0.304
	8	0.566	0.624	0.438	0.502	0.506	0.259
	10	0.583	0.680	0.485	0.498	0.524	0.289
10	1	0.921	0.595	0.970	0.885	1.534	2.812
	2	0.721	0.676	0.642	0.694	1.026	1.163
	4	0.663	0.681	0.555	0.625	0.606	0.396
	6	0.637	0.691	0.533	0.568	0.578	0.335
	8	0.595	0.648	0.493	0.538	0.554	0.307
	10	0.573	0.626	0.431	0.486	0.510	0.279

Chapter 4

Application: Measles, Mumps and Rubella Vaccination and Idiopathic Thrombocytopaenic Purpura

In this chapter, we apply the methods discussed in the previous chapters to analyze a data set from medicine. Our goal is to illustrate the use of the self-controlled case series (SCCS) design in the investigation of the association between the measles, mumps, rubella (MMR) vaccination and idiopathic thrombocytopenic purpura (ITP). We consider this issue in different settings including parametric, semiparametric, and piecewise-constant baseline rate functions methods.

We first discuss the background information about MMR vaccination and ITP in the next section. In Section 4.2, we introduce the models, and analyze the data. In the final section, we present the conclusion.

4.1 Background

Measles is a contagious infection caused by the measles virus. It is an airborne disease which spreads easily through the contact with saliva or nasal secretions, coughs and sneezes of those infected. Nine out of ten people who are not immune who share living space with an infected person will catch it. Testing for the virus in suspected cases is important for public health efforts (Atkinson, 2011).

Mumps is another viral disease of childhood, which is caused by the mumps virus. Mumps is highly contagious and spreads rapidly. The virus is transmitted by respiratory droplets or direct contact with an infected person (Atkinson, 2012). Without immunization about 0.1% to 1% of the population are affected per year. Widespread vaccination has resulted in a more than 90% decline in rates of disease (Junghanss, 2013).

Rubella is an infection caused by the rubella virus. Rubella is usually spread through the air via coughs of people who are infected. Once recovered, people are immune to future infections. Rubella is a common infection in many areas of the world. Each year about 100,000 cases of congenital rubella syndrome occur (Lambert et al, 2015).

The MMR vaccine administrated via injection is an immunization vaccine against MMR. It is a mixture of live attenuated viruses of the three diseases. The MMR vaccine is ideally administered to children around the age of one year, with a second dose at the age 4 or 5 years. According to Immunization Action Coalition (IAC), the second dose is a dose to produce immunity in the small number of persons (2–5%) who fail to develop measles immunity after the first dose. It is usually considered a childhood vaccination. It is widely used around the world; since introduction of its earliest versions in the 1970s, over 500 million doses have been used in over 60 countries.

The association of the administration of the MMR vaccine and adverse events in

children has been investigated by many reserachers. According to National Institutes of Health (NIH), idiopathic thrombocytopenic purpura (ITP) is a rare, potentially recurrent bleeding disorder in which the immune system destroys platelets, which are necessary for normal blood clotting. ITP is considered as an disease such as its ocurrence rate increases within a few weeks of MMR immunisation, and its occurrence prior MMR vaccination does not affect the rate of administartion fo the MMR vaccine (Miller et al., 2001). It is an important research question to reveal whether there is a relation between the administartaion of the MMR vaccine and inrease rate in the occurence of ITP within a short time period after administartion of MMR vaccine in children. There are some studies considered this issue (for example, see Miller et al., 2001; Black et al., 2003). Our approach is similar to the one given by Farrington and Whitaker (2006) who discussed this issue under an outcome dependent sampling design setting as well.

4.2 Data Analysis

The data set (presented in Appendix) is obtained through a study in the UK in 1988, and an updated version of it was considered by Farrington and Whitaker (2006). Our goal here is to use this data set to illustrate the estimation based on some models we discussed in Chapter 2.

The data set includes information about 35 children (i.e, cases with $m = 35$) aged between 366 and 730 days. These children developed ITP, and therefore, included in this study. The distribution of the number of observed events (i.e., the occurrence of ITP) over the follow-up times showed that there is some heterogeneity in the observed number of events per subject. One of the advantages of the SCCS method is that it automatically controls for such variations as it is self-controlled. Otherwise, a model

based on a cohort should include terms such as different parametric baseline rate functions or random effects for each subject to address variations in the number of observed events.

We start our analysis with the estimation of the cumulative mean function $\mu(t)$, $t \geq 0$. To this end, we use the Nelson-Aalen estimator $\hat{\mu}(t)$, a nonparametric maximum likelihood estimator of $\mu(t)$ (Cook and Lawless, 2007, Section 3.4.1). Figure 4.1 shows a plot of this estimator with 95% pointwise confidence intervals of $\mu(t)$ based on robust variance estimator of $\hat{\mu}(t)$ (Cook and Lawless, 2007, Section 3.6.1). The concave down shape of the plot of the Nelson-Aalen estimate of the mean function reveals that there is a monotonically decreasing trend in the rate of occurence of failure as time increases.

In addition, we consider a parametric nonhmogeneous Poisson model to check whether there is a trend in the baseline rate function. For this purpose, we fit the Power-Law-Process (PLP) model, in which the intensity function, conditional on the path of the covariate is given by

$$\rho_i(t \,|\, x_i) = \alpha_i \, \gamma \, t^{\gamma-1} \, e^{\beta \, x_i(t)}, \qquad i = 1, \ldots, m, \tag{4.1}$$

where $\gamma > 1$ ($\gamma < 1$) indicates that there is an increasing (a decreasing) trend in the rate of occurrance of events as t increases. Also, a test for the null hypothesis $H_0 : \gamma = 1$ against the alternative hypothesis $H_a : \gamma \neq 1$ can be developed for a monotonic trend in the baseline rate function. In the intensity function (4.1), the covariate $x_i(t)$ is an indicator function for the risk period. That is, it takes the value of 1 after the administration of the MMR vaccination and stays at 1 for a Δ time period, what is defined as the risk period. Figure 4.2 displays the life history of a subject with two events occurred over the observation window with start time

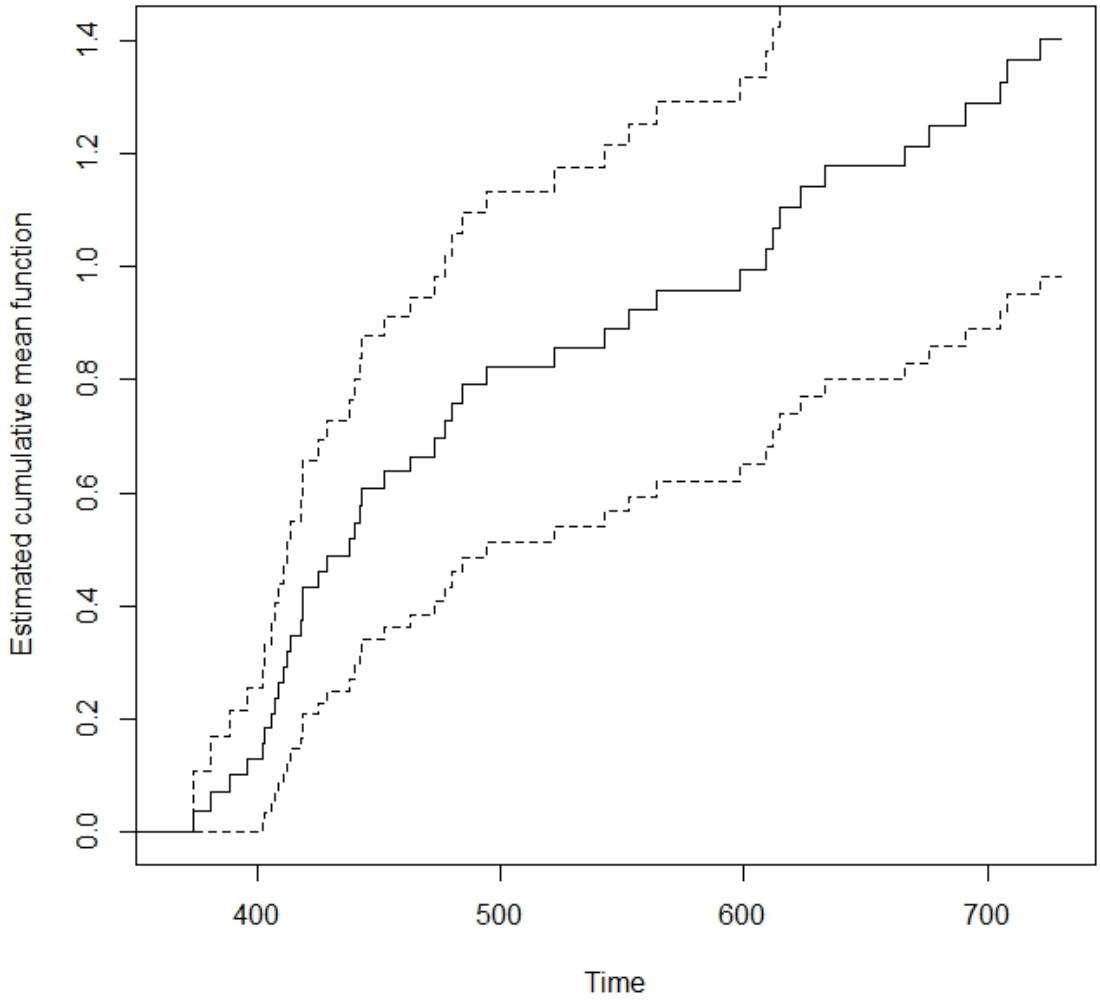

Figure 4.1: The Nelson-Aalen plot of the cumulative mean function $\mu(t)$, where time t denotes days.

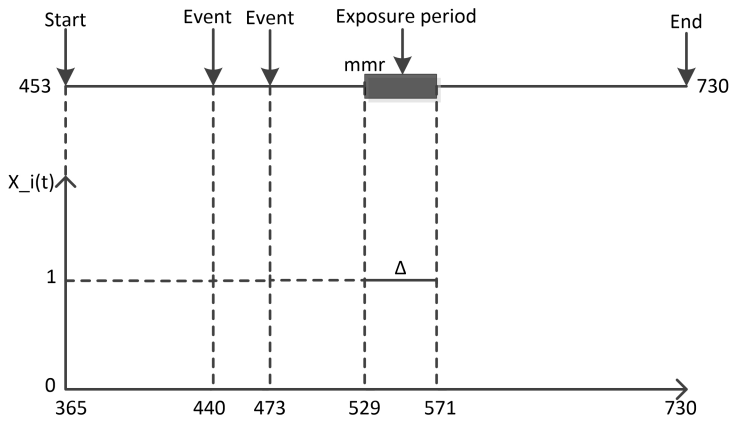

Figure 4.2: Event history plot of a subject with its corresponding path of the covariate $x_i(t)$, where the exposure (risk) period Δ is taken as 42 days.

$\tau_{0i} = 365$ and end-of-followup time $\tau_i = 720$ in days. For this particular subject, the MMR vaccination is on the 529th day and the risk period Δ is 42 days. The same figure also shows the corresponding covariate path of the same subject. We consider a conditional likelihood function to estimate the parameters in the model (4.1). Given the number of events over observation windows and covariate path, the conditional likelihood function is

$$L_c(\gamma, \beta) = \prod_{i=1}^{35} \left\{ \prod_{j=1}^{n_i} \left[\frac{\gamma\, t_{ij}^{\gamma-1}\, e^{\beta\, x_i(t_{ij})}}{\int_{\tau_{0i}}^{\tau_i} \gamma\, t^{\gamma-1} e^{\beta\, x_i(t)}\, dt} \right] \right\}. \tag{4.2}$$

Since the likelihood function (4.2) is based on the SCCS design, the α_i in the model (4.1) are cancel each out in the conditional likelihood function. We fit the conditional log likelihood $\ell_c(\gamma, \beta) = \log L_c(\gamma, \beta)$, and obtain the maximum likelihood estimates $\hat{\gamma}$ and $\hat{\beta}$ and their standard errors with the `nlm` function in R. The estimates of γ and β are presented in Table 4.1 for four different Δ values (in days). A Wald type test statistics rejects the null hypothesis $H_0 : \gamma = 1$ in favor of $H_a : \gamma \neq 1$ at 0.05 level for all Δ values considered in Table 4.1. Therefore, we conclude that there is a significant, monotonically decreasing trend in the rate function. It should be also noted

Table 4.1: Estimation results for Model (4.1) are given. Δ denotes risk period in days. $SE(\hat{\gamma})$ and $SE(\hat{\beta})$ denote the standard errors of the maximum likelihood estimates $\hat{\gamma}$ and $\hat{\beta}$, respectively. $-\ell_c^{\max}$ is the negative of the log of $L_c(\hat{\gamma}, \hat{\beta})$ given in (4.2).

Δ	$\hat{\gamma}$	$SE(\hat{\gamma})$	$\hat{\beta}$	$SE(\hat{\beta})$	$-\ell_c^{\max}$
21	-1.201	0.802	0.643	0.498	253.12
42	-0.739	0.840	1.285	0.370	248.55
63	-0.818	0.850	1.092	0.363	249.63
84	-0.896	0.857	1.016	0.360	250.00

that, since the value of the conditional log likelihood $\ell_c(\gamma, \beta)$ is the highest, the data support the value of Δ at 42 days.

We next consider three models:

M1: Semiparametric SCCS model with the intensity function

$$\text{Model 1:} \qquad \rho_0(t)\, \exp[\beta x_i(t)], \tag{4.3}$$

where $\rho_0(t)$ is left parametrically unspecified.

M2: Piecewise-constant rate model for the SCCS design with the intensity function

$$\text{Model 2:} \qquad \eta\, \psi(t;\alpha) \exp[\gamma_i + \beta x_i(t)], \tag{4.4}$$

where $\alpha = (\alpha_1, \alpha_2, \alpha_3)'$ and

$$\psi(t;\alpha) = \begin{cases} e^{\alpha_1}, & \text{if } t \in (365, 486.67] \\ e^{\alpha_2}, & \text{if } t \in (486.67, 608.33] \\ e^{\alpha_3}, & \text{if } t \in (608.33, 730]. \end{cases} \tag{4.5}$$

M3: Parametric SCCS model with the intensity function

$$\text{Model 3:} \qquad \eta \exp[\gamma_i + \beta x_i(t)]. \tag{4.6}$$

Since there is a trace of a significant trend in the baseline rate function, a semiparametric SCCS model (i.e., Model 1) is useful. Since the event of interest is rare, this model is computationally not demaning to fit in this example. Second model is a flexible parametric Poisson model. We specify the baseline with three pieces (i.e., $K = 3$). There is going to be some amount of bias with respect to misspecification of the number of pieces. This bias decreses as the number of pieces K increases. However, as noted by Cook and Lawless (2007), depending on the shape of the baseline hazard function, a choice of K between 3 and 10 gives close approximation of estimates with those obtained with a semiparametric model. Model 2 is computationally more efficient than Model 1 so it can easily be implemented when the event of interest is not rare. We also fitted Model 3 which assumes no trend in the baseline rate function.

The estimates of the parameter β are presented in Table 4.2 under four different risk periods Δ. Table 4.2 also includes the total number of events observed within a Δ time period after MMR vaccination (denoted by $Obs(\Delta)$), as well as the total expected number of events during the same risk period (denoted by $Exp(\Delta)$). To calculate $Exp(\Delta)$, we use Model 3 when there is no adverse effect of the vaccination; that is, when $\beta = 0$.

A simple comparison of $Obs(\Delta)$ with $Exp(\Delta)$ shows that the total number of events observed over the risk periods is larger than the expected number of events, which indicates an increased number of events for a short period of time after the vaccination. A Wald type test for $H_0 : \beta = 0$ against $H_a : \beta \neq 0$ shows that β is significant at the 0.05 level when $\Delta = 42$, 63 and 84 days. The estimates are very close under

Table 4.2: Estimates of β for Model 1, Model 2 and Model 3. Δ is the length value of risk period. $Obs(\Delta)$ and $Exp(\Delta)$ are the observed number of events in exposure time period and expected number of events in exposure time period when $\beta = 0$. $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$ are the estimate of β under Model 1, Model 2 and Model 3, respectively, and SE denotes their standard errors.

Δ	$Obs(\Delta)$	$Exp(\Delta)$	$\hat{\beta}_1$	$SE(\hat{\beta}_1)$	$\hat{\beta}_2$	$SE(\hat{\beta}_2)$	$\hat{\beta}_3$	$SE(\hat{\beta}_3)$
21	5	2.598	0.501	0.520	0.558	0.497	0.979	0.483
42	13	5.197	1.103	0.396	1.099	0.385	1.536	0.351
63	15	7.795	1.015	0.405	0.967	0.387	1.335	0.344
84	17	10.394	1.022	0.401	1.018	0.393	1.242	0.343

Model 1 and Model 2 especially when $\Delta = 42$ days. Model 3 does not produce close estimates of β to those based on Model 1. It should be noted that, since Model 3 does not include a trend in the baseline, it produces poor results. However, the model is improved by fitting a piecewise-constant baseline model with only 3 pieces. Furthermore, since Model 2 includes less parameters to be estimated compared with Model 1, the standard errors of the estimate of β based on Model 2 is smaller than those based on Model 1.

4.3 Conclusion

Our preliminary analysis of the data set showed that there is a monotonically decreasing trend in the rate of occurrence of ITP disease as time increases. Therefore, a semiparametric model would be a good choice for the investigation of the association between MMR vaccine and ITP disease. We therefore consider a semiparametric SCCS model (Model 1). This model was also used by Farrington and Whitaker (2006), but the resuts are given only when $\Delta = 42$. In our analysis, we also fit a Poisson model with piecewise-constant baseline functions (Model 2). With only 3 pieces ($K = 3$), we show that Model 1 and Model 2 give close estimates of the relative incidence rate;

that is, $\exp(\beta)$. When the risk period Δ is 42 days, we show that there is a positive significant association between the MMR vaccine and ITP occurrrence rate within a short time period after the administration of the MMR vaccine. Our conlcusion is also similar to the conclusion of Farrington and Whitaker (2006).

Chapter 5

Summary and Conclusion

In this chapter, we give the summary and conclusion of the thesis. The self-controlled case series (SCCS) design is a relatively new outcome-dependent sampling design introduced by Farrington (1995). The main objective in a SCCS design is to investigate the association between time-varying exposures and outcome events. This design automatically adjusts for all fixed covariates acting multiplicatively on the intensity function of a subject. Since it is computationally efficient when the event of interest is rare and provides consistent estimates of the relative incidence rate, the SCCS design has received condsiderable recent attention. Therefore, the main objective of this thesis is to investigate the SCCS design through simulations. We consider parametric, semiparamteric and weakly parametric SCCS model, and compare them with well-known models based on the classical cohort design. We also illustrate the methods with a real life data set from medicine.

The main advantage of the SCCS design is that it only depends on the cases. In other words, only the subjects with at least one event needs to be sampled, and controls can be safely ignored in the SCCS design. Since only cases are included, it is economically and computationally efficient compared with a cohort design. This property of the

SCCS design also helps protect data privacy. Because of these reasons, the SCCS design is an important alternative to the cohort design especially when the outcome of interest is a rare event, and has been used in many studies in medicine, epidemiology and pharmacoepidemiology.

There is an increasing interest in using large administrative health care databases. In these studies, computational effciency of a method may become critical. This issue is especially important in the semiparametric modeling of SCCS design when the event of interest is not rare. For example, in a small simulation study with the Setting C of Chapter 3, we show that, when the risk period Δ is 50 days, the relative incidence rate e^{β} is 8 and the number of cases m is 25, the semiparametric SCCS model needed 226 minutes to estimate β. Whereas, using the same computer system, the SCCS model with piecewise-constant baseline rates (PWC-SCCS) with 4 pieces needed about 8 minutes to estimate β. Under the same setup, when we increase m to 50, the semiparametric SCCS model needed 2456 minutes, but the PWC-SCCS model with 4 pieces needed only 21 minutes to estimate β. Our simulation studies show that, under various settings, the PWC-SCCS models with the number of pieces between 4 to 10 gives close estimates of β to that of obtained from the semiparametric SCCS design. Therefore, we recommend the use of the PWC-SCCS in investigation of the association between time-varying transient exposures and an event, especially when the event of interest is not rare. Depending on the shape of the baseline rate function, more flexible PWC-SCCS models can be obtained by including more parameters in the baseline rate functions.

The SCCS design is not perfect. There are important research questions related to the estimation and statistical efficiency. For example a key assumption in the develoment of the SCCS design is that the observation periods $(\tau_{0i}, \tau_i]$ need to be independent of the event processes $\{N_i(t); t \geq 0\}$ for all process under observation. Otherwise,

the SCCS likelihood given in Chapter 2 is no longer valid. For example, if the event of interest, such as stroke, increases the mortality rate, this assumption is violated. Farrington et al. (2009) developed a method which relaxes this assumption when the event of interst is non-reccurrent. This issue limits the use of the SCCS design with recurrent events in many applications. Therefore, we will discuss this issue as a future work. Another limitation of the SCCS design is that it only provides estimates of relative incidence, while the estimate of the absolute incidence cannot be obtained. Since in many applications the relative incidence rate is the main objective, this is not a major limitation.

Another important problem is to investigate in which situations a SCCS design would be preferable to other outcome-dependent sampling designs that use a sample of controls as well, such as nested case-control design. Therefore, as a future work we will investigate this issue with comparing the statistical efficiency of the SCCS design with that of the nested case-control design.

Bibliography

Aalen, O., Borgan, O., Gjessing, H., 2008. Survival and Event History Analysis: A Process Point of View. Springer Science and Business Media.

Andersen, P.K., Borgan, O., Gill, R.D., Keiding, N., 1993. Statistical Models Based on Counting Processes. Springer Verlag, New York.

Andersen, P.K., Gill, R.D., 1982. Cox's regression model for counting processes: A large sample study. The Annals of Statistics , 1100–1120.

Asmussen, S., Glynn, P.W., 2007. Stochastic Simulation: Algorithms and Analysis. volume 57. Springer Science and Business Media.

Atkinson, W., 2011. Epidemiology and Prevention of Vaccine Preventable Diseases (12 ed.). Public Health Foundation.

Atkinson, W., 2012. Mumps Epidemiology and Prevention of Vaccine Preventable Diseases (12 ed.). Public Health Foundation.

Becker, N.G., Salim, A., Kelman, C.W., 2006. Analysis of a potential trigger of an acute illness. Biostatistics (Oxford, England) 7(1), 16–28.

Black, C., Kaye, J.A., Jick, H., 2003. Mmr vaccine and idiopathic thrombocytopaenic purpura. British journal of clinical pharmacology 55(1), 107–111.

Borgan, O., Goldstein, L., Langholz, B., 1995. Methods for the analysis of sampled cohort data in the cox proportional hazards model. The Annals of Statistics , 1749–1778.

Cook, R.J., Lawless, J.F., 2007. The Statistical Anlaysis of Recurrent Events. Springer, New York.

Cox, D.R., 1972. Regression models and life tables. Journal of the Royal Statistical Society , 187–220.

Cox, D.R., Lewis, P.A.W., 1966. The Statistical Analysis of Series of Events. London: Chapman and Hall.

Daley, D.J., Vere-Jones, D., 2003. An Introduction to the Theory of Point Processes. volume 1 of *Elementary Theory and Methods*. Springer, New York.

Douglas, I.J., Smeeth, L., 2008. Exposure to antipsychotics and risk of stroke: Self controlled case series study. British Medical Journal 337, a12272.

Farrington, C.P., 1995. Relative incidence estimation from case series for vaccine safety evaluation. Biometrics 51(1), 228–235.

Farrington, C.P., Anaya Lzquierdo, K., Whitaker, H.J., Hocine, M.N., Douglas, I., Smeeth, L., 2011. Self controlled case series analysis with event-dependent observation periods. Journal of the American Statistical Association 106(494), 417–426.

Farrington, C.P., Hocine, M.N., 2010. Within individual dependence in self controlled case series models for recurrent events. Journal of the Royal Statistical Society.Series C (Applied Statistics) 59(3), 457–475.

Farrington, C.P., Whitaker, H.J., 2006. Semiparametric analysis of case series data. Journal of the Royal Statistical Society.Series C (Applied Statistics) 55(5), 553–594.

Farrington, C.P., Whitaker, H.J., Hocine, M.N., 2009. Case series analysis for censored, perturbed, or curtailed post event exposures. Biostatistics 10(1), 3–16.

Grandell, J., 1997. Mixed Poisson Processes. volume 77. CRC Press.

Grosso, A., Douglas, I., MacAllister, R., Petersen, I., Smeeth, L., Hingorani, A.D., 2011. Use of the self-controlled case series method in drug safety assessment. Expert Opinion on Drug Safety 10(3), 337–340.

Hocine, M., Guillemot, D., Tubert Bitter, P., Moreau, T., 2005. Testing independence between two poisson generated multinomial variables in case series and cohort studies. Statistics in Medicine 24(24), 4035–4044.

Hu, X.J., Lawless, J.F., 1996. Estimation of rate and mean functions from truncated recurrent event data. Journal of the American Statistical Association 91(433), 300–310.

Junghanss, T., 2013. Manson's Tropical Diseases. Oxford: Elsevier Saunders.

Keogh, R.H., Cox, D.R., 2014. Case Control Studies. volume 4. Cambridge University Press.

Kingman, J.F., 1993. Poisson Processes. volume 3. Oxford University Press.

Kuhnert, R., Hecker, H., Poethko, M.C., Schlaud, M., Vennemann, M., Whitaker, H.J., Farrington, C.P., 2011. A modified self controlled case series method to examine association between multidose vaccinations and death. Statistics in Medicine 30, 666–677.

Lambert, N., Strebel, P., Orenstein, W., Icenogle, J., Poland, G.A., 2015. Rubella. Lancet 385(9984), 2297 – 2307.

Lawless, J.F., 1987. Regression methods for poisson process data. Journal of the American Statistical Association 82, 808–815.

Lawless, J.F., 2003. Statistical Models and Methods for Lifetime Data, Second Edition. John Wiley and Sons, Inc.

Lewis, P.A.W., Shedler, G.S., 1976. Simulation of nonhomogeneous poisson processes with log linear rate function. Biometrika 63(3), 501–505.

Li, B., Huang, X., 2006. Existence and uniqueness of relative incidence estimates in case series analysis. Computational Statistics and Data Analysis 50(7), 1807–1817.

Miller, E., Waight, P., Farrington, C.P., Andrews, N., Stowe, J., Taylor, B., 2001. Idiopathic thrombocytopenic purpura and mmr vaccine. Archives of disease in childhood 84(3), 227–229.

Musonda, P., 2006. The self controlled case series method: Performance and design in studies of vaccine safety. PhD Thesis, The Open University .

Musonda, P., Hocine, M.N., Andrews, N.J., Tubert Bitter, P., Farrington, C.P., 2008a. Monitoring vaccine safety using case series cumulative sum charts. Vaccine 26(42), 5358–5367.

Musonda, P., Hocine, M.N., Whitaker, H.J., Farrington, C.P., 2008b. Self controlled case series analyses: Small sample performance. Computational Statistics and Data Analysis 52(4), 1942–1957.

Musonda, P., Paddy Farrington, C., Whitaker, H.J., 2006. Sample sizes for self controlled case series studies. Computational Statistics and Data Analysis 25(15), 2618–2631.

Oakes, D., Cui, L., 1994. On semiparametric inference for modulated renewal processes. Biometrika 81(1), 83–90.

Prentice, R.L., 1986. A case cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika 73(1), 1–11.

Rigdon, S.E., Basu, A.P., 2000. Statistical Methods for the Reliability of Repairable Systems. New York: Wiley.

Simpson, S.E., 2013. A positive event dependence model for self controlled case series with applications in postmarketing surveillance. Biometrics 69(1), 128–136.

Vines, S.K., Farrington, C.P., 2001. Within subject exposure dependency in case crossover studies. Statistics in Medicine 20(20), 3039–3049.

Weldeselassie, Y.G., Whitaker, H.J., Farrington, C.P., 2011. Use of the self-controlled case-series method in vaccine safety studies: Review and recommendations for best practice. Epidemiology and Infection 139(12), 1805–1817.

Whitaker, H.J., Farrington, C.P., Spiessens, B., Musonda, P., 2006. Tutorial in biostatistics: The self-controlled case series method. Statistics in Medicine 25(10), 1768–1797.

Whitaker, H.J., Hocine, M.N., Farrington, C.P., 2007. On case crossover methods for environmental time series data. Environmetrics 18(2), 157–171.

Whitaker, H.J., Hocine, M.N., Farrington, C.P., 2009. The methodology of self-controlled case series studies. Statistical Methods in Medical Research 18(1), 7–26.

Xu, S., Hambidge, S.J., McClure, D.L., Daley, M.F., Glanz, J.M., 2013. A scan statistic for identifying optimal risk windows in vaccine safety studies using self controlled case series design. Statistics in medicine 32(19), 3290–3299.

Xu, S., Zhang, L., Nelson, J.C., Zeng, C., Mullooly, J., McClure, D., Glanz, J., 2011. Identifying optimal risk windows for self controlled case series studies of vaccine safety. Statistics in Medicine 30(7), 742–752.

Appendix

ITP data set: indiv is serial number of individual. itp is event time. start and stop is start and stop time of following up time. mmr is the start time of exposure period.

indiv	itp	start	stop	mmr
1	691	453	730	670
2	722	365	730	868
3	442	365	730	540
4	429	365	730	378
5	414	365	730	710
5	418	365	730	710
6	708	438	730	487
7	615	365	730	461
8	463	365	730	526
9	440	365	730	529
9	473	365	730	529
10	477	365	730	458
11	396	365	730	374
12	676	365	730	428
13	480	365	730	446
14	633	365	730	423
15	403	365	730	365
16	419	365	730	369
16	443	365	730	369
17	553	365	730	889
17	666	365	730	889
18	705	365	730	389
19	419	365	730	389
20	402	365	730	385
21	406	365	730	458
22	494	365	730	468
23	374	365	730	819
23	389	365	730	819
23	452	365	730	819
23	522	365	730	819
23	564	365	730	819
24	598	365	730	430
25	409	365	730	384
26	612	365	730	398
27	381	365	723	427
28	438	365	730	427
29	425	365	677	647
30	543	365	677	422
31	609	365	674	860
32	412	365	730	387
33	407	365	730	396
34	484	365	730	408
34	623	365	730	408
35	411	365	730	383