STATISTICAL ANALYSIS FOR SURVIVAL DATA WITH MISSING INFORMATION

A Dissertation

Presented to

the Faculty of the Graduate School

University of Missouri

In Partial Fulfillment

Of the Requirements for the Degree

Doctor of Philosophy

---

by

BIN ZHANG

Dr. (Tony) Jianguo Sun, Dissertation Supervisor

JULY, 2009

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

STATISTICAL ANALYSIS FOR SURVIVAL DATA
WITH MISSING INFORMATION

presented by Bin Zhang

A candidate for the degree of Doctor of Philosophy

And hereby certify that in their opinion it is worthy of acceptance.

Dr. (Tony) Jianguo Sun ______________________________

Dr. Nancy Flournoy ______________________________

Dr. Athanasios C. Micheas ______________________________

Dr. Min Yang ______________________________

Dr. Xiaoguang Ni ______________________________

Dedicated to my family,

Xiying Zhang, Junxin Gao, Jingjing Zhang and Jing Zhang

ACKNOWLEDGEMENTS

Here I would like to express my deep and sincere gratitude to my advisor Dr. Tony Sun. It is Dr. Sun who told me what a good researcher should do and what good research should be. He showed an amazing world of statistical research which I never imagined. His wonderful guidance, generous support, and endless help enable me to complete this work. His encouragement and instruction make this work possible. Without his unselfish dedication I could never achieve the success today.

I extend my gratitude to the members of my advisory committee: Dr. Nancy Flournoy, Dr. Min Yang, Dr. Athanasios Micheas and Dr. Xiaoguang Ni for their insightful comments and suggestions on my work. I also want to thank Dr. Xingwei Tong for his discussion and help.

I want to thank our department for offering us such a wonderful opportunity studying here. Thank our fantastic faculty for providing so many great courses. Thank our great staff Judy and Tracy for their kind help during the last four years. I also want to thank my friends in this department. They make my PhD life full of fun.

I am grateful to my beloved wife, Jing Zhang, for her love and strong support throughout my time at school.

Most importantly, I am forever indebted to my family, my parents Xiying Zhang, Junxin Gao and my sister Jingjing Zhang, for their understanding and moral support. It is to them that I dedicate this work.

STATISTICAL ANALYSIS FOR SURVIVAL DATA WITH MISSING INFORMATION

BIN ZHANG

Dr. (Tony) Jianguo Sun, Dissertation Supervisor

ABSTRACT

In survival analysis, the random variables of interest are the times to certain events. The occurrence of an event is usually referred to as a failure. While in practical problems, sometimes the failure time can not be recorded exactly or some information might be missing. Due to the different structures, missing information can be classified into two types. The first type is caused by the censoring scheme. For example, in AIDS or cancer studies, the failure time is often not observed directly, but only known to lie within an interval. This may happen, for example, when a subject misses one or more visits in a medical study with periodic follow-ups. This type of data is usually referred to as interval-censored data. The other type of missing is caused by the sampling structure. For example, we are interested in the relationship between hypertension and weight. If the scale of a researcher can only measure the weight from 0 to 300 pounds, then only the people who are not heavier than 300 pounds can be selected. This dissertation discusses the statistical analysis for survival data with missing information,

it is organized as following.

Chapter 1 provides some basic concepts and commonly used models in survival analysis. Several examples are provided to illustrate different types of failure time and missing data. The existing methods for analyzing the interval-censored data are reviewed. Also the biased sampling models with their applications will be introduced.

Chapter 2 discusses efficient estimation for the linear transformation models with current status data. It is well-known that in survival analysis, the proportional hazards model (Cox model) may not be appropriate for modeling failure times in some applications. We propose regression analysis for current status data using the linear transformation models. The main advantage of the linear transformation models is their flexibility. Several authors have considered the fitting of the linear transformation models to right-censored data. However, in those approaches, the estimators are not efficient. We will fit these models to current status data. We also derive the maximum likelihood estimates and their information bound and establish the efficiency of the estimates. Simulation studies show that the presented approach works well for practical use and an application of our methodology to a tumorigenicity study will be given.

Chapter 3 considers efficient estimation for the proportional odds model with bivariate current status data. The analysis about bivariate current status data has draw much attention recently. Due to the complexity of the censoring scheme and the difficulties in modeling the correlation between the two failure times, only few work has been done for this type of data. We develop an efficient estimator for the situations that the marginal distributions of the failure times follow proportional odds model. To model the correlation between two failure times, we employ the copula models as

the joint survival function of two related failure time variables. Simulation studies are conducted to show the good performance of this method and the methodology is applied to a set of data from a tumorigenicity experiment.

Chapter 4 studies the biased sample problem with empirical likelihood method. The information about the parameters β of an unknown distribution F is given by unbiased estimating equations. The likelihood ratio statistic for the biased sample problem is proved to follow a chi-square distribution asymptotically. This statistic can be used to do the hypothesis test or to compute the confidence intervals of the parameters. Simulation studies indicate that the estimates of parameters derived perform well, the large sample properties of the statistic are also supported by the simulation results.

Some discussion and future work can be found in Chapter 5.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

As a branch of statistics, survival analysis, which is often referred to as “reliability theory” in engineering, has a long history. In early 1960’s, life table methods were already used to analyze the survival data. But the bridge between the classical life table methods and the modern regression analysis is due to the Cox’s paper (1972) on the proportional hazards model, which stimulated the research activity and revolutionized the practice of medical statistics. After that, more and more models and methods were proposed. Survival analysis has been well developed in the past 35 years and it is widely used not only in medicine but also in epidemiology, insurance, sociology, engineering, biology, environmental science, manufacturing, education, etc..

This dissertation discusses statistical analysis for survival data with missing information. The remaining of this chapter is organized as follows. Section 1.1 introduces the failure time data and the types of missing information. Section 1.2 focuses on the types of interval censoring and gives some examples. The multivariate case is discussed at the end of this section. Section 1.3 presents some commonly used regression models which play a very important role in the sequel. The models and properties are all demonstrated in this section. In Section 1.4, we review some existing methods for analyzing current status failure time data. The literatures about the regression methods for current status data are reviewed and several examples are shown after that. Section 1.5 introduces the definition of biased sampling. Three examples are provided and existing analysis methods are summarized. Section 1.6 gives the outline of this dissertation.

1.1 Failure Time and Missing Data

When we talk about failure time data, we mean the data concern some random variables which represent the time to some event. The occurrence of the event is usually referred to as a failure. For example, Pearl (1938) investigated the effects of tobacco on human longevity among white males over the age of 30. In this study, the individuals were divided into three categories: nonsmokers, moderate smokers and heavy smokers. Comparisons of the longevity of the three categories were made. In this case, the death time is of interest, thus it is the failure time. However, sometimes the individual may not develop such failure at the end of the study, such individuals are referred to as censored observations. For the censored subjects, the failure times are not observed directly. Sometimes a subpopulation do not have a chance to be selected because of the sampling structure. Unlike the censored data, the missing information in this case is the response. The types of missing information will be discussed later in this section. Before we discuss them in detail, an example is given below to illustrate failure time data.

1.1.1 Acute Leukemia Study

This leukemia study was presented by Freireich et al. (1963) and Gehan (1965). In this study, 42 acute leukemia patients were assigned to two treatment groups. One was the drug 6-mercaptopurine (6-MP) group and the other one was the placebo group. After one-year period, the remission times in weeks were collected and shown in Table 1.1. The purpose of this study is to detect whether the remission times of the two treatment groups are different.

In Table 1.1, the numbers with stars are censoring times, which means the times of those individuals were censored because those patients were still in the state of remission at the end of the study. Other times were exactly observed. Each number is the time from the patient entered the study to the end of the study. This type of censoring is known as right-censoring, which is a very common situation in survival analysis. Freireich et al. (1963), Gehan (1965) and Kalbfleisch and Prentice (2002) analyzed and discussed this data set in details.

1.1.2 Types of Missing Information

At the beginning of this section, we talked about the possible missing information in practical problems. There are two types of missing. The first one is caused by the censoring scheme. The missing information is the exact failure time. The other type is caused by the sampling structure. The missing information is the response. We explain each type in details.

For censoring, as we have seen in the previous example in Section 1.1.1, the failure time may not always be observed exactly. Instead, the information we may only know is whether the failure time is grater than or less than some censoring time or it falls between two time points. When the failure time is only known to be equal to or larger than a censoring time, we call it right-censored failure time data. The typical case is that the failure is not developed at the end of the study. While if the failure time is known only to be less than a certain time, we call it left-censored failure time. Interval-censored failure time data arise when the subjects in the study are not continuously observed. Instead, the time is only known to lie within an interval. For example, when a medical or health study with periodic follow-ups. An individual who is monitored weekly or monthly for a clinically observed change ("response") may miss one or more prespecified visits and return with a changed state, thus we only know the failure developed between the two visits, contributing an "interval-censored" observation. This type of data occurs in many fields such as medical, epidemiological and financial studies. Hoel and Walburg (1972) and Finkelstein (1986) provided examples from animal carcinogenicity and epidemiology studies. More examples can be found in Kim et al. (1993), Jewell, Malani and Vittinghoff (1994), Goggins and Finkelstein (2000).

For sampling procedure, sometimes not all the subjects in the population have the chance to be selected, the outcome of a subpopulation might be missing. It happens not only in survival analysis, but also in economics, survey sampling, etc.. For example, if we are interested in the relationship between the wages and the education levels of people. Then only the people who have jobs will have the chance to be selected. More generally, if the subjects in the population do not have an equal chance to be selected, instead the items are observed with different probabilities that depend on the outcome. This type of sampling problems are often referred to as "biased sample problems". For example, in the case-control study, since the different probabilities of disease and non-disease. The diseased individuals and the diseased-free individuals do not have the same chance to be selected. More applications can be found in Vardi (1982, 1985) and Gilbert (1996) among others.

1.2 Interval Censoring

This section will focus on interval censoring of the first type of missing information, which is most widely seen in practice.

1.2.1 Case I Interval-Censored Data or Current Status Data

If every subject is observed only once and the occurrence of the failure is recorded, then the resulting data are usually referred to as case I interval-censored data (Groeneboom and Wellner 1992) or current status data .

Table 1.2 presents the data introduced in Hoel and Walberg (1972) about the lung tumor experiment for 144 male RFM mice. In this experiment, 144 male RFM mice were assigned to two treatment groups, the conventional environment (CE) which contained 96 mice and the germ-free environment (GE) which contained 48 mice. For each mouse, only the death time or sacrifice time (in days) was observed. An indicator was used to denote the status of lung tumor at the time of death. The indicator is 1 when lung tumor present, it is 0 when lung tumor absent. The interest of this experiment is to determine whether the time until the tumor developed differs between the animals in the two environments. In this data set, only the death times (in days) for the mice are known with the status of the tumor at the death time (with tumor or no tumor). By the definition above, this data are case I interval-censored or current status data.

1.2.2 Case II Interval-Censored Data

For case II or general interval-censored data, we only know the failure occurs within one interval. Both examine time points of the interval belong to $(0, \infty)$. Let us use T to denote the failure time of interest, U and V are the two examine time points, where $U \leq V$. Then the observations of case II interval-censored data have the form

$$\{U, V, \delta_1 = I(T \leq U), \delta_2 = I(U < T \leq V), \delta_3 = I(T > V)\},$$

where $\delta_i, i = 1, 2, 3$ is an indicator. Case I interval censoring can be expressed by the above form by taking $U = V$.

Finkelstein and Wolfe (1985) gave a set of data from a study on early breast cancer patients at Joint Center for Radiation Therapy in Boston between 1976 and 1980. This study contains 94 patients who were divided into two treatment groups, 46 of them were given radiation therapy (RT) alone, other 48 were assigned to the group receiving radiation therapy plus adjuvant chemotherapy (RCT). The data are shown in Table 1.3. Although patients were supposed to be seen at clinic visits every 4 to 6 months, the real visit times and the time between two visits are different for the patients. At each visit, the cosmetic appearance of the patient such as breast retraction was evaluated. The purpose of this study is to evaluate the difference between the two treatments with respect to their cosmetic effects. The data in Table 1.3 gave the time to breast retraction. The intervals without right end points mean those patients did not experience breast retraction during the whole study.

1.2.3 Bivariate and Multivariate Interval-Censored Data

Multivariate interval-censored data occur when two or more failure times are of interest and each time variable has a structure of interval censoring. The failure times usually depend on one another. The data are usually referred to as bivariate interval-censored data when only two failure times are of interest. Bivariate analysis is often used to investigate the dependence between two variables.

Wang and Ding (2000) analyzed the data based on a study from 1991 to 1993 which was designed to determine risk factors for cardiovascular diseases in two towns, Chu-Dung and Pu-Tze, in Taiwan. 6314 people were involved in this study, 2904 males and 3410 females. The individual's age at the censoring time was measured with the indicators of three diseases, hypertension, diabetes mellitus and hypercholesterolaemia. Since only the censoring time and whether the three diseases occurred before or after the censoring time are known, this data are multivariate current status data. Ding and Wang (2004) provided a test of pairwise independence of the three diseases. It showed that the association between each pair is very strong. Another example can be found in Goggins and Finkelstein (2000). They discussed a data set from an AIDS clinical trial, ACTG 181. It is a substudy of a comparative clinical trial of three anti-pneumocystis drugs and concerns the opportunistic infection cytomegalovirus (CMV). During the study, 204 patients provided at least one urine and blood samples. For some subjects, the shedding times are left-censored because the shedding had already occurred before these patients entered the study. While some are right-censored because the shedding did not occur at the end of the study. Some shedding times were just interval-censored. We are interested in the two correlated failure times and both of them are interval-censored. It is an example of the general case of bivariate interval censoring.

1.3 Some Regression Models for Failure Time Data

Regression models are commonly used in analyzing covariate effects. Several widely used continuous semiparametric regression models will be described below, which include the proportional hazards model or Cox model, the proportional odds model and the linear transformation model. Before we introduce these regression models, several basic concepts will be given first.

In this section, we use T to denote failure time which is a nonnegative random variable. Let $S(t)$ denote the survival function of T which is defined as the probability that T exceeds a value t (Sun 2006). We have

$$S(t) = P(T > t),\ 0 < t < \infty.$$

Besides survival function, hazard function, probability density and distribution function are often used. While the definition varies for continuous and discrete T.

First assume T is absolutely continuous, then the probability density function $f(t)$ is defined as

$$f(t) = -\frac{dS(t)}{dt},$$

which gives

$$S(t) = \int_t^\infty f(s)ds.$$

The hazard function $\lambda(t)$ is the probability that the failure happens at time t given the failure has not developed before time t. It has the following from

$$\lambda(t) = \lim_{\Delta t \to 0+} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}.$$

Since

$$\lim_{\Delta t \to 0+} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \lim_{\Delta t \to 0+} \frac{P(t \leq T < t + \Delta t)}{\Delta t} \frac{1}{P(T \geq t)} = \frac{f(t)}{S(t)} = -\frac{1}{S(t)} \frac{dS(t)}{dt},$$

we have

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{d \log S(t)}{dt}.$$

Let $\Lambda(t) = \int_0^t \lambda(s)ds$, which is the cumulative hazard function of T. Then we have the following relationships

$$S(t) = e^{-\int_0^t \lambda(s)ds} = e^{-\Lambda(t)},$$

$$f(t) = \lambda(t)e^{-\Lambda(t)}.$$

Now suppose T is discrete with the support points $s_1 < s_2 < \ldots$ and the probability function $f(s_i) = P(T = s_i), i = 1, 2, \cdots$. Then by the definition, we have

$$S(t) = \sum_{i:t<s_i} f(s_i), 0 < t < \infty.$$

Then the hazard of T at s_i is defined as

$$p_i = P(T = s_i | T \geq s_i) = \frac{f(s_i)}{S(s_{i^-})} \qquad i = 1, 2, \cdots.$$

In the discrete case, we have the following relationship

$$S(t) = \prod_{i:t\geq s_i} (1 - p_i)$$

and

$$f(s_i) = p_i \prod_{j=1}^{i-1} (1 - p_j).$$

In the rest of this section we will discuss several regression models commonly used in survival analysis.

1.3.1 The Proportional Hazards Model

The proportional hazards model (PH), also known as Cox model, is named after D. R. Cox (1972). Let Z be the covariates. The PH model or Cox model assumes the covariates affect the failure time through the following equation

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

where $\lambda_0(t)$ is an arbitrary unspecified baseline function and β is the vector of regression parameters. From the definition, the ratio of the hazard functions for two subjects with different covariates is constant, in another word, the two hazard functions are proportional to each other.

Under the PH model (1.1), the conditional density and survival functions of T given Z satisfy

$$f(t;Z) = \lambda_0(t)e^{Z'\beta}e^{-\Lambda_0(t)\exp(Z'\beta)}$$

and

$$S(t;Z) = e^{-\Lambda_0(t)\exp(Z'\beta)} = S_0(t)^{\exp(Z'\beta)},$$

where $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$ is the baseline cumulative hazard function and $S_0(t) = e^{-\Lambda_0(t)}$ is the unknown baseline survival function which denotes the survival function for the subjects with $Z = 0$.

The proportional hazards model may be the most widely used regression model in survival analysis. It was first proposed by Cox (1972, 1975) and has been studied by many authors. Anderson and Gill (1982) analyzed the PH model for counting processes. Chen and Little (1999) discussed the PH model with missing covariates. Struthers and Kalbfleisch (1986) and Anderson and Fleming (1995) proposed methods for detecting model misspecification in Cox model. Cox and Oakes (1984), Kalbfleisch and Prentice (2002) and Sun (2006) provide more references and applications for the proportional hazards model in right-censored failure time data. There are also a lot of authors who applied Cox model to interval-censored failure time data. Finkelstein (1986) used maximum likelihood estimation and tested the covariate effect. Huang (1996) provided an efficient estimation for PH model with interval censoring. Farrington (2000) gave a residual analysis and Sun (1999) proposed a nonparametric test for current status data with unequal censoring.

1.3.2 The Proportional Odds Model

Another commonly used model is the proportional odds model. For a given Z, the proportional odds model can be expressed as

$$\frac{S(t;Z)}{1-S(t;Z)} = e^{Z'\beta}\frac{S_0(t)}{1-S_0(t)} \tag{1.2}$$

or

$$\text{logit}[S(t;Z)] = \text{logit}[S_0(t)] + Z'\beta,$$

where $S_0(t)$ is again the baseline survival function and $\text{logit}(x) = \log(\frac{x}{1-x})$.

Unlike the proportional hazards model, the ratio of the hazard functions for two subjects with different covariates are no longer a constant, i.e. the two hazard functions are not proportional to each other. However, for model (1.2), the odds of the survival functions for two subjects with different covariates are proportional to each other. Assuming $Z = 0$ or 1, under the model (1.2), we have

$$\frac{\lambda(t; Z = 1)}{\lambda(t; Z = 0)} = \frac{1}{1 + (e^{\beta} - 1)S_0(t)}.$$

The literatures for the applications of the proportional odds model are abundant. For right-censored data, Murphy et al. (1997) studied the maximum likelihood estimation. Yang and Prentice (1999) proposed several classes of regression estimators. Chen (2001) applied the proportional odds model to fit data from case-cohort design by using the weighted likelihood. While for interval-censored data, Rossini and Tsiatis (1996) used the MLE to analyzed the proportional odds model for current status data by treating the baseline log-odds function as a step function. Huang and Rossini (1997) studied the MLE of the proportional odds model for interval-censored data by using the sieve estimator for the baseline odds function, which was treated as a piece-wise linear function. Rabinowitz et al. (2000) fitted the model to interval-censored data by using conditional logistic regression. As an alternative to the Cox model, the proportional odds model plays a very important role in survival analysis.

1.3.3 The Linear Transformation Model

For a given baseline function, both the proportional hazards model and the proportional odds model are singly specified models. Now a class of regression models will be introduced which are called the linear transformation model. Let $H(t)$ be an unknown strictly increasing function. Then the linear transformation model is defined as below

$$H(T) = -Z'\theta + \epsilon, \tag{1.3}$$

where θ is the vector of regression parameters and ϵ is a random variable which follows a known distribution function F. Actually, both the PH model and the proportional odds model are special cases of the linear transformation models. By taking $F(t) = 1 - \exp[-\exp(t)]$, the extreme value distribution, (1.3) gives the PH model. The proportional odds can be obtained if F is the standard logistic distribution. The linear transformation model is general because F in the model can be any distribution function. In practice, the PH model or the PO model may not be appropriate (Lin and Ying 1994 and Jin et al. 2003), the linear transformation model is a good choice.

Chen et al. (2002) and Cheng et al. (1995) developed some estimation procedures for fitting the linear transformation models to right-censored failure time data. Fine et al. (1998) modified Cheng's estimation procedures for regression parameters and provided a new prediction procedure for the survival probabilities of future subjects. Kong et al. (2004) used weighted estimating equations to analyze the censored data from a case-cohort design. Lu and Ying (2004) applied the linear transformation model with a cure rate.

1.4 Review of the Existing Methods for Current Status Data

1.4.1 Analysis of Univariate Current Status Data

The earliest work on nonparametric likelihood estimation (NPMLE) with current status data was given by Ayer et al. (1955). They introduced the pool-adjacent-violators algorithm to compute the NPMLE of a distribution function. Groeneboom (1987) and Groeneboom and Wellner (1992) provided the asymptotic properties of the NPMLE. After that, more semiparametric methods were developed to current status data. For example, Huang (1996) showed the MLE for the regression parameter of the PH model is efficient and asymptotically normal. Rossini and Tsiatis (1996) used the sieve estimators for the PO model and proved the estimators are normal and efficient. Martinussen and Scheike (2002) proposed an efficient estimation for the additive hazards model. Sun and Sun (2005) considered fitting the linear transformation model with the estimating equation approach.

1.4.2 Analysis of Multivariate Current Status Data

1.4.2.1 Copula Models for Bivariate Data

As described in Section 1.2.3, bivariate current status data arise when two dependent failure times are of interest and each time has a structure of case I interval censoring. Clayton (1978) first discussed the association for bivariate failure time data.

In 1986, Genest and Mackay introduced the general form of copula models as follows.

Let Φ be a class of function $\phi : [0,1] \rightarrow [0,\infty]$ that has two continuous derivatives on $(0,1)$ and satisfy

$$\phi(1) = 0, \quad \phi'(t) < 0, \quad \phi''(t) > 0$$

for all $0 < t < 1$. Thus ϕ has an inverse ϕ^{-1}. For every function in class Φ, copula model is defined as the following bivariate distribution function for a pair of random variables (X, Y)

$$H(X,Y) = \begin{cases} \phi^{-1}[\phi(x) + \phi(y)] & \text{if } \phi(x) + \phi(y) \leq \phi(0) \\ 0 & \text{otherwise.} \end{cases} \tag{1.4}$$

As an example, let $\phi(t) = (t^{-\alpha} - 1)/\alpha$, the copulas given by (1.4) become the well known Clayton model (Clayton 1978 and Oakes 1982).

Let T_1, T_2 be the two failure times of interest and suppose $S_1(t)$ and $S_2(t)$ are the marginal survival functions for T_1 and T_2 respectively. To measure the association between T_1 and T_2, we employ the copula model

$$S(s,t) = C_\alpha(S_1(s), S_2(t))$$

as the joint survival function for (T_1, T_2), where $C_\alpha(u,v) : [0,1]^2 \rightarrow [0,1]$.

In survival analysis, copula models are often used to provide a relationship between the joint distribution and the marginal distributions. Ding and Wang (2004) checked the independence for bivariate current status data by using Clayton's and Frank's family. Wang (2003) estimated the association parameter for copula model under dependent censoring. Wang et al. (2008) compared the efficient estimations of different copula models for bivariate current status data.

1.4.2.2 Regression Analysis of Multivariate Data

Marginal approach is a very important method for the analysis of multivariate failure time data. It uses the marginal likelihood functions directly without considering the dependence structure among the failure times. Cai and Prentice (1995) used weighted partial likelihood score equations for the inference about the regression parameters. Dunson and Dinse (2002) proposed Bayesian models for multivariate current status data with informative censoring. Guo and Lin (1994) provided the regression analysis of multivariate grouped data. Wei et al. (1989) modeled the marginal distributions with the PH model. Goggins and Finkelstein (2000) applied the method for a data set from an AIDS observational study that was conducted by the AIDS Clinical Trails Group (the ACTG).

Another well known method for analyzing multivariate data is the random effect model. This approach assumes there exists a common and unobserved latent random variable and the correlated failure times are independent given this latent variable. The latent variable, also known as frailty, models the dependence of the failure times. Oakes (1989) applied frailty model for bivariate failure time data. Huang and Wolf

(2002) used this model to deal with informative censoring. But the literatures about the multivariate analysis are still limited due to the difficulties of modeling the structure of the correlation among the failure times.

1.5 Biased Sampling

In this section, we discuss several common sampling problems in biostatistics. When an investigator records an observation, the recorded observations will not have the same original distribution unless every observation is given an equal chance of being recorded. But in practical problems, this assumption is often violated. For a simple example, suppose two scientists, independently of each other, are recording measurements of a certain natural phenomenon whose cumulative distribution function (cdf) is F. The first scientist, because of limited experimental conditions, can observe the phenomenon only in the range 10 to 20. Outside this range, even the phenomenon occurred, it can not be recorded. He reports his measurements to be 13, 16, 18. The second scientist can observe the phenomenon throughout its entire range and he reports his measurements to be 8, 14, 17, 23. These two sets of measurements are independent, but obviously not identical. The question is how to combine the two samples to make inference. Biased sampling scheme arise when the items are observed with different probabilities that depend on the outcome. It is frequently used in biostatistics, economics, survey sampling, etc. The biased sampling problems can be described as following. Let Y be a real-valued random variable with cdf $F(Y)$. In biased sampling situation we do not observe Y_i's independent and identically-distributed (i.i.d.) $F(Y)$, but instead we observe $Y_1, \ldots, Y_n$ from G, where G is a distribution from biased sampling of F according to some known biasing or weight function w. If w is a nonnegative function, then

$$G(y) = \frac{\int_{-\infty}^{y} w_i(t)dF(t)}{\int_{-\infty}^{\infty} w_i(t)dF(t)} = \int_{-\infty}^{y} \frac{w_i(t)}{W} dF(t), \tag{1.5}$$

where $W = \int w(t)dF(t)$. The following will provide three very widely used biased sampling problems.

1.5.1 Length-Biased Sampling

Suppose that F is a cdf on $R^+ = [0, \infty)$ with positive finite mean $\mu = \int_0^\infty ydF(y)$. Let $w(y) = y$, then $W = \mu$ and (1.5) becomes

$$G(y) = \frac{1}{\mu} \int_0^y tdF(t).$$

This is the length-biased distribution corresponding to F. Because the selection probability depends on the actual values y of the sampled items, it makes the estimation of F a nonstandard problem. This problem was first studied by Vardi (1982) by assuming two samples are selected with $w_1(t) = 1$ and $w_2(t) = t$. He provided a nonparametric maximum likelihood estimate (NPMLE) of two-sample problem in the presence of length bias and proved the estimate $\hat{F}$ converges weekly to a Gaussian process. Vardi (1985) generalized the problem to several independent samples, the NPMLE was derived and shown to be asymptotically efficient. Both assume that the weight functions are completely known. A good survey of real-life application of length biased distributions can be found in Patil and Rao (1977).

1.5.2 Stratified Sampling

Suppose $\{D_1, \ldots, D_s\}, s \geq 3$, is a partition of sample space $Y = R^1$ with $\bigcup_{i=1}^s D_i = Y = R^1$ and $D_i \bigcap D_j = \emptyset$ for $i \neq j$. If the weight functions in the biased sampling model are $w_i(y) = I_{D_i}(y)$ for $i = 1, \ldots, s$, then

$$G_i(y) = F(y|D_i) = \frac{F\left((-\infty, y] \bigcap D_i\right)}{F(D_i)}.$$

It is just the conditional distribution given the event D_i. This is stratified sampling from the strata $D_1, \ldots, D_s$. The estimation of F itself is impossible without the knowledge of the stratum probabilities $F(D_i)$. This special case of biased sampling model corresponds to stratified sampling in the survey sampling literature. A standard assumption is that the stratum sizes $N_i, i = 1, \ldots, s$ are known (e.g. Cochran 1963). If $s+1$ samples are drawn from the $s+1$ distributions $(G_1, \ldots, G_{s+1}) = (G_1, \ldots, G_s, F)$ with weight functions $(w_1, \ldots, w_{s+1}) = (I_{D_1}, \ldots, I_{D_s}, 1)$, then we can use the last sample to estimate $F(D_i)$. This type of model is often referred to as “Enriched stratified sampling”. This problem was studied by Vardi (1985) and Gill, Vardi and Wellner (1988), among others.

1.5.3 Case-Control Study

Case-control design is very useful in biostatistics, it consists of a sample of cases (i.e., diseased individuals) and a sample of controls (i.e., disease-free individuals). To analyze binary data which arise in studying relationships between diseases and environment, logistic regression models are commonly used. Let y be a binary response variable which denote the status of disease and x be the covariate. Then

$$P(y = 1|x) = \frac{\exp(\alpha^* + x\beta)}{1 + \exp(\alpha^* + x\beta)} = \psi(x),$$

where β is the parameter and the marginal distribution of x is unspecified. For case-control data, we have two independent groups of sample data as below. Let $x_1, \ldots, x_{n_0}$ be a random sample from $F(x|y = 0)$ and $z_1, \ldots, z_{n_1}$ be a random sample from $F(x|y = 1)$ independently. Denote $n = n_0+n_1$ and let $\{x_1, \ldots, x_{n_0}; z_1, \ldots, z_{n_1}\}$ be the combined sample. If $\pi = P(y = 1) = 1 - P(y = 0)$ and $f(x|y = i) = dF(x|y + i)/dx, i = 0, 1$ represents the conditional density, then by Bayes' rule we have

$$f(x|y = 1) = \frac{\psi(x)}{\pi} f(x), \quad f(x|y = 0) = \frac{1 - \psi(x)}{1 - \pi} f(x),$$

where $f(x)$ is the marginal distribution of x. So

$$\frac{f(x|y = 1)}{f(x|y = 0)} = \frac{1 - \pi}{\pi} \frac{\psi(x)}{1 - \psi(x)}.$$

Let $g(x) = f(x|y = 0)$ and $h(x) = f(x|y = 1)$, we have

$$h(x) = f(x|y = 1) = \frac{1-\pi}{\pi}\frac{\psi(x)}{1-\psi(x)}g(x) = \exp(\alpha + x\beta)g(x),$$

where $\alpha = \alpha^* + \log\{(1 - \pi)/\pi\}$. As a result, we arrive at the two-sample model in which $\{x_1, \ldots, x_{n_0}\}$ and $\{z_1, \ldots, z_{n_1}\}$ are independent and

$$\begin{aligned} x_1, \ldots, x_{n_0} &\overset{iid}{\sim} g(x) \\ z_1, \ldots, z_{n_1} &\overset{iid}{\sim} h(x) = \exp(\alpha + x\beta)g(x). \end{aligned}$$

This is a biased sampling model with weight function $\exp(\alpha+x\beta)$. There are numerous applications for case-control study. Recently a lot of publications on genetic studies have been seen, most of them came from case-control association studies of complex diseases (e.g., Lin and Zeng 2009).

1.6 Outline of the Dissertation

This dissertation contains four parts, including two chapters on semiparametric analysis of current status data, one chapter on the biased sampling problem and one chapter about the future work.

In Chapter 2, we consider the linear transformation model for analyzing univariate current status data. This chapter begins with the introduction of the background and basic concept used for the whole chapter following by the main results of the estimators.

Then simulation studies will be provided to illustrate the good performance of this method. A real data analysis will be shown afterward.

In Chapter 3, bivariate current status data will be studied by modeling the marginal distributions with the proportional odds model. Copula models are used to define the dependence between the two failure times. The efficient sieve estimators are developed by treating the marginal distribution functions as piecewise linear functions.

In Chapter 4, we discuss the biased sample problem with empirical likelihood method with estimating equations. The likelihood ratio statistic for biased sample problem is proved to follow a chi-square distribution asymptotically. Simulation studies indicate that the estimate of parameter derived performs well, the large sample property of the statistic is also supported by the simulation results.

Some discussions and future work are addressed in Chapter 5.

Chapter 2

Efficient Estimation for Linear Transformation Models with Current Status Data

2.1 Introduction

In this chapter we will discuss regression analysis of current status data using the linear transformation model. As mentioned in Section 1.3.3, the main advantage of the linear transformation models is their flexibility since they include many well-known regression models as special cases. For example, one can get the proportional hazards model and the proportional odds model (Cheng et al., 1995) from the linear transformation models as shown in Chapter 1. For regression analysis of failure time data, the proportional hazards model is commonly used and a number of inference approaches have been developed (Kalbfleisch and Prentice, 2002). However, it is well-known that the proportional hazards model may not be appropriate for modeling failure times in some applications and other models are needed. For example, Lin and Ying (1994) and Jin et al. (2003) considered the fitting of the additive hazards model and the accelerated failure time model, respectively, to right-censored data. Chen et al. (2002) and Cheng et al. (1995) developed some estimation procedures for fitting the linear transformation models to right-censored failure time data.

In current status data, each subject is observed only once and the failure time of interest is either left- or right-censored (Keiding, 1991; Sun, 2006). They are a special case of interval-censored failure time data, in which the failure time is observed only to belong to an interval (Sun, 2006), and often occur in, for example, cross-sectional studies and tumorigenicity experiments (Keiding, 1991; Sun and Kalbfleisch, 1996). Several authors have considered the fitting of the linear transformation model to interval-censored data using the estimating equation approach (Sun and Sun, 2005; Younes and Lachin, 1997; Zhang et al., 2005). A main drawback of these approaches is that they may not be efficient. In the following, we consider the maximum likelihood approach and establish its efficiency.

The remainder of this chapter is organized as follows. Section 2.2 introduces the models and assumptions that are used in this chapter. The likelihood approach will be proposed in Section 2.3. The estimators with their information bound will be given. Their large sample properties and main results about the efficiency of the estimators will be provided as well. Also a algorithm will be given at the end of this section. In Section 2.4, simulation results are presented to show that the approach is appropriate for practical use. The method is applied to a real life data and the results are shown in Section 2.5. Some discussions can be found in Section 2.6.

2.2 Models and Assumptions

Similar to the notations as in Chapter 1, let T denote the failure time of interest and Z be a vector of covariates. The linear transformation model assumes that given Z, T follows

$$H_0(T) \,=\, -Z'\theta_0 \,+\, \epsilon \tag{2.1}$$

(Chen et al., 2002; Cheng et al., 1995; Lu and Tsiatis, 2006; Lu and Ying, 2004; Ma and Kosorok, 2005; Yin and Zeng, 2006).

Let $S(t) \,=\, 1 - F(t)$ and $\Lambda(t) \,=\, -\, \log\{S(t)\}$, the survival and cumulative hazard functions of ϵ, respectively. Suppose that $\Lambda(t)$ is twice differentiable and let $\lambda(t) \,=\, d\Lambda(t)/dt$. Then it follows from model (2.1) that the survival function of T has the form

$$S(t|Z) \,=\, \exp[-\Lambda(H_0(t) \,+\, Z'\theta_0)]$$

given Z. As mentioned above, for current status data, each subject is observed only once. Let C denote the observation time. Then current status data only provide $X \,=\, (C, \delta, Z)$, where $\delta \,=\, I(T \,\leq\, C) \,=\, 1$ if T is left-censored and 0 if T is right-censored. In the following, we will assume that T and C are independent given Z and that the joint distribution of (C, Z) does not involve θ_0 and H_0.

2.3 Inference Procedure

Consider a survival study giving n i.i.d. copies of X denoted by $\{ X_i = (c_i, \delta_i, z_i) \, ; \; i = 1, ..., n \}$. Under model (2.1), the log-likelihood function has the form

$$l_n(\theta, H) = \sum_{i=1}^{n} \{\delta_i \log[1 - \exp(-\Lambda(H(c_i) + z_i'\theta))] - (1 - \delta_i)\Lambda(H(c_i) + z_i'\theta)\} \, .$$

Without loss of generality, we will assume that $c_1, \cdots, c_n$ are the order statistics, that is, $c_1 \le c_2 \le \cdots \le c_n$. It is easy to see that only the values of H at the $c_i's$ can be estimated. In the following, for estimation of H_0, we will only consider these H that are right continuous increasing step functions with jumps at the c_i's and $H(t) = -\infty$ for all $t < c_1$ and $H(t) = H(c_n)$ for $t > c_n$. The maximum likelihood estimators of θ_0 and H_0 are defined as $\hat{\theta}_n$ and $\hat{H}_n$ with $\hat{H}_n(c_i) = h_i$ that maximize

$$\phi(\theta, \mathbf{h}) = \sum_{i=1}^{n} \{ \delta_i \log[1 - \exp(-\Lambda(h_i + z_i'\theta))] - (1 - \delta_i)\Lambda(h_i + z_i'\theta) \} \qquad (2.2)$$

subject to $h_1 \le h_2 \le \cdots \le h_n$, where $\mathbf{h} = (h_1, ..., h_n)$. Following Huang (1996), we assume that $\delta_1 = 1$ and $\delta_n = 0$. Otherwise, if $\delta_1 = 0$ or $\delta_n = 1$, maximizing $l_n(\theta, H)$ would give $\hat{H}_n(c_1) = -\infty$ or $\hat{H}_n(c_n) = \infty$.

To maximize $\phi(\theta, \mathbf{h})$, following Huang (1996), we propose to apply a two-stage procedure that maximizes $\phi(\theta, \mathbf{h})$ iteratively with respect to θ and $\mathbf{h}$. First we fix H. To obtain $\hat{\theta}_n$, note that if θ_0 is an interior point of the parameter space Θ and $\hat{\theta}_n$ is consistent (shown below), $\hat{\theta}_n$ will be an interior point of Θ for large n. Thus $(\hat{\theta}_n, \hat{H}_n)$

must be the solution to the score function $\partial l_n(\theta, H)/\partial\theta \,=\, 0$, which has the form

$$\sum_{i=1}^{n} z_i \lambda(H(c_i) + z_i'\theta) \left[\delta_i \frac{\exp(-\Lambda(H(c_i) + z_i'\theta)}{1 - \exp(-\Lambda(H(c_i) + z_i'\theta)} - (1 - \delta_i)\right] \,=\, 0\,. \tag{2.3}$$

That is, one can solve this equation for $\hat{\theta}_n$ for given H using, for example, the Newton-Raphson algorithm.

For the fixed θ, the determination of $\hat{H}_n$ or $\hat{\mathbf{h}} \,=\, (\hat{h}_1, ..., \hat{h}_n)$ is not as straightforward as that of $\hat{\theta}_n$ due to the constraint. To this end, we develop an ICM algorithm, which was first introduced for interval-censored data by Groeneboom and Wellner (1992). The others that discussed this algorithm for interval-censored data include Huang (1996) and Sun (2006). To describe the algorithm, define $G(t) \,=\, \exp(H(t))$. Then $G(0) \,=\, 0$ and the maximization of $\phi(\theta, \mathbf{h})$ with respect to $\mathbf{h}$ is equivalent to maximizing it with respect to $\mathbf{g} \,=\, \{\, g_i \,=\, \exp(h_i)\,\}_i^n$ over

$$C_g \,=\, \{\mathbf{g} = (g_1, \cdots, g_n)' \in \mathcal{R}^n : 0 \le g_1 \le \cdots \le g_n\}\,.$$

In the terms of $\mathbf{g}$, $\phi(\theta, \mathbf{h})$ in (2.2) can be rewritten as

$$\phi(\mathbf{g}) \,=\, \sum_{i=1}^{n} \{\, \delta_i \log[1 - \exp(-\Lambda(\log(g_i) + z_i'\theta))] - (1 - \delta_i)\Lambda(\log(g_i) + z_i'\theta)\,\}\;. \tag{2.4}$$

Let $W(\mathbf{g}) \,=\, -\mathrm{diag}(w_1, \cdots, w_n)$, a diagonal matrix, with $w_i \,=\, w_i(\mathbf{g}) \,=\, \partial^2\phi(\mathbf{g})/\partial^2 g_i$ and define

$$\phi^*(\mathbf{g}_1, \mathbf{g}_2) \,=\, -\frac{1}{2}\,[\mathbf{g}_1 - y(\mathbf{g}_2)]\,'W(g_2)\,[\mathbf{g}_1 - y(\mathbf{g}_2)]$$

for $\mathbf{g}_1, \mathbf{g}_2 \in C_g$, where $y(\mathbf{g}) = \mathbf{g} - W^{-1}(\mathbf{g})\,[\partial\phi/\partial\mathbf{g}]$. At the lth iteration, the ICM algorithm defines the updated estimate $\mathbf{g}^{(l)}$ as $\mathbf{g}$ maximizes $\phi^*(\mathbf{g}, \mathbf{g}^{(l-1)})$, which is given by the derivative of the convex minorant of the cumulative sum diagram $\{P_u; u = 0, 1, \cdots, n\}$, where $P_0 = (0, 0)$ and

$$P_u = \Big(\sum_{i=1}^{u} w_i^{(l-1)}, \sum_{i=1}^{u}(w_i^{(l-1)} g_i^{(l-1)} - \frac{\partial}{\partial g_i}\phi^*(\mathbf{g}^{(l-1)}))\Big)$$

for $u = 1, ..., n$. In above, $\mathbf{g}^{(l-1)}$ denotes the estimate of $\mathbf{g}$ obtained at the $(l-1)$th iteration and $w_i^{(l-1)} = w_i(\mathbf{g}^{(l-1)})$.

In summary, to determine $\hat{\theta}_n$ and $\hat{H}_n$, one can apply the following iterative algorithm.

Step 1. Select initial estimates $\theta_n^{(0)}$ and $H_n^{(0)}$.

Step 2. At the lth iteration, let $\theta = \theta_n^{(l-1)}$ and apply the ICM algorithm described above with $H_n^{(l-1)}$ being the initial estimate to obtain $H_n^{(l)}$.

Step 3. Fix $H = H_n^{(l)}$ and solve equation (2.3) using an iterative algorithm with $\theta_n^{(l-1)}$ being the initial estimate to obtain $\theta_n^{(l)}$.

Step 4. Check if $\|\theta_n^{(l)} - \theta_n^{(l-1)}\|^2 + \|H_n^{(l)} - H_n^{(l-1)}\|^2 \leq \epsilon$ for a given ϵ. If so, set $\hat{\theta}_n = \theta_n^{(l)}$ and $\hat{H}_n = H_n^{(l)}$ and stop. Otherwise, go back to step 2.

Now we begin to investigate the asymptotic properties of the maximum likelihood estimates $\hat{\theta}_n$ and $\hat{H}_n$ and determine the information bound for θ_0. First, we derive the efficient score function and information bound for θ_0. For this, define

$$Q(C, \delta, Z) \;=\; \delta \, \frac{\exp(-\Lambda(H(C) + Z'\theta))}{1 - \exp(-\Lambda(H(C) + Z'\theta))} - (1 - \delta)$$

and

$$O(C|Z) \;=\; E[Q^2(C, \delta, Z)|C, Z] \;=\; \frac{\exp(-\Lambda(H(C) + Z'\theta))}{1 - \exp(-\Lambda(H(C) + Z'\theta))} \,.$$

Also we need the following conditions.

(A1) The parameter space Θ is a bounded subset of R^p.

(A2) For covariate Z, there exists a positive constant z_0 such that $|Z| \;=\; Z'Z \;\leq\; z_0$ with probability 1. Also for any $\theta \neq \theta_0$, $P(Z'\theta \neq Z'\theta_0) \;>\; 0$.

(A3) The support of C is given by $S[C] = [l_c, u_c]$ with $0 < l_c < u_c < \infty$.

(A4) The function H_0 has strictly positive first derivative on $S[C]$ that is uniformly bounded and $H_0(0) = -\infty$.

(A5) The function Λ has strictly positive first derivative, that is, $\lambda(t) \;=\; d\Lambda(t)/dt > 0$. Furthermore, for any $0 < \tau < \infty$, there exists constants $0 < M_\tau^L < M_\tau^U < \infty$ such as $M_\tau^L < \lambda(t) < M_\tau^U$ for all $t \in [-\tau, \tau]$.

Suppose that assumptions (A2) and (A3) hold. Then the efficient score function for θ_0 has the form

$$\dot{l}_\theta^*(X) \;=\; \lambda(H(C) + Z'\theta)Q(C, \delta, Z)\Big(Z - \frac{E[Z\lambda^2(H(C) + Z'\theta)O(C|Z)|C]}{E[\lambda^2(H(C) + Z'\theta)O(C|Z)|C]}\Big) \qquad (2.5)$$

and the information bound $I(\theta)$ for θ_0 is given by

$$I(\theta) \;=\; E\Big\{R(C,Z)\Big[Z - \frac{E(ZR(C,Z)|C)}{E(R(C,Z)|C)}\Big]^{\otimes 2}\Big]\Big\}\,, \tag{2.6}$$

where $R(C,Z) \;=\; \lambda^2(H(C) + Z'\theta)O(C|Z)$ and $a^{\otimes 2} \;=\; aa'$ for a vector a..

For the consistency of $\hat{\theta}_n$ and $\hat{H}_n$, we proved under conditions (A1), (A2), (A3) and (A5), as $n \;\rightarrow\; \infty$, we have

$$d\Big[(\widehat{\theta}_n, \widehat{H}_n), (\theta_0, H_0)\Big] \;= |\widehat{\theta}_n - \theta_0| + ||\widehat{H}_n - H_0||_2 = \,O_p(n^{-1/3})\,,$$

This implies that

$$\hat{\theta}_n \;\rightarrow\; \theta_0 \;\; \text{a.s.} \tag{2.7}$$

and

$$\hat{H}_n(c) \;\rightarrow\; H_0(c) \;\; \text{a.s.} \tag{2.8}$$

for any $c \in S[C]$ at which the distribution of T is continuous. That is, $\hat{\theta}_n$ and $\hat{H}_n$ are consistent with the $n^{1/3}$ convergence rate rather than the typical $n^{1/2}$ convergence rate.

Moreover, suppose that θ_0 is an interior point of Θ and all of the conditions (A1) -

(A5) hold. Then as $n \rightarrow \infty$, we have

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \ \rightarrow \ N(0, I(\theta_0)^{-1})$$

in distribution.

For inference about θ_0, one needs to estimate the information bound $I(\theta_0)$, which involves three terms, $R(C, Z)$, $E[ZR(C, Z)|C]$ and $E[R(C, Z)|C]$. Given the observed data and the estimates $\hat{\theta}_n$ and $\hat{H}_n$, it is easy to see that $R(C, Z)$ can be consistently estimated by

$$\hat{R}(c, z) \ = \ \lambda^2(\hat{H}_n(c) + z'\hat{\theta}_n) \, \frac{\exp(-\Lambda(\hat{H}_n(c) + z'\hat{\theta}_n))}{1 - \exp(-\Lambda(\hat{H}_n(c) + z'\hat{\theta}_n))} \, .$$

Let $\mu_1(c) \ = \ E[ZR(C, Z)|C \ = \ c]$ and $\mu_2(c) \ = \ E[R(C, Z)|C \ = \ c]$. In general, it is hard to estimate $\mu_1(c)$ and $\mu_2(c)$ and the discussion given in Huang (1996) for the proportional hazards model applies here.

Suppose that Z is a binary variable taking value 0 or 1 with $P(Z \ = \ 1) \ = \ \gamma$ and let $f_0(c)$ and $f_1(c)$ denote the conditional density functions of C given $Z \ = \ 0$ and $Z \ = \ 1$, respectively. Then by following Huang (1996), one can estimate $\mu_1(c)/\mu_2(c)$ by

$$\hat{\mu}(c) \ = \ \frac{\hat{R}(c, 1) \, \hat{f}_1(c) \, \hat{\gamma}}{\hat{R}(c, 1) \, \hat{f}_1(c) \, \hat{\gamma} + \hat{R}(c, 0) \, \hat{f}_0(c) \, (1 - \hat{\gamma})} \, , \tag{2.9}$$

where $\hat{\gamma} \ = \ \sum_{i=1} z_i / \sum_{i=1} (1 - z_i)$, and $\hat{f}_0(c)$ and $\hat{f}_1(c)$ are kernel estimates of $f_0(c)$ and $f_1(c)$, respectively. For the special case where C and Z are independent, instead of using the estimate (2.9), one can estimate $I(\theta_0)$ by

$$\hat{I}(\hat{\theta}_n) = \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{R}(c_i, z_i)\left[z_i - \frac{\sum_{j=1}^{n} z_j \hat{R}(c_i, z_j)}{\sum_{j=1}^{n} \hat{R}(c_i, z_j)}\right]^{\otimes 2}\right\}, \qquad (2.10)$$

which is much simpler than the estimate based on (2.9). All the proof about the large sample properties can be found in Appendix A.

2.4 Simulation Studies

Simulation studies were conducted to assess the performance of the proposed maximum estimators and their asymptotic properties for finite sample situations. In the study, the covariate Z was assumed to be a binary variable generated from the Bernoulli distribution with $P(Z = 1) = P(Z = 0) = 0.5$. For the survival time T, it was supposed that the baseline hazard function of the error term ϵ takes the form

$$\lambda(t) = \frac{\exp(t)}{1 + r\exp(t)}$$

(Chen et al., 2002), where r is a constant. This model corresponds to the proportional hazards model if $r = 0$ and gives the proportional odds model if $r = 1$. Furthermore, we took $H(t) = \log(t)$ and generated the censoring time C from the uniform distribution $U(0, 2)$. That is, we assumed that C and Z are independent. The results given below are for $n = 100$ and 200 and based on 1000 replications.

Table 2.1 presents the results obtained based simulated data for $\hat{\theta}_n$, the maximum likelihood estimate of θ_0, with $\theta_0 = -1$, 0, or 1 and $r = 0$, 0.5 or 1. The table gives the estimated bias (BIAS) defined as the averages of the estimates $\hat{\theta}_n$ minus the true value, the sample standard deviations (SSD) of $\hat{\theta}_n$, and the averages of the estimated standard errors (ESE) obtained based on (2.10). The results indicate that the bias seems reasonably small for all situations except the case of $n = 100$ and $r = 1$ and the estimated standard errors are close to the sample standard deviations. It is clear that as expected, both the bias and the standard error become smaller when the sample size increases. Also it seems that the bias and standard error go up when r is moving away from 0 or the true model moves away from the proportional hazards model. One possible reason for this is the increasing of the percentage of right-censored observations when r is moving from 0. For example, with $\beta_0 = -1$, the percentages of right-censored failure times are approximately 57%, 61% and 64% for $r = 0, 0.5$ and 1, respectively. In other words, relatively large sample sizes are needed for r away from 0.

2.5 An Application of Lung Tumor Data

In this section, we apply the methodology proposed in the previous sections to the tumorigenicity data given in Table 1.3 of Sun (2006). Tumorigenicity experiments are usually designed to determine whether a suspected agent or environment accelerates the time until tumor onset in experimental animals. In these situations, the time to tumor onset is usually of interest but not directly observable. Instead, only the death or sacrifice time of an animal is observed along with the presence or absence of a tumor at the time. If the tumor can be considered to be rapidly lethal, meaning that its occurrence kills the animal right away, it is reasonable to treat the time as an exact or right-censored observation of the tumor onset time. On the other hand, if the tumor is nonlethal, then it is reasonable to assume that the tumor onset time and death time are independent given treatments.

The data considered here arose from 144 male RFM mice and consist of the death time of each animal measured in days and an indicator of lung tumor presence ($z_i = 1$) or absence ($z_i = 0$) at the time of death. The experiment involves two treatments, conventional environment (96 mice) and germ-free environment (48 mice). One of the objectives of the study was to compare the lung tumor incidence rates of the two treatment groups and lung tumors in RFM mice are predominantly nonlethal. The authors who discussed this set of data include Hoel and Walburg (1972) and Huang (1996).

To analyze the data, define $z_i = 1$ if the ith animal was in the conventional environment and 0 otherwise. Suppose that the error term in model (2.1) has the same baseline hazard function used in the simulation study. Using the triangular kernel function for estimates $\hat{f}_0(c)$ and $\hat{f}_1(c)$ with the bandwidth 200, we obtained $\hat{\theta}_n = -0.5576$ with the estimated standard error of 0.3201 for $r = 0$. By taking $r = 0.5$ and 1, we had $\hat{\theta}_n = -1.0679$ and -1.2226 with the corresponding standard errors being 0.5533 and 0.6315, respectively. These yielded p-values around 0.05 for testing $\theta_0 = 0$ or no treatment difference, which indicates that the animals in the conventional environment had higher tumor incidence rate than those in the germ-free environment.

We used different kernel functions and bandwidths and got similar results. The result is also similar to that given in Huang (1996) based on the proportional hazards model.

2.6 Concluding Remarks

As discussed before, current status data occur in many fields and for their regression analysis, several methods have been proposed under some specific regression models including the proportional hazards and the additive hazards model. It is well-known, however, that these specific models may fail in some situations and also model selection or diagnosis is quite difficult in general. Thus in practice one may prefer to apply a more general model like the linear transformation model considered in this chapter. One major advantage of the linear transformation model is that it is flexible and includes many models as special cases. For estimation of regression parameters in the linear transformation model, we derived their efficient estimates and the simulation study suggested that the estimates perform well in practice. As pointed in Section 2.3, the results include those given by Huang (1996) as a special case.

Chapter 3

Efficient Estimation for the Proportional Odds Model with Bivariate Current Status Data

3.1 Introduction

In Chapter 2, we discussed regression analysis of the linear transformation model for univariate current status data. While in practice bivariate current status data arise when two different failure times are of interest. This chapter applies the efficient estimation approach to regression analysis of bivariate current status data with the proportional odds model.

Efficient estimation approach provides a useful tool for semiparametric analysis of failure time data if the main interest is estimation of regression parameters (Bickel et al., 1993; Huang, 1996; Martinussen and Scheike, 2002). For example, Huang (1996) and Martinussen and Scheike (2002) considered the use of the approach for regression analysis of univariate current status data arising from the proportional hazards model and the additive hazards model, respectively. The main advantage of the approach is that it yields efficient estimates of regression parameters. For current status data, a number of authors have considered regression analysis of univariate case as listed in Section 1.4. Also for bivariate current status data, Jewell et al. (2005) investigated estimation of univariate cumulative distribution functions.

In the following, we will discuss regression analysis of bivariate current status data. We will begin in Section 3.2 with introducing the notation and describing the models used throughout this chapter. For modeling the joint survival function, we assume it follows a copula model with the association parameter that may depend on covariates. The marginal survival functions are supposed to follow the proportional odds model. In Section 3.3, we derive the efficient score and information bound for all regression parameters and the estimation procedure is presented in Section 3.4. Section 3.5 gives some results from the simulation studies conducted for assessing the performance of the procedure and an illustration is provided in Section 3.6. Section 3.7 concludes with some discussion.

3.2 Notation and Models

Let T_1 and T_2 be the two related failure times of interests and suppose that both variables are only observed at a monitoring time C. That is, the only information available for them are C, $\delta_1 = I(T_1 \geq C)$ and $\delta_2 = I(T_2 \geq C)$, indicating whether the survival events represented by T_1 and T_2 have occurred before C. Note that here for simplicity, we assume that T_1 and T_2 have the same observation time and the methodology given below can be easily generalized to situations where they have different observation times. More comments on this will be given below. Also it will be assumed that all T_1, T_2 and C are continuous variables.

Let X be a vector of covariates and $S_k(t)$ denote the marginal survival function of T_k, $k = 1, 2$. To describe the covariate effects on T_k, it will be assumed that given X, $S_k(t)$ has the form

$$\frac{S_k(t)}{1 - S_k(t)} = \exp(X'\beta) \, \frac{S_{0k}(t)}{1 - S_{0k}(t)} \tag{3.1}$$

where S_{0k} is an unknown baseline survival function and β denotes the vector of regression parameters. That is, T_k follows the proportional odds model (Yang and Prentice, 1999). Note that in model (3.1), without loss of generality, it is supposed that the covariate effects are the same for T_1 and T_2. If they are different, one can easily define a common β through the introduction of extra type-specific covariates. Define $O_k(t) = S_k(t)/\{1 - S_k(t)\}$ and $O_{0k}(t) = S_{0k}(t)/\{1 - S_{0k}(t)\}$. Then we have

$$S_k(t) = \frac{\exp(x'\beta) O_{0k}(t)}{1 + \exp(x'\beta) O_{0k}(t)} \, .$$

It will be assumed that T_1 and T_2 are independent of C given covariates.

To model the joint survival function of T_1 and T_2, several approaches can be applied (Hougaard, 2000). A common one, which will be used here, is the copula model approach that assumes

$$S(s,t) = P(T_1 > s, T_2 > t) = C_\alpha(S_1(s), S_2(t)) \, , \tag{3.2}$$

where C_α: $[0,1]^2 \to [0,1]$ is a genuine survival function on the unit square and α is a global association parameter. The copula model has attracted considerable attention in failure time data analysis (Genest and MacKay, 1986; Oakes, 1989; Wang, 2003) and includes many useful bivariate failure time models as special cases. For example, one special case is the Clayton model given by

$$C_\alpha(u,v) \;=\; (u^{1-\alpha} + v^{1-\alpha} - 1)^{1/(1-\alpha)}$$

(Clayton, 1978) and a more general example is the Archimedean copula family defined as

$$C_\alpha(u,v) = \phi_\alpha^{-1}(\phi_\alpha(u) + \phi_\alpha(v))\,,\;\; 0 \le u,\;\; v \le 1,$$

where $\phi(\cdot)$ is a decreasing convex function defined on $[0,1]$ with $\phi(1) \;=\; 0$. The global association parameter α is related to the Kendall's τ through $\tau \;=\; 4\int_0^1\int_0^1 C_\alpha(u,v)dudv-1$.

One special and desirable feature of the copula model is that one can model the association and the marginal survival functions separately. This is convenient as sometimes α may depend on the covariates. In the following, we will suppose that

$$\alpha \;=\; \exp(X'\gamma) \,+\, 1\,, \tag{3.3}$$

where γ is a vector of regression parameters representing the effects of covariates on the association between T_1 and T_2.

3.3 The Efficient Score and the Information Bound

In this section, we derive the efficient score function and the information bound for regression parameters $\theta' = (\beta', \gamma')$ under models (3.1)-(3.3) based on bivariate current status data $\{C, \delta_1, \delta_2, X\}$. For this, define the following counting processes

$$\begin{aligned} N_{00}(t) &= \delta_1\delta_2 I(C \le t) \ , & N_{10}(t) &= (1-\delta_1)\delta_2 I(C \le t)\,, \\ N_{01}(t) &= \delta_1(1-\delta_2) I(C \le t) \ , & N_{11}(t) &= (1-\delta_1)(1-\delta_2) I(C \le t)\,. \end{aligned}$$

Also define

$$S_{00}(\theta, t) = P(T_1 > t, T_2 > t) = C_{\alpha(\gamma)}(S_1(\beta, t), S_2(\beta, t))\,,$$

$$S_{10}(\theta, t) = P(T_1 < t, T_2 > t) = S_2(\beta, t) - C_{\alpha(\gamma)}(S_1(\beta, t), S_2(\beta, t))\,,$$

$$S_{01}(\theta, t) = P(T_1 \ge t, T_2 \le t) = S_1(\beta, t) - C_{\alpha(\gamma)}(S_1(\beta, t), S_2(\beta, t))\,,$$

and

$$S_{11}(\theta, t) = P(T_1 \le t, T_2 \le t) = 1 - S_1(\beta, t) - S_2(\beta, t) + C_{\alpha(\gamma)}(S_1(\beta, t), S_2(\beta, t))\,.$$

Then it can be easily shown that for $j = 0, 1$ and $m = 0, 1$, the intensity function for N_{jm} is given by $Y(t)\,\lambda_c(t)\,S_{jm}(\theta, t)$, where $Y(t) = I(C \ge t)$, $\lambda_c(t)$ denotes the hazard function for C. The log likelihood function based on $\{C, \delta_1, \delta_2, X\}$ is proportional to

$$l(\theta, O_{01}, O_{02}) = \sum_{j=0}^{1} \sum_{m=0}^{1} \int \log\{S_{jm}(\theta, t)\}\, dN_{jm}(t)$$

since $S_{jm}(t)$ depends on $O_{0k}(t)$.

Let

$$D_\alpha = \partial C_\alpha(S_1, S_2)/\partial\alpha, \ D_u = \partial C_\alpha(u,v)/\partial u|_{u=S_1,v=S_2},$$

and

$$D_v = \partial C_\alpha(u,v)/\partial v|_{u=S_1,v=S_2}.$$

Then we have $\partial S_k(\beta,t)/\partial\beta = Q_k(t)X$, $\partial C_{\alpha(\gamma)}(S_1,S_2)/\partial\beta = (D_uQ_1 + D_vQ_2)X$, and $\partial C_{\alpha(\gamma)}(S_1,S_2)/\partial\gamma = (\alpha-1)D_\alpha X$, where $Q_k(t) = S_k(t)(1-S_k(t))$. It follows from

$$S_{jm}(\theta,t) = jm + (-1)^j mS_1 + (-1)^m jS_2 + (-1)^{m+j}C_{\alpha(\gamma)}(S_1,S_2)$$

that

$$\begin{aligned} \frac{\partial S_{jm}(\theta,t)}{\partial\beta} &= \Big[\big((-1)^j m + (-1)^{m+j}D_u\big)\,Q_1 + \big((-1)^m j + (-1)^{m+j}D_v\big)\,Q_2\Big]X \\ \frac{\partial S_{jm}(\theta,t)}{\partial\gamma} &= (-1)^{j+m}\alpha_1 D_\alpha X\,, \end{aligned}$$

where $\alpha_1 = \alpha - 1$.

For $j = 0,1$ and $m = 0,1$, define $a_{jm1} = [(-1)^j m + (-1)^{m+j}D_u]Q_1$, $a_{jm2} = [(-1)^m j + (-1)^{m+j}D_v]Q_2$, and $a_{jm} = a_{jm1} + a_{jm2}$. Also define $A_{jm} = (a_{jm1}, a_{jm2})'$,

$$X_{jm1} = \begin{pmatrix} X \\ (-1)^{j+m}a_{jm}^{-1}\alpha_1 D_\alpha X \end{pmatrix}, \quad X_{jm2} = \begin{pmatrix} X \\ (-1)^{j+m}a_{jm}^{-1}\alpha_1 D_\alpha X \end{pmatrix},$$

$X_{jm} = (X_{jm1}, X_{jm2})$, $G = E\sum_{j,m} A_{jm}A'_{jm}S_{jm}^{-1}Y\lambda_c$ and $H = E\sum_{j,m} X_{jm}A_{jm}A'_{jm}S_{jm}^{-1}Y\lambda_c$.

We show in Appendix B that the efficient score function for θ is given by

$$\dot{l}_{\theta^*} = \sum_{j=0}^{1}\sum_{m=0}^{1}\int \frac{X_{jm}A_{jm} - HG^{-1}A_{jm}}{S_{jm}} dM_{jm} \,,$$

where

$$dM_{jm}(t) \; = \; dN_{jm}(t) - Y(t)\lambda_c(t)S_{jm}dt, \;\; j = 0, 1; \;\; m = 0, 1$$

are martingales. Furthermore, the information bound for θ has the form

$$I(\theta) \; = \; E\dot{l}_{\theta^*}^{\otimes 2} \; = \sum_{j=0}^{1}\sum_{m=0}^{1} E\int \Big((X_{jm} - HG^{-1})A_{jm}\Big)^{\otimes 2} S_{jm}^{-1} Y\lambda_c dt \,,$$

where $a^{\otimes 2} \; = \; aa'$ for a vector a. Note that $\sum_{j=0}^{1}\sum_{m=0}^{1} A_{jm} \; = \; 0$ and $\sum_{j=0}^{1}\sum_{m=0}^{1} a_{jm} = 0$. This suggests that the efficient score function can be rewritten as

$$\dot{l}_{\theta^*} \; = \sum_{j=0}^{1}\sum_{m=0}^{1}\int \frac{X_{jm}A_{jm} - HG^{-1}A_{jm}}{S_{jm}} dN_{jm} \,. \tag{3.4}$$

3.4 Estimation of Regression Parameters

Now we consider the estimation of regression parameters θ. Assume that n i.i.d. copies of $\{C, \delta_1, \delta_2, X\}$ are available and denoted by $\{\, (C_i, \delta_{1i}, \delta_{2i}, X_i)\, ; \, i = 1, ..., n\,\}$. In the following, we will discuss situations where the distribution of the C_i's is independent of covariates X_i's and then situations where the C_i's may depend on the X_i's.

3.4.1 Estimation with Covariate-Independent Monitoring Times

Suppose that the monitoring times C_i's follow the same distribution. Let $N_{jmi}(t)$, $S_{jmi}(\theta, t)$, A_{jmi} and X_{jmi} be defined as in the previous section but with respect to subject i, $i = 1, ..., n$. To estimate θ, it is natural to use the empirical version of the efficient score function given in (3.4), which has the form

$$U(\theta, O_{01}, O_{02}) \;=\; \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=0}^{1} \sum_{m=0}^{1} \int \frac{X_{jmi}A_{jmi} - H_n G_n^{-1} A_{jmi}}{S_{jmi}} \, dN_{jmi} \,,$$

where

$$G_n = \frac{1}{n} \sum_{i=1}^{n} \sum_{j,m} A_{jmi} A'_{jmi} S_{jmi}^{-1} Y_i \lambda_c$$

and

$$H_n = \frac{1}{n} \sum_{i=1}^{n} \sum_{j,m} X_{jmi} A_{jmi} A'_{jmi} S_{jmi}^{-1} Y_i \lambda_c$$

are the empirical estimators of H and G defined before.

It can be easily seen and as indicated above, $U(\theta, O_{01}, O_{02})$ depends on the unknown functions $O_{01}(t)$ and $O_{02}(t)$ (or $S_{01}(t)$ and $S_{02}(t)$) and thus are not available yet. For this, we propose to replace them with their consistent estimates. Let $\hat{O}_{01}(t)$ and $\hat{O}_{02}(t)$ denote some consistent estimates of $O_{01}(t)$ and $O_{02}(t)$, respectively, with the convergence rate $O(n^{1/3})$, which will be discussed below. Then we can define an estimator $\hat{\theta}$ as the solution to the following estimating function

$$U(\theta, \hat{O}_{01}, \hat{O}_{02}) \;=\; \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=0}^{1} \sum_{m=0}^{1} \int \frac{\hat{X}_{jmi}\hat{A}_{jmi} - \hat{H}_n \hat{G}_n^{-1} \hat{A}_{jmi}}{\hat{S}_{jmi}} dN_{jmi} \;=\; 0 \,,$$

where $\hat{X}_{jmi}$, $\hat{A}_{jmi}$, $\hat{S}_{jmi}$, $\hat{H}_n$ and $\hat{G}_n$ denote X_{jmi}, A_{jmi}, S_{jmi}, H_n and G_n with $O_{01}(t)$ and $O_{02}(t)$ replaced by $\hat{O}_{01}(t)$ and $\hat{O}_{02}(t)$.

Let θ_0 denote the true value of θ. To derive the asymptotic distribution of $\hat{\theta}$, we show in Appendix B that

$$U(\theta_0, \hat{O}_{01}, \hat{O}_{02}) \;=\; U(\theta_0, O_{01}, O_{02}) \;+\; o_p(1)\,. \tag{3.5}$$

It then follows from the central limit theorem for martingales that $U(\theta_0, \hat{O}_{01}, \hat{O}_{02})$ converges in distribution to a normal random vector with zero mean and covariance matrix $I(\theta_0)$. Thus it can be shown that under some regularity conditions and as $n \;\to\; \infty$,

$$\sqrt{n}\,(\hat{\theta} \;-\; \theta_0) \to N(0, I^{-1}(\theta_0))$$

in distribution with $I(\theta_0)$ consistently estimated by

$$I(\hat{\theta}) \;=\; \frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{1}\sum_{m=0}^{1}\int \left(\frac{\hat{X}_{jmi}\hat{A}_{jmi} - \hat{H}_n\hat{G}_n^{-1}\hat{A}_{jmi}}{\hat{S}_{jmi}}\right)^{\otimes 2} dN_{jmi}\,.$$

3.4.2 Estimation with Covariate-Dependent Monitoring Times

In reality, the distribution of C_i may depend on covariates. In this case, the hazard function of C for subject i, $\lambda_c(t) \;=\; \lambda_{ci}(t)$ can depend on X_i and the empirical efficient score function is given by

$$U_D(\theta, O_{01}, O_{02}) \;=\; \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{j=0}^{1}\sum_{m=0}^{1}\int \frac{(X_{jmi} - H_n^D(G_n^D)^{-1})A_{jmi}}{S_{jmi}}\,dN_{jmi}\,,$$

where

$$H_n^D = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{1}\sum_{m=0}^{1} X_{jmi}A_{jmi}A'_{jmi}S_{jmi}^{-1}Y_i\lambda_{ci}$$

and

$$G_n^D = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{1}\sum_{m=0}^{1} A_{jmi}A'_{jmi}S_{jmi}^{-1}Y_i\lambda_{ci}\,.$$

It can be seen that in addition to O_{01} and O_{02}, $U_D(\theta, O_{01}, O_{02})$ also involves the unknown function λ_{ci} and thus unknown H_n^D and G_n^D. To estimate H_n^D and G_n^D, following Martinussen and Scheike (2002), we propose to use the following kernel estimates

$$\hat{H}_n^D(t) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{1}\sum_{m=0}^{1}\int \hat{X}_{jmi}(s)\hat{A}_{jmi}(s)\hat{A}'_{jmi}(s)\hat{S}_{jmi}^{-1}(s)Y_i(s)K_b(s-t)dN_{jmi}(s)$$

and

$$\hat{G}_n^D(t) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{1}\sum_{m=0}^{1}\int \hat{A}_{jmi}(s)\hat{A}'_{jmi}(s)\hat{S}_{jmi}^{-1}(s)Y_i(s)K_b(s-t)dN_{jmi}(s)$$

with O_{01} and O_{02} replaced by $\hat{O}_{01}(t)$ and $\hat{O}_{02}(t)$ defined above. In these estimates, K_b is a kernel function with bandwidth $b > 0$, $\int uK_b(u)du = 0$ and $\int K_b(u)du = 1$, which will be discussed below.

As before, one could naturally estimate θ by the solution to

$$U_D(\theta, \hat{O}_{01}, \hat{O}_{02}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{j=0}^{1}\sum_{m=0}^{1}\int \frac{(\hat{X}_{jmi} - \hat{H}_n^D(\hat{G}_n^D)^{-1})\hat{A}_{jmi}}{\hat{S}_{jmi}}dN_{jmi} = 0\,.$$

Let $\hat{\theta}^D$ denote the estimate defined above. Then similarly as $\hat{\theta}$, one can show that under some regularity conditions, $n^{1/2}(\hat{\theta}^D - \theta_0)$ follows an asymptotically normal distribution with mean zero and covariance matrix $I(\theta_0)$ that can be consistently estimated by

$$I(\hat{\theta}^D) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{1}\sum_{m=0}^{1}\int\left\{\frac{\hat{X}_{jmi}\hat{A}_{jmi} - \hat{H}_n^D(\hat{G}_n^D)^{-1}\hat{A}_{jmi}}{\hat{S}_{jmi}}\right\}^{\otimes 2} dN_{jmi}\,.$$

To implement the estimation procedures proposed above, one needs to estimate $O_{01}(t)$ and $O_{02}(t)$ (or $S_{01}(t)$ and $S_{02}(t)$). Since the main interest here is the estimation of regression parameters, one simple approach for estimating $O_{0k}(t)$ is to base it on the univariate current status data $\{\,(C_i, \delta_{ki}, X_i)\,;\, i = 1, ..., n\,\}$ on T_k, $k = 1, 2$. For this, one can employ the sieve estimation method given in Huang and Rossini (1997) and they showed that the sieve estimator of $\log(O_{0k}(t))$ given there is consistent and can achieve the convergent rate $O(n^{1/3})$.

In the case of covariate-dependent monitoring times, we also need to choose a kernel function $K_b(t) = b^{-1}K(t/b)$ along with a bandwidth b to determine $\hat{\theta}^D$. For this, many choices for $K(t)$ are available. For example, two popular choices are the normal and triangular kernel functions given by

$$K(t) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}t^2)\,,\; K(t) = (1 - |t|)I(|t| \leq 1)\,,$$

respectively. For the selection of the bandwidth b, one can employ either a manual or data-driven approach. For simplicity, we will use the manual approach in the following numerical studies and more comments on this will be given below.

3.5 Numerical Results

Several simulation studies were conducted for the assessment of the estimation procedures proposed in the previous sections. In the studies, we focused on situations where the covariate X is a binary variable taking value 0 or 1 with probability 0.5. For the survival variables T_1 and T_2, it was assumed that T_k follows the proportional odds model (3.1) with $S_{0k}(t)$ being the exponential survival function with the hazard function $\lambda_{0k}(t) = 1$. Furthermore, we assumed that their joint survival function is specified by the Clayton model as

$$S(t_1, t_2) = \{S_1^{1-\alpha}(t_1) + S_2^{1-\alpha}(t_2) - 1\}^{1/(1-\alpha)} .$$

For the monitoring time C, it was supposed that it follows either the uniform distribution over $(0, 1)$ or the exponential distribution with the hazard function $\lambda_c(t) = \exp(X)$. Note that in the latter case, C is dependent of X. We also considered two cases for the association parameter α. One is that α is independent of X and the other is that α is dependent of X and taken to be $\alpha = \exp(X'\gamma) + 1$. In all situations, the sieve method discussed in the previous section was applied for estimation of $O_{01}(t)$ and $O_{02}(t)$. The results given below are based on 1000 replications with $n = 200$.

Table 3.1 presents the results obtained for estimation of β and α for the situations where both C and α are independent of the covariate X with $\beta_0 = 0.5$, 0, or -0.5 and $\alpha_0 = 2$. The table includes the estimated bias (Bias) given by the average of the estimates minus the true value, the average of the estimated standard errors (ESR), the sample standard deviation of the estimates (SSD), and the 95% empirical coverage probability (CP). The results obtained for situations where both C and α depend on X are given in Table 3.2 with $\beta_0 = 0.5$, 0, or -0.5 and $\gamma_0 = 1$. The results indicate that the estimation procedure seems to perform well for the situations considered. In particular, the estimates seem to be unbiased and the variance estimates are close to the sample variances. It can also be seen from the tables that for the given sample size, it seems that the parameters β can be estimated more accurately than the parameter α or γ.

To investigate the normal approximation to the distributions of the regression parameters, we obtained the quantile plots of the standardized estimates against the standard normal variable. These plots, which are not shown here, indicate that the approximation seems reasonable. In the simulation study, we also considered other sample sizes and kernel functions and obtained similar results. For the results given here for the case where C depends on X, the normal kernel function was used for estimation of H_n^D and G_n^D with the bandwidth b set to be equal to the range of C divided by 10. We also tried the triangle kernel function and a few other values for the bandwidth and obtained similar results.

3.6 An Illustrative Example

To illustrate the proposed estimation procedures, we consider a two-year tumorigenicity study conducted by the National Toxicology Program on groups of 50 male and female F344/N rats and B6C3F$_1$ mice (Dunson and Dinse, 2002). In the study, the animals were exposed to chloroprene at different concentrations by inhalation, 6 hours per day, 5 days per week, for two years with the goal of examining the effect of chloroprene on tumor growth. At their death or sacrifice, the animals were examined for tumor presence or absence. For the illustration here, we will focus on two lung tumors, Alveolar/Bronchiolar Carcinoma (A/B C) and Alveolar/Bronchiolar Adenoma (A/B A), and consider the 100 male B6C3F$_1$ mice in the control and high dose groups.

For the analysis, define T_{1i} and T_{2i} to be the occurrence times of A/B C and A/B A, respectively, for the ith animal and $X_i = 1$ if the ith animal was in the high dose group and $X_i = 0$ otherwise, $i = 1, ..., 100$. Here C_i is the death or sacrifice time for the ith animal. Assume that T_{1i} and T_{2i} follow the models (3.1)-(3.3). Then the application of the estimation procedure proposed in Section 3.2 yielded $\hat{\beta}^D = -2.3259$ and $\hat{\gamma}^D = -13.3645$ with the estimated standard errors being 0.2747 and 19.1911, respectively. These results indicate that the association parameter does not seem to depend on the treatment. Next we applied the estimation procedure given in Section 3.1 and obtained $\hat{\beta} = -2.3188$ and $\hat{\alpha} = 0.6085$ with the estimated standard errors of 0.1990 and 0.1666. Here for both cases, we assumed the Clayton model for the joint distribution of T_{1i} and T_{2i} and used the same kernel function and bandwidth as those used in the simulation studies. The results suggest that the high dose chloroprene had a significant effect of increasing the tumor occurrence rates and the occurrence rates of the two lung tumors, A/B C and A/B A, were negatively significantly correlated.

To investigate the possible dependence of the analysis results given above on the kernel function and the copula model, we tried different kernel functions and copula models. For example, with the use of the triangular kernel function and the Clayton model, we obtained $\hat{\beta} = -2.2916$ and $\hat{\alpha} = 0.6023$ and the the estimated standard errors are 0.2164 and 0.1785, respectively. In the case of the FGM model defined as $C_\alpha(u,v) = uv + \alpha uv(1-u)(1-v)$ and the normal kernel function, the results become $\hat{\beta} = -2.3314$ and $\hat{\alpha} = -0.7921$ with the estimated standard errors of 0.4393 and 0.3409. These results gave similar conclusions as before.

3.7 Discussion

The proportional odds model is an important and commonly used model in failure time data analysis (Yang and Prentice, 1999; Rabinowitz et al., 2000). This chapter discussed the application of the model to the analysis of bivariate current status data and for inference, an efficient estimation approach was developed. We showed that the resulting estimates of regression parameters have asymptotic normal distributions and can achieve the information bound. The approach makes use of the copula model and allows the association parameter to depend on covariates. The simulation study indicates that the estimates perform well for practical situations.

In the estimation procedures, for simplicity, we assumed that the two related failure variables of interest have the same monitoring time. This is true for many situations such that the two variables represent two different events on the same subject as in the example discussed in Section 3.6. On the other hand, it is straightforward to generalize the methodology to situations where the two variables may have different monitoring times. Also for simplicity, we only considered the manual selection approach for the bandwidth in estimation of H_n^D and G_n^D. In practice, one may want to apply some data-driven approaches, whose investigation is beyond the scope of this chapter, such as the cross validation approach.

Chapter 4

General Estimating Equations in Biased Sample Problems

4.1 Introduction

In the previous chapters, all the samples used are regarded as independent and identical distributed. As we all know, when an investigator records an observation, the recorded observations will not have the same original distribution unless every observation is given an equal chance of being recorded. But in practical problems, this assumption is often violated because of the different selection probabilities. Biased sampling scheme arise when the items are observed with a probability that depends on the outcome. It is frequently used in biostatistics, for example, the case-control studies (Prentice and Pyke, 1979). More examples can be found in Section 1.5. Cox (1969) first discussed the nonparametric inference about the cumulative distribution function (cdf) F for length biased sampling case. Vardi (1982, 1985) and Gill, Vardi and Wellner (1988) provided a nonparametric maximum likelihood estimate (NPMLE) of length bias sampling problem and proved the the estimate $\hat{F}$ converges weekly to a Gaussian process. In a more recent paper, Zhou et al. (2002) discussed a semiparametric method for a biased sampling problem with continuous outcome. More applications about the biased sampling models can be found in Patil and Rao (1978), Gilbert (1996) and Gilbert, Lele and Vardi (1999).

The empirical likelihood method was presented by Owen (1988, 1990, 1991) based on some previous work by Thomas and Grunkemeier (1975). Qin (1993) applied the empirical likelihood method in biased sampling problems. But the paper only considered the case when there is just one random variable of interest and only two biased samples were selected. In survey sampling or economic problems, it is very common to estimate the properties of some variable Y with some known information about the auxiliary variable X. For example, the mean of X, $E(x) = c$ is known and the mean of Y, $E(y) = \beta$ is to be estimated. In this chapter, we will consider the following type of data. Let Y and X be the random variables of interest, $(x_{ij}, y_{ij}),\ i = 1, 2, \ldots, I;\ j = 1, 2, \ldots, n_i$, are random samples from the probability density function (pdf)

$$f_i(y,x) = \frac{w_i(y)f(y,x)}{\int\int w_i(y)f(y,x)dydx}, i = 1, \ldots, I,$$

where $w_i(y)$'s are given functions. Suppose we are interested in the parameter β which is associated with $F(y,x)$. The information about β is given through independent unbiased estimating functions $g(y,x,\beta)$ defined by $E[g(y,x,\beta)] = 0$, where $g(y,x,\beta) = (g_1(y,x,\beta), \ldots, g_r(y,x,\beta))^T$ is an $r \times 1$ vector function and β is a $p \times 1$ parameter with

$r \geq p$. For the example described above, if $E(x) = c$ is known and $E(y) = \beta$ is to be estimated, then we have $g(y, x, \beta) = (x - c, y - \beta)^T$. Because of the different probability of selection in this problem, the standard statistical analysis can be misleading. Qin and Lawless (1994) linked the general estimating equations and empirical likelihood and developed new methods for the analysis. The authors showed the likelihood ratio statistic for testing the parameter is asymptotically chi-square distributed in certain cases. But they only considered the simple random sampling problem in that paper. Here we will discuss the biased sampling problem by using the empirical likelihood method with estimating equations.

The rest of this chapter is organized as follows: Section 4.2 gives the main results for the biased sampling problems. A likelihood-ratio based test for the parameter will be proposed as well as the large sample properties for the test statistic. The confidence interval of the parameter can be computed based on the asymptotic properties of the test statistic. Some simulation results are shown in Section 4.3. Section 4.4 provides some discussion.

4.2 Likelihood Ratio Test for Biased Sample Problems

Consider the biased sampling data

$$(x_{ij}, y_{ij}) \sim f_i = \frac{w_i(y)dF(y,x)}{\int\int w_i(y)dF(y,x)}, \; i = 1, 2, \ldots, I; \; j = 1, 2, \ldots, n_i,$$

where $w_i(y)$'s are given functions. Denote $N = n_1 + \cdots + n_I$, $k_i = \frac{n_i}{N}, i = 1, \ldots, I$, and $(y_1, x_1), \ldots, (y_N, x_N)$ be the combined observations. Then the probability of the data is

$$P(data) = \prod_{i=1}^{I} \prod_{j=1}^{n_i} \frac{w_i(y_{ij}) dF(y_{ij}, x_{ij})}{\theta_i},$$

where $\theta_i = \int \int w_i(y) dF(y, x)$. Now we want to maximize $P(data)$ with respect to the cdf F. Let $p_j = dF(y_j, x_j),\ j = 1, 2, \ldots, N$ and

$$L = \left\{ \prod_{j=1}^{N} p_j \right\} \left\{ \prod_{i=1}^{I} \theta_i^{-n_i} \right\}.$$

Then

$$l = \log L = \sum_{j=1}^{N} \log p_j - \sum_{i=1}^{I} n_i \log \theta_i.$$

The NPMLE problem is to maximize l with respect to $p_j,\ j = 1, 2, \ldots, N$ with the constraints

$$\sum_{j=1}^{N} p_j = 1,\ p_j \geq 0, j = 1, \ldots, N \text{ and } \theta_i = \sum_{j=1}^{N} p_j w_i(y_j),\ i = 1, \ldots, I.$$

As introduced in Section 4.1, suppose the information about the parameter of interest $\beta(F)$ is given by $E[g(y, x, \beta)] = 0$, where $g(y, x, \beta) = (g_1(y, x, \beta), \ldots, g_r(y, x, \beta))^T$ is an $r \times 1$ vector function. Thus an additional constraint will be given by $\sum_{i=1}^{N} p_i g(y_i, x_i, \beta) = 0$. To maximize l with respect to p_j, Lagrange multiplier argument can be used by defining

$$\begin{aligned} H &= \sum_{j=1}^{N} \log p_j - \sum_{i=1}^{I} n_i \log \theta_i - N\lambda_1^T \sum_{j=1}^{N} p_j g(y_j, x_j, \beta) \\ &\quad -N\lambda_2^T [\sum_{j=1}^{N} p_j w(y_j) - \theta] + \rho(1 - \sum_{j=1}^{N} p_j), \end{aligned}$$

where $\theta = (\theta_1, \dots, \theta_I)^T$, $w = (w_1, \dots, w_I)^T$ and λ_1, λ_2, ρ are Lagrange multipliers. Since $\sum_{j=1}^{N} p_j w_i(y_j) - \theta_i = \sum_{j=1}^{N} p_j(w_i(y_j) - \theta_i)$, taking derivatives with respect to p_j, we have

$$\frac{\partial H}{\partial p_j} = \frac{1}{p_j} - N\lambda_1^T g(y_j, x_j, \beta) - N\lambda_2^T [w(y_j) - \theta] - \rho = 0,\ j = 1, \dots, N.$$

Then $\rho = N$ by noticing $\sum_{j=1}^{N} p_j \frac{\partial H}{\partial p_j} = 0$. So

$$p_j = \frac{1}{N} \frac{1}{1 + \lambda_1^T g(y_j, x_j, \beta) + \lambda_2^T [w(y_j) - \theta]},$$

where

$$\begin{aligned} \sum_{j=1}^{N} \frac{g(y_j, x_j, \beta)}{1 + \lambda_1^T g(y_j, x_j, \beta) + \lambda_2^T [w(y_j) - \theta]} &= 0 \qquad (4.1) \\ \sum_{j=1}^{N} \frac{w(y_j) - \theta}{1 + \lambda_1^T g(y_j, x_j, \beta) + \lambda_2^T [w(y_j) - \theta]} &= 0. \end{aligned}$$

Thus the empirical log likelihood function can be written as

$$l(\theta,\beta) = -\sum_{j=1}^{N} \log\left[1+\lambda_1^T g(y_j,x_j,\beta)+\lambda_2^T[w(y_j)-\theta]\right] - \sum_{i=1}^{I} n_i \log\theta_i - N\log N.$$

Let $l(\hat{\theta},\hat{\beta})$ be the maximum value of $l(\theta,\beta)$ and $l(\tilde{\theta},\beta)$ be the maximum value of $l(\theta,\beta)$ when β is fixed. Then we define

$$R(\beta) = 2[l(\hat{\theta},\hat{\beta}) - l(\tilde{\theta},\beta)].$$

Now we develop the limiting distribution of the empirical likelihood ratio statistic $R(\beta)$ and derive the confidence intervals for β.

To this end, we define

$$l_E = \sum_{j=1}^{N} \log\left[1+\lambda_1^T g(y_j,x_j,\beta)+\lambda_2^T[w(y_j)-\theta]\right] + \sum_{i=1}^{I} n_i \log\theta_i.$$

Since $l_E = -l - N\log N$, to maximize l, it is equivalent to minimize l_E. Now we reparameterize the parameter. Note that

$$\begin{aligned} & 1+\lambda_1^T g(y_j,x_j,\beta)+\lambda_2^T[w(y_j)-\theta] \\ = & 1+\lambda_1^T g(y_j,x_j,\beta)+(\lambda_2-\frac{k}{\theta})^T[w(y_j)-\theta]+\left(\frac{k}{\theta}\right)^T[w(y_j)-\theta] \\ = & \left(\frac{k}{\theta}\right)^T w(y_j)+v_1^T g(y_j,x_j,\beta)+v_2^T[w(y_j)-\theta]+(1-\sum_{i=1}^{I}k_i) \end{aligned}$$

$$= \quad \left(\frac{k}{\theta}\right)^T w(y_j) + v_1^T g(y_j, x_j, \beta) + v_2^T[w(y_j) - \theta]$$

where $v_1 = \lambda_1,\ v_2 = \lambda_2 - \frac{k}{\theta}$ and $\frac{k}{\theta} = \left(\frac{k_1}{\theta_1}, \ldots, \frac{k_I}{\theta_I}\right)^T$. Then (4.1) becomes

$$\begin{aligned}\sum_{j=1}^N \frac{h_1(y_j, x_j, \beta)}{1 + v_1^T h_1(y_j, x_j, \beta) + v_2^T h_2(y_j, \theta)} &= 0 \\ \sum_{j=1}^N \frac{h_2(y_j, \theta)}{1 + v_1^T h_1(y_j, x_j, \beta) + v_2^T h_2(y_j, \theta)} &= 0, \end{aligned} \tag{4.2}$$

and

$$p_j = \frac{1}{N} \frac{1}{\left(\frac{k}{\theta}\right)^T w(y_j)} \frac{1}{1 + v_1^T h_1(y_j, x_j, \beta) + v_2^T h_2(y_j, \theta)},$$

where

$$\begin{aligned} h_1 &= \frac{g(y_j, x_j, \beta)}{\left(\frac{k}{\theta}\right)^T w(y_j)} \\ h_2 &= \frac{w(y_j) - \theta}{\left(\frac{k}{\theta}\right)^T w(y_j)}. \end{aligned}$$

Lemma 4.1. Suppose that the distribution function F is nondegenerate and $E_{F_i}||(x, y)||^3 < \infty$ and $k_i \to k_{i0}$, as $N \to \infty$, where $0 < k_{i0} < 1$. Then at the true value $\beta = \beta_0$, with probability 1 in the interior of the interval $||\theta - \theta_0|| \leq N^{-1/3}$ for N large enough, (4.2) uniquely determines v_1 and v_2 and $l_E(\theta, \beta_0)$ attains a local minimum value at some point $\tilde{\theta}$ in the interior of the interval $||\theta - \theta_0|| \leq N^{-1/3}$, where θ_0 is the true value of θ.

Proof. When $\beta = \beta_0$ and $\theta = \theta_0$, we have

$$\begin{aligned}
\frac{1}{N}\sum_{j=1}^{N} h_1 &= \int\int g(x, y, \beta_0) dF(x, y) + O_p(N^{-1/2}) = O_p(N^{-1/2}) \\
\frac{1}{N}\sum_{j=1}^{N} h_2 &= \int\int (w(x, y) - \theta_0) dF(x, y) + O_p(N^{-1/2}) = O_p(N^{-1/2})
\end{aligned}$$

By the approach of Owen (1990), we have when $\theta = \theta_0 + O_p(N^{-1/3})$ and $\beta = \beta_0$, $v_1 = O_p(N^{-1/3})$, $v_2 = O_p(N^{-1/3})$. By the implicit function theorem, (4.2) uniquely determines v_1 and v_2.

Now we rewrite l_E as $l_E = l_{1E}(\theta) + l_{2E}(\theta, \beta_0)$, where $l_{1E}(\theta) = \sum_j \log[(\frac{k}{\theta})^T w(y_j)] + \sum n_i \log(\theta_i)$ and $l_{2E}(\theta, \beta_0) = \sum \log[1 + v_1^T h_1(y_j, x_j, \beta) + v_2^T h_2(y_j, \theta)]$. Let $\theta = \theta_0 + uN^{-1/3}$, where $||u|| = 1$. Denote $ab = (a_1 b_1, \ldots, a_n b_n)^T$ if a and b are column vectors with the same dimension n. Then

$$\begin{aligned}
\frac{\partial l_{1E}}{\partial \theta} &= \sum_{j=1}^{N} \frac{-(\frac{k}{\theta^2} w(y_j))}{(\frac{k}{\theta})^T w(y_j)} + \frac{n}{\theta} = \sum_{j=1}^{N} diag\left\{-\frac{k_1}{\theta_1^2}, \ldots, -\frac{k_I}{\theta_I^2}\right\} \frac{w(y_j) - \theta + \theta}{(\frac{k}{\theta})^T w(y_j)} + \frac{n}{\theta} \\
&= diag\left\{-\frac{k_1}{\theta_1^2}, \ldots, -\frac{k_I}{\theta_I^2}\right\} \sum_{j=1}^{N} h_2 + \sum_{j=1}^{N} diag\left\{-\frac{k_1}{\theta_1^2}, \ldots, -\frac{k_I}{\theta_I^2}\right\} \frac{\theta}{(\frac{k}{\theta})^T w(y_j)} + \sum_{j=1}^{N} \frac{k}{\theta} \\
&= diag\left\{-\frac{k_1}{\theta_1^2}, \ldots, -\frac{k_I}{\theta_I^2}\right\} \sum_{j=1}^{N} h_2 + \sum_{j=1}^{N} \left(\frac{k}{\theta}\right) \left[\frac{-1}{(\frac{k}{\theta})^T w(y_j)} + 1\right]
\end{aligned}$$

Note that $1 = \left(\frac{k}{\theta}\right)^T \theta$, so

$$\frac{\partial l_{1E}}{\partial \theta} = diag\left\{-\frac{k_1}{\theta_1^2}, \ldots, -\frac{k_I}{\theta_I^2}\right\} \sum_{j=1}^{N} h_2 + \sum_{j=1}^{N} \left(\frac{k}{\theta}\right) \left[\frac{-\left(\frac{k}{\theta}\right)^T \theta + (\frac{k}{\theta})^T w(y_j)}{(\frac{k}{\theta})^T w(y_j)}\right]$$

$$= \left[diag\left\{-\frac{k_1}{\theta_1^2}, \ldots, -\frac{k_I}{\theta_I^2}\right\} + \left(\frac{k}{\theta}\right)\left(\frac{k}{\theta}\right)^T \right] \sum_{j=1}^{N} h_2.$$

And

$$\begin{aligned}
\frac{\partial^2 l_{1E}}{\partial\theta\partial\theta^T} &= \sum_{j=1}^{N} \frac{diag\left\{\frac{2k_1}{\theta^3}w_1(y_j), \ldots, \frac{2k_I}{\theta^3}w_I(y_j)\right\}\left(\frac{k}{\theta}\right)^T w(y_j) - \left(\frac{k}{\theta^2}w(y_j)\right)\left(\frac{k}{\theta^2}w(y_j)\right)^T}{\left[(\frac{k}{\theta})^T w(y_j)\right]^2} \\
&+ diag\left\{-\frac{n_1}{\theta_1^2}, \ldots, -\frac{n_I}{\theta_I^2}\right\} \\
&= diag\left\{\frac{2k_1}{\theta_1^3}, \ldots, \frac{2k_I}{\theta_I^3}\right\} \sum_{j=1}^{N} \frac{diag\left\{w_1(y_j) - \theta_1, \ldots, w_I(y_j) - \theta_I\right\}}{(\frac{k}{\theta})^T w(y_j)} \\
&+ \sum_{j=1}^{N} \frac{diag\left\{\frac{2k_1}{\theta_1^2}, \ldots, \frac{2k_I}{\theta_I^2}\right\}}{(\frac{k}{\theta})^T w(y_j)} + diag\left\{-\frac{n_1}{\theta_1^2}, \ldots, -\frac{n_I}{\theta_I^2}\right\} \\
&- diag\left\{\frac{k_1}{\theta_1^2}, \ldots, \frac{k_I}{\theta_I^2}\right\} \sum_{j=1}^{N} \frac{w(y_j)w(y_j)^T}{\left[(\frac{k}{\theta})^T w(y_j)\right]^2} diag\left\{\frac{k_1}{\theta_1^2}, \ldots, \frac{k_I}{\theta_I^2}\right\}.
\end{aligned}$$

Since when $N \to \infty$, $k_i \to k_{i0}$ and

$$\sum_{j=1}^{N} \frac{w(y_j)w(y_j)^T}{\left[(\frac{k}{\theta})^T w(y_j)\right]^2} \to N \sum_{i=1}^{I} k_{i0} E_{F_i}\left(\frac{w(y)w(y)^T}{\left[(\frac{k}{\theta})^T w(y)\right]^2}\right),$$

then

$$\begin{aligned}
l_{1E}(\theta_0 + uN^{-1/3}) &= l_{1E}(\theta_0) + \left(\frac{\partial l_{1E}(\theta_0)}{\partial\theta}\right) uN^{-1/3} + \frac{1}{2}(u^T N^{-1/3}) \frac{\partial^2 l_{1E}(\theta^*)}{\partial\theta\partial\theta^T}(uN^{-1/3}) \\
&= l_{1E}(\theta_0) + O\left((N \log\log N)^{1/2}\right) uN^{-1/3} + O\left(N\right) u^T u N^{-2/3} + o(N^{1/3}) \ a.s. \\
&> l_{1E}(\theta_0)
\end{aligned}$$

where θ^* is between θ_0 and $\theta_0 + uN^{-1/3}$. Similarly, we have $l_{2E}(\theta_0 + uN^{-1/3}, \beta_0) > l_{2E}(\theta_0, \beta_0)$. Thus $l_E(\theta_0 + uN^{-1/3}, \beta_0) > l_E(\theta_0, \beta_0)$. So $l_E(\theta, \beta_0)$ attains a local minimum value at some point $\tilde{\theta}$ in the interior of the interval $||\theta - \theta_0|| \leq N^{-1/3}$.

Note that $\tilde{\theta}$ satisfies

$$\frac{\partial l_E(\theta, \beta_0)}{\partial \theta} = \sum_{j=1}^{N} \frac{-\lambda_2}{1 + \lambda_1^T g(y_j, x_j, \beta) + \lambda_2^T [w(y_j) - \theta]} + \frac{n}{\theta} = 0,$$

so $\lambda_2 = \frac{k}{\theta}$, thus $v_2 = 0$ and we have the following estimating equations

$$\begin{aligned} Q_{1N}(\theta, \lambda) &= \frac{1}{N} \sum_{j=1}^{N} \frac{h_1(x_j, y_j, \theta)}{1 + \lambda^T h_1(x_j, y_j, \theta)} = 0 \\ Q_{2N}(\theta, \lambda) &= \frac{1}{N} \sum_{j=1}^{N} \frac{h_2(x_j, y_j, \theta)}{1 + \lambda^T h_1(x_j, y_j, \theta)} = 0. \end{aligned} \tag{4.3}$$

And

$$l_E(\tilde{\theta}) = l_{1E}(\tilde{\theta}) + l_{2E}(\tilde{\theta}), \quad l_{2E}(\tilde{\theta}) = \sum_{j=1}^{N} \log[1 + \tilde{\lambda}^T h_1(x_j, y_j, \theta)],$$

where $\tilde{\theta}$ and $\tilde{\lambda}$ satisfy (4.3) and $h_1(x_j, y_j, \theta)$ is evaluated at $\beta = \beta_0$.

Lemma 4.2. With the same condition of Lemma 4.1, under $H_0 : \beta = \beta_0$ is true, then

$$\begin{pmatrix} \sqrt{N}(\tilde{\lambda} - 0) \\ \sqrt{N}(\tilde{\theta} - \theta_0) \end{pmatrix} \to N(0, U)$$

Proof. With Taylor expansion for $Q_{iN}(\tilde{\theta}, \tilde{\lambda}), i = 1, 2$, we have

$$0 = Q_{iN}(\theta_0, 0) + \frac{\partial Q_{iN}(\theta_0, 0)}{\partial \theta^T}(\tilde{\theta} - \theta_0) + \frac{\partial Q_{iN}(\theta_0, 0)}{\partial \lambda^T}(\tilde{\lambda} - 0) + o_p(\varepsilon_N),$$

where $\varepsilon_N = ||\tilde{\theta} - \theta_0|| + ||\tilde{\lambda} - 0||$. So

$$\begin{pmatrix} \tilde{\lambda} - 0 \\ \tilde{\theta} - \theta_0 \end{pmatrix} = -S_N^{-1} Q_N(\theta_0, 0) + o_p(\varepsilon_N),$$

where $Q_N(\theta, \lambda) = \begin{pmatrix} Q_{1N}(\theta, \lambda) \\ Q_{2N}(\theta, \lambda) \end{pmatrix}$ and

$$\begin{aligned} S_N &= \begin{pmatrix} \frac{\partial Q_{1N}(\theta_0,0)}{\partial \lambda} & \frac{\partial Q_{1N}(\theta_0,0)}{\partial \theta} \\ \frac{\partial Q_{2N}(\theta_0,0)}{\partial \lambda} & \frac{\partial Q_{2N}(\theta_0,0)}{\partial \theta} \end{pmatrix} \\ &= \begin{pmatrix} -\frac{1}{N}\sum_{j=1}^N h_1(\theta_0)h_1(\theta_0)^T & \frac{1}{N}\sum_{j=1}^N \frac{\partial h_1(\theta_0)}{\partial \theta} \\ -\frac{1}{N}\sum_{j=1}^N h_2(\theta_0)h_1(\theta_0)^T & \frac{1}{N}\sum_{j=1}^N \frac{\partial h_2(\theta_0)}{\partial \theta} \end{pmatrix} \\ &\to \begin{pmatrix} -\sum_{i=1}^I k_i E_{F_i}(h_1 h_1^T) & \sum_{i=1}^I k_i E_{F_i}(\frac{\partial h_1}{\partial \theta}) \\ -\sum_{i=1}^I k_i E_{F_i}(h_2 h_1^T) & \sum_{i=1}^I k_i E_{F_i}(\frac{\partial h_2}{\partial \theta}) \end{pmatrix} = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} = S. \end{aligned}$$

Also note that $Q_N(\theta_0, 0) = \begin{pmatrix} \frac{1}{N}\sum_{j=1}^N h_1(\theta_0) \\ \frac{1}{N}\sum_{j=1}^N h_2(\theta_0) \end{pmatrix}$, then we have

$$\sqrt{N} Q_N(\theta_0, 0) \to N(0, V),$$

where

$$V = \begin{pmatrix} \sum_{i=1}^{I} k_i \left[E_{F_i}(h_1 h_1^T) - E_{F_i}(h_1)E_{F_i}(h_1)^T\right] & \xi \\ \xi^T & \sum_{i=1}^{I} k_i \left[E_{F_i}(h_2 h_2^T) - E_{F_i}(h_2)E_{F_i}(h_2)^T\right] \end{pmatrix},$$

here

$$\begin{aligned} \xi &= \sum_{i=1}^{I} k_i^2 E_{F_i}(h_1 h_2^T) + \sum_{i=1}^{I}\sum_{j\neq i} k_i k_j E_{F_{ij}}(h_1 h_2^T) - \left[\sum_{i=1}^{I} k_i E_{F_i}(h_1)\right]\left[\sum_{i=1}^{I} k_i E_{F_i}(h_2)\right]^T \\ &= \sum_{i=1}^{I} k_i \left[E_{F_i}(h_1 h_2^T) - E_{F_i}(h_1)E_{F_i}(h_2^T)\right]. \end{aligned}$$

Then $U = S^{-1}V(S^{-1})^T$.

By the above lemmas, we can prove the following result.

Theorem 4.3. Under the conditions in Lemma 4.1, if $H_0 : \beta = \beta_0$ is true, then the empirical likelihood statistic satisfies

$$R(\beta_0) \to \chi^2_{(p)}.$$

Proof. Since $l(\hat{\theta}) = -l_{1E}(\hat{\theta}) - N \log N$, where $Q_{2N}(\hat{\theta}, 0) = 0$, by Taylor expansion, we have

$$\begin{aligned} \hat{\theta} - \theta_0 &= -\left[\frac{\partial Q_{2N}(\theta_0, 0)}{\partial \theta^T}\right]^{-1} Q_{2N}(\theta_0, 0) + o_p(N^{-1/2}) \\ &= S_{22}^{-1}(0, I)Q_N(\theta_0, 0) + o_p(N^{-1/2}). \end{aligned}$$

From the equation above and the result of Lemma 4.2,

$$\begin{aligned}
\tilde{\theta}-\hat{\theta} &= -(0,I)S_N^{-1}Q_N(\theta_0,0)+S_{22}^{-1}(0,I)Q_N(\theta_0,0)+o_p(N^{-1/2}) \\
&= -(0,I)\left[I-\begin{pmatrix}0\\ I\end{pmatrix}S_{22}^{-1}(0,I)S_N\right]S_N^{-1}Q_N(\theta_0,0)+o_p(N^{-1/2}) \\
&= -S_{22}^{-1}S_{21}(I,0)S_N^{-1}Q_N(\theta_0,0)+o_p(N^{-1/2}) \\
&= -S_{22}^{-1}S_{21}\tilde{\lambda}+o_p(N^{-1/2})
\end{aligned}$$

Now we expand $l_{1E}(\tilde{\theta})$ at $\hat{\theta}$, then

$$l_{1E}(\tilde{\theta})-l_{1E}(\hat{\theta})=\frac{\partial l_{1E}(\hat{\theta})}{\partial\theta^T}(\tilde{\theta}-\hat{\theta})+\frac{1}{2}(\tilde{\theta}-\hat{\theta})^T\frac{\partial^2 l_{1E}(\hat{\theta})}{\partial\theta\partial\theta^T}(\tilde{\theta}-\hat{\theta})+o_p(1).$$

Note that $Q_{2N}(\hat{\theta},0)=0$, so we have

$$\frac{\partial l_{1E}(\hat{\theta})}{\partial\theta^T}=\sum_{j=1}^{N}\frac{-\frac{kw}{\hat{\theta}^2}}{(\frac{k}{\hat{\theta}})^T w(y_j)}+\frac{n}{\hat{\theta}}=-N\left[diag\left\{-\frac{k_1}{\theta_1^2},\ldots,-\frac{k_I}{\theta_I^2}\right\}-\left(\frac{k}{\theta}\right)\left(\frac{k}{\theta}\right)^T\right]Q_{2N}=0.$$

From $Q_{1N}(\tilde{\theta},\tilde{\lambda})=0$, we have

$$\begin{aligned}
l_{2E}(\tilde{\theta},\beta_0) &= \sum_{j=1}^{N}\log[1+\tilde{\lambda}^T h_1(\tilde{\theta})] \\
&= \sum_{j=1}^{N}\tilde{\lambda}^T h_1(\tilde{\theta})-\frac{1}{2}\sum_{j=1}^{N}\tilde{\lambda}^T h_1(\tilde{\theta})h_1(\tilde{\theta})^T\tilde{\lambda}+o_p(1) \\
&= \frac{N}{2}\tilde{\lambda}^T S_{11}\tilde{\lambda}+o_p(1)
\end{aligned}$$

So

$$\begin{aligned}
l_E(\tilde{\theta}, \beta_0) - l_{1E}(\hat{\theta}) &= l_{1E}(\tilde{\theta}) - l_{1E}(\hat{\theta}) + l_{2E}(\tilde{\theta}, \beta_0) \\
&= \frac{N}{2}\tilde{\lambda}^T \left[(S_{22}^{-1}S_{21})^T \left(\frac{1}{N} \frac{\partial^2 l_{1E}(\hat{\theta})}{\partial\theta\partial\theta^T} \right) (S_{22}^{-1}S_{21}) + S_{11} \right] \tilde{\lambda} + o_p(1)
\end{aligned}$$

By Lemma 4.2, we know that $\tilde{\lambda} \sim N(0, U_{11})$, so we only need to show

$$\left[(S_{22}^{-1}S_{21})^T \left(\frac{1}{N} \frac{\partial^2 l_{1E}(\hat{\theta})}{\partial\theta\partial\theta^T} \right) (S_{22}^{-1}S_{21}) + S_{11} \right] U_{11}$$

is idempotent. Let $A = (S_{22}^{-1}S_{21})^T \left(\frac{1}{N} \frac{\partial^2 l_{1E}(\hat{\theta})}{\partial\theta\partial\theta^T} \right) (S_{22}^{-1}S_{21}) + S_{11}$. From the proof of Lemma 4.1, we know that

$$\begin{aligned}
\frac{\partial^2 l_{1E}}{\partial\theta\partial\theta^T} &= diag\left\{ \frac{2k_1}{\theta_1^3}, \ldots, \frac{2k_I}{\theta_I^3} \right\} \sum_{j=1}^N \frac{diag\left\{ w_1(y_j) - \theta_1, \ldots, w_I(y_j) - \theta_I \right\}}{(\frac{k}{\theta})^T w(y_j)} \\
&+ \sum_{j=1}^N \frac{diag\left\{ \frac{2k_1}{\theta_1^2}, \ldots, \frac{2k_I}{\theta_I^2} \right\}}{(\frac{k}{\theta})^T w(y_j)} + diag\left\{ -\frac{n_1}{\theta_1^2}, \ldots, -\frac{n_I}{\theta_I^2} \right\} \\
&- diag\left\{ \frac{k_1}{\theta_1^2}, \ldots, \frac{k_I}{\theta_I^2} \right\} \sum_{j=1}^N \frac{w(y_j)w(y_j)^T}{\left[(\frac{k}{\theta})^T w(y_j) \right]^2} diag\left\{ \frac{k_1}{\theta_1^2}, \ldots, \frac{k_I}{\theta_I^2} \right\}.
\end{aligned}$$

Note that

$$\sum_{j=1}^N \frac{diag\left\{ \frac{k_1}{\theta_1^2}, \ldots, \frac{k_I}{\theta_I^2} \right\}}{(\frac{k}{\theta})^T w(y_j)} + diag\left\{ -\frac{n_1}{\theta_1^2}, \ldots, -\frac{n_I}{\theta_I^2} \right\}$$

$$
\begin{aligned}
&= \sum_{j=1}^{N} \frac{diag\left\{\frac{k_1}{\theta_1^2},\ldots,\frac{k_I}{\theta_I^2}\right\}\left[1-(\frac{k}{\theta})^T w(y_j)\right]}{(\frac{k}{\theta})^T w(y_j)} \\
&= \sum_{j=1}^{N} \frac{diag\left\{\frac{k_1}{\theta_1^2},\ldots,\frac{k_I}{\theta_I^2}\right\}\left[(\frac{k}{\theta})^T (w(y_j)-\theta)\right]}{(\frac{k}{\theta})^T w(y_j)} \\
&= diag\left\{\frac{k_1}{\theta_1^2},\ldots,\frac{k_I}{\theta_I^2}\right\}\left[(\frac{k}{\theta})^T \sum_{j=1}^{N} \frac{(w(y_j)-\theta)}{(\frac{k}{\theta})^T w(y_j)}\right]
\end{aligned}
$$

Since $Q_{2N}(\hat{\theta},0)=0$, then

$$
diag\left\{\frac{2k_1}{\theta_1^3},\ldots,\frac{2k_I}{\theta_I^3}\right\}\sum_{j=1}^{N}\frac{diag\left\{w_1(y_j)-\theta_1,\ldots,w_I(y_j)-\theta_I\right\}}{(\frac{k}{\theta})^T w(y_j)}\Bigg|_{\theta=\hat{\theta}} = 0
$$

and

$$
\sum_{j=1}^{N}\frac{diag\left\{\frac{k_1}{\theta_1^2},\ldots,\frac{k_I}{\theta_I^2}\right\}}{(\frac{k}{\theta})^T w(y_j)} + diag\left\{-\frac{n_1}{\theta_1^2},\ldots,-\frac{n_I}{\theta_I^2}\right\}\Bigg|_{\theta=\hat{\theta}} = 0.
$$

So

$$
\frac{1}{N}\frac{\partial^2 l_{1E}(\hat{\theta})}{\partial\theta\partial\theta^T} = diag\left\{\frac{k_1}{\hat{\theta}_1^2},\ldots,\frac{k_I}{\hat{\theta}_I^2}\right\} + \sum_{i=1}^{I} k_{i0} E_{F_i}\left(\frac{w(y)w(y)^T}{\left[(\frac{k}{\hat{\theta}})^T w(y)\right]^2}\right) + o_p(1).
$$

Thus we have $AU_{11}A = A$. Hence

$$
R(\beta_0) = 2[l(\hat{\beta}) - l(\hat{\theta},\beta_0)] = 2[l_E(\tilde{\theta},\beta_0) - l_{1E}(\hat{\theta})] \to \chi^2_{(p)}.
$$

4.3 Simulation Results

In this section, we will show the numerical results to evaluate our method. We generate the data (X, Y) from bivariate normal distribution with mean vector μ and variance-covariance matrix Σ. We set $\mu_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $\mu_2 = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$ and $\Sigma_1 = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$, $\Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ for dependent and independent cases, respectively. Also we chose $I = 3$ and $w_1(y) = I_{[y \geq d]}, w_2(y) = I_{[y \leq c]}, w_3(y) = 1$. The estimating function is given by $g(x, y, \beta) = x + y - \beta$.

To check the normality of the estimate for λ and θ, Table 4.1, 4.2 and 4.3 present the results obtained from the dependent and independent cases with different c and d. The tables include the estimated bias (Bias) given by the average of the estimates minus the true value, the average of the estimated standard errors (ASE), the sample standard error of the estimates (SE), and the 95% empirical coverage probability (CP). The estimations and the confidence intervals for β are given in Table 4.4 as well as the empirical coverage probability. All the results are based on 1000 replications.

Figure 4.1 is the Q-Q plot of standardized λ and θ versus the standard normal distribution. Figure 4.2 shows the Q-Q plot for $R(\beta_0)$ versus standard $\chi^2_{(1)}$. The plot shows the approximation appears satisfactory.

4.4 Discussion

In this chapter, we proposed a nonparametric estimate for the biased sample problems with empirical likelihood method. One big advantage of our method is it does not rely on the distribution assumption of the variables. The inference shows the likelihood ratio statistic follows a chi-square distribution asymptotically. This result is very useful in data analysis, especially for the hypothesis testing.

The simulation results show our method performs well. When more concentration was put on the sampling with extreme tails in the output, larger sample size is needed to get the same efficiency. The results are stable regardless the dependence of X and Y.

The information about the parameter of interest is given by the estimation equations in this chapter. But sometimes the parameter may be defined differently. For example, the β can be defined by regression quantile

$$E[I(y \leq x\beta) - 0.5] = 0.$$

Furthermore, the weight functions can be assumed to be parametric, say $w(\cdot, \xi)$, for some parameter ξ (Sun and Woodroofe, 1997 and Gilber, Lele and Vardi, 1999). This makes the weight functions more flexible. These situations will be considered in the near future.

Chapter 5

Future Research

As mentioned in the previous chapters, current status data, as a special case of interval-censored data, play a very important role in real life. We proposed different methods dealing with univariate and bivariate current status data in Chapter 2 and Chapter 3, respectively. We also discussed another type of missing information, the biased sampling problem, in Chapter 4. But there are still a lot of unsolved problems, some of them may be the direction of my future research work. In this chapter, I will list only three of them.

5.1 Regression Analysis of Multivariate Current Status Data

At the end of Section 1.4, we discussed the recent development of the analysis for multivariate failure time data. The literature is very limited especially for multivariate current status data. Marginal approach and random effect approach are the most widely used methods for the regression analysis with multivariate interval-censored data. And most of the existing methods are based on the proportional hazards models. Since the numerous applications of different models in the analysis of univariate or bivariate current status data, such as the linear transformation model we used in Chapter 2, we hope to apply different models to multivariate case to model the marginal survival functions for the failure times. But one difficulty is to model the structures of the correlation among the failure times. There is not a good way to simply measure the correlation so far. It will be a big challenging.

Thus the regression analysis for multivariate current status data with the linear transformation model may be a future research topic.

5.2 Model Checking and Model Selection

In Chapter 3, we discussed the efficient estimation for the proportional odds model with bivariate current status data. We assume the marginal distributions follow the PO models. This may not be true in some cases. In the application, we used the nonparametric way to check the validation about the PO model. But the method is very rough. The existing method for model checking is usually based on counting processes which is quite complicated. Moreover, the residual analysis can only be used after a model is presented. But in practice, some prior study just based on the data may be more of interest. Selecting the "best" model to fit the data set is somewhat important.

The model checking and model selection is a difficult topic in the regression analysis especially for interval-censored data. It may be a very challenging but very helpful topic of my future work.

5.3 Empirical Likelihood Method in Survival Analysis

In Chapter 4, we discussed the biased sampling problems with empirical likelihood method. The advantage of the nonparametric method is it does not rely on the assumption of the underlying distribution. There are a lot of papers about empirical likelihood method, however, its application in survival analysis is still in a very underdeveloped stage. Pan and Zhou (2002) studied empirical likelihood ratios for right censored data. The parameters in that paper were given by linear functionals of the cumulative hazard function. They also showed the empirical likelihood ratio in some special setting can be obtained by solving a one parameter monotone equation. This is one of the few works about empirical likelihood in survival analysis. To my best knowledge, there is no research study about the application of empirical likelihood for interval-censored data. Even for right censored data, the study is very limited. One difficulty of the application of empirical method on survival data is the complicated form of the empirical likelihood function. Another difficult part for this problem is how to choose an appropriate constrain. One research topic I am very interested in and planning to continue in the future is to develop an easy way to apply empirical likelihood in survival analysis, especially for interval-censored data.

There exists several other directions or topics for future research. As one extension of the topic in Chapter 2, we are interested in the estimation of the information bound $I(\theta_0)$ for general situations, which involves some conditional expectations given observation variable. Another is that in Chapter 2, we only considered the situation where covariates are time-independent. In some cases, this may not be true and it would be useful to develop approaches that can handle time-dependent covariates. Also it would be helpful to derive the asymptotic process of $\hat{H}_n$. All the problems described above are very interesting and may become the directions of my future research.

Appendix

Appendix A Proof of asymptotic properties of $\hat{\theta}$ in Chapter 2

A.1 Proof of equations (2.5) and (2.6)

Without loss of generality, we only consider the case of $p = 1$, that is $Z \in R$. The proof is similar for general situations. To derive the efficient score function for θ_0, we first need to obtain the score functions for θ and H. By definition, the score function for θ is given by

$$\dot{l}_\theta = \frac{\partial l_n(\theta, H)}{\partial \theta} = Z\lambda(H(C) + Z'\theta)\,[\delta \frac{q}{1-q} - (1-\delta)]\,,$$

where $q = \exp(-\Lambda(H_0(C) + Z'\theta_0)))$. Suppose that $\mathcal{H}_0 = \{H_\eta, |\eta| < 1\}$ is a regular parametric subfamily of the class of all monotone functions H over R^+. Set $\partial H_\eta(t)/\partial\eta|_{\eta=0} = b(t)$. Then the score function for H is given by

$$\dot{l}_H(b) = \frac{\partial l_n(\theta, H_\eta)}{\partial \eta} = b\lambda(H(C) + Z'\theta)\,[\delta \frac{q}{1-q} - (1-\delta)]$$

with $b = b(C)$. To determine the efficient score for θ_0, we need to find b^* such that $\dot{l}^*_\theta \,=\, \dot{l}_\theta - \dot{l}_H(b^*)$ is orthogonal to $\dot{l}_H(b)$ for any b. That is, $E\{\,[\dot{l}_\theta - \dot{l}_H(b^*)]\dot{l}_H(b)\,\} \,=\, 0$. This yields

$$E\lambda^2(H(C) + Z'\theta)[\delta\frac{q}{1-q} - (1-\delta)]^2[Z - b^*]b \,=\, 0\,.$$

Solving this equation gives

$$b^*(C) \,=\, \frac{E[Z\lambda^2(H(C)+Z'\theta)(\delta\frac{q}{1-q} - (1-\delta))^2|C]}{E[\lambda^2(H(C)+Z'\theta)(\delta\frac{q}{1-q} - (1-\delta))^2|C]} \,=\, \frac{E[Z\lambda^2(H(C)+Z'\theta)\frac{q}{1-q}|C]}{E[\lambda^2(H(C)+Z'\theta)\frac{q}{1-q}|C]}\,.$$

It follows that the efficient score for θ_0 has the form

$$\dot{l}^*_\theta \,=\, \lambda(H(C)+Z'\theta)[\delta\frac{q}{1-q} - (1-\delta)]\,\{Z - b^*(C)\}$$

and the information bound for θ_0 is given by

$$I(\theta) \,=\, E[\dot{l}^*_\theta]^{\otimes 2} \,=\, E\lambda^2(H(C)+Z'\theta)\frac{q}{1-q}\Big\{Z - \frac{E[Z\lambda^2(H(C)+Z'\theta)\frac{q}{1-q}|C]}{E[\lambda^2(H(C)+Z'\theta)\frac{q}{1-q}|C]}\Big\}^{\otimes 2}\,.$$

This completes the proof of equations (2.5) and (2.6).

A.2 Proof of the normality of $\hat{\theta}$

For the proof, the key is to verify

$$S_{1n}(\widehat{\theta}_n, \widehat{H}_n) = o_p(n^{-1/2})\ ,\ S_{2n}(\widehat{\theta}_n, \widehat{H}_n)[\psi^*] = o_P(n^{-1/2})\,,$$

the first assumption of Theorem 6.1 of Huang (1996), where S_{1n}, S_{2n} and ψ^* are defined below. Then the normality of $\hat{\theta}_n$ directly follows Theorem 6.1 of Huang (1996) since all other conditions there can be easily shown to hold.

For this, for any fixed $H \in \mathcal{H}$, let $\{H(t) : t \text{ in a neighborhood of } 0 \in R\}$ be a smooth curve in $\mathcal{H}$ running through H at $t = 0$. Let $\psi = \partial H(t)/\partial|_{t=0}$ and let Ψ be the collection of all ψ defined above. Define

$$l_1(\theta, H; x) = \frac{\partial}{\partial\theta} l_{\theta,H}(x)\ ,\ l_2(\theta, H; x)[\psi] = \frac{\partial}{\partial t} l_{\theta,H(t)}(x)|_{t=0}\ ,$$

$S_{1n}(\theta, H) = P_n l_1(\theta, H, X)$ and $S_{2n}(\theta, H)[\psi] = P_n l_2(\theta, H, X)[\psi]$. Here $P = P_{\theta_0, H_0}$ and $Pf = \int f(x)dP(x)$. Also let

$$r(c, z; \theta, H) = \frac{\exp(-\Lambda(H(c) + z'\theta))}{1 - \exp(-\Lambda(H(c) + z'\theta))}\,.$$

Note that the score function for θ is $l_1(\theta, H; x) = z\lambda(H(c) + z'\theta))(\delta r(c, z; \theta, H) - (1 - \delta))$ from the proof of (2.5) and (2.6). This together with (2.5) implies that

$$S_{1n}(\widehat{\theta}_n, \widehat{H}_n) = 0\,.$$

To prove that $S_{2n}(\widehat{\theta}_n, \widehat{H}_n)[\psi^*] = o_P(n^{-1/2})$, note that for any $\psi \in L_2(P_C)$ (P_C is the marginal distribution of C), we have

$$l_2(\theta, H; x)[\psi] = \psi(c)\lambda(H(c) + z'\theta))[\delta r(c, z; \theta, H) - (1 - \delta)]$$

and l_2 can be considered as the derivative $(\partial/\partial\epsilon)l(\theta, H + \epsilon\psi; x)|_{\epsilon=0}$. Define

$$\psi^*(c) = \frac{E[Z\lambda^2(H(C) + Z'\theta)r(C, Z; \theta, H)|C = c]}{E[\lambda^2(H(C) + Z'\theta)r(C, Z; \theta, H)|C = c]}.$$

From the proof of (2.5) and (2.6), $l_1(\theta, H; x) - l_2(\theta, H; x)[\psi^*]$ is orthogonal to $l_2(\theta, H; x)[\psi]$ in $L_2^0(P)$ for any $\psi \in L_2(P_C)$.

Furthermore, from the condition (A4), $\xi_0 = \psi^* \circ H_0^{-1}$ is well defined on the range of H_0 and $\xi(\widehat{H}_n)$ is a right continuous step function at the jump point c_i. Also from the property of the estimator $\widehat{H}_n$, we have

$$P_n[\xi(\widehat{H}_n(c))\lambda(\widehat{H}_n(c) + z'\widehat{\theta}_n))[\delta r(c, z; \widehat{\theta}_n, \widehat{H}_n) - (1 - \delta)]] = 0\,.$$

and ξ_0 also has bounded derivative since ψ^* has bounded derivative from the conditions (A2), (A3), (A4) and (A5). It follows from $\psi^* = \xi_0 \circ H_0$ that

$$S_{2n}(\widehat{\theta}_n, \widehat{H}_n)[\psi^*] = P_n\Big[\psi^*(c)\lambda(\widehat{H}_n(c) + z'\widehat{\theta}_n))[\delta r(c, z; \widehat{\theta}_n, \widehat{H}_n) - (1 - \delta)]\Big]$$

$$= P_n\Big[\Big(\xi_0 \circ H_0(c) - \xi_0 \circ \widehat{H}_n(c)\Big)\lambda(\widehat{H}_n(c) + z'\widehat{\theta}_n)\Big(\delta r(c, z; \widehat{\theta}_n, \widehat{H}_n) - (1 - \delta)\Big)\Big]$$

$$= (P_n - P)\phi(x; \widehat{\theta}_n, \widehat{H}_n) + P\phi(x; \widehat{\theta}_n, \widehat{H}_n),$$

where

$$\phi(x; \theta, H) = [(\xi_0 \circ H_0(c) - \xi_0 \circ H(c))\lambda(H(c) + z'\theta)(\delta r(c, z; \theta, H) - (1 - \delta))].$$

For any $\eta > 0$, define the class of functions $\Phi(\eta) = \{\phi(x;\theta,H) : \quad |\theta - \theta_0| + \|H - H_0\|_2 \leq \eta$ and $H \in \mathcal{H}\}$. Using the same arguments in the proof of Lemma 7.1 in Huang (1996), one can show that Φ is a Donsker class, which implies

$$\sup_{\phi\in\Phi(M_0 n^{-1/3})} (P_n - P)\phi(x;\theta,H) = o_p(n^{-1/2}).$$

Furthermore, by using the Cauchy-Schwartz inequality, (2.7)-(2.8) and the mean value theorem, we have that

$$P\phi(x;\widehat{\theta}_n,\widehat{H}_n) \leq M_0(P[H_0(c) - \widehat{H}(c)]^2)^{1/2}(P[\exp(-\Lambda(\widehat{H}_n(c) + z'\widehat{\theta}_n))$$

$$- \exp(-\Lambda(H_0(c) + z'\theta_0))]^2)^{1/2} = O_p(n^{-2/3})\,,$$

where $M_0 = \sup|\lambda(\cdot)| \times \sup|d\xi_0(t)/dt| \times \sup[1 - \exp(-\Lambda(\cdot))]^{-1}$. The three supreme exist from the conditions (A2), (A3), (A4) and (A5). The above inequality plus $S_{1n}(\widehat{\theta}_n,\widehat{H}_n) = 0$ shows that the first assumption of Theorem 6.1 in Huang (1996) holds. This completes the proof.

Appendix B Proof of asymptotic properties of $\hat{\theta}$ in Chapter 3.

B.1 Derivation of the Efficient Score and the Information Bound

To derive the efficient score function for θ, first note that

$$\begin{aligned} \dot{l}_\theta &= \frac{\partial l(\theta, O_{01}, O_{02})}{\partial \theta} = \sum_{j=0}^{1}\sum_{m=0}^{1}\int \frac{a_{jm1}X_{jm1} + a_{jm2}X_{jm2}}{S_{jm}}\, dN_{jm} \\ &= \sum_{j=0}^{1}\sum_{m=0}^{1}\int \frac{a_{jm1}X_{jm1} + a_{jm2}X_{jm2}}{S_{jm}}\, dM_{jm}\,. \end{aligned}$$

If we set $\partial log O_{0k}(t)/\partial\eta = b_k(t)$ for $k = 1, 2$, the score function for S_k has the form

$$\dot{l}_{S_1,S_2}(b_1, b_2) = \sum_{j=0}^{1}\sum_{m=0}^{1}\int \frac{a_{jm1}b_1 + a_{jm2}b_2}{S_{jm}} dN_{jm} = \sum_{j=0}^{1}\sum_{m=0}^{1}\int \frac{a_{jm1}b_1 + a_{jm2}b_2}{S_{jm}} dM_{jm}\,.$$

By definition, to determine the efficient score function for θ, we need to find a function b_1^* and b_2^* so that for any b_1 and b_2, we have $\dot{l}_{\theta^*} \perp \dot{l}_{S_1,S_2}(b_1, b_2)$ or

$$E\Big\{\Big(\dot{l}_\theta - \dot{l}_{S_1,S_2}(b_1^*, b_2^*)\Big)\, \dot{l}_{S_1,S_2}(b_1, b_2)\Big\} = 0\,, \qquad (B.1)$$

where

$$\dot{l}_{\theta^*} = \dot{l}_\theta - \dot{l}_{S_1,S_2}(b_1^*, b_2^*) = \sum_{j=0}^{1}\sum_{m=0}^{1}\int \frac{a_{jm1}(X_{jm1} - b_1^*) + a_{jm2}(X_{jm2} - b_2^*)}{S_{jm}} dM_{jm}\,.$$

Note that

$$
\begin{aligned}
&E\Big\{(\dot{l}_\theta - \dot{l}_{S_1,S_2}(b_1^*, b_2^*))\dot{l}_{S_1,S_2}(b_1, b_2)\Big\} \\
= \quad & \sum_{j=0}^{1}\sum_{m=0}^{1} E\Big\{\int \frac{a_{jm1}(X_{jm1} - b_1^*) + a_{jm2}(X_{jm2} - b_2^*)}{S_{jm}} dM_{jm} \int \frac{a_{jm1}b_1 + a_{jm2}b_2}{S_{jm}} dM_{jm}\Big\} \\
= \quad & \sum_{j=0}^{1}\sum_{m=0}^{1} E\Big\{\int \frac{a_{jm1}(X_{jm1} - b_1^*) + a_{jm2}(X_{jm2} - b_2^*)}{S_{jm}} (a_{jm1}b_1 + a_{jm2}b_2) Y\lambda_c dt\Big\}\,.
\end{aligned}
$$

Thus it follows from (A.1) that we have

$$
\sum_{j=0}^{1}\sum_{m=0}^{1} E\left\{\frac{a_{jm1}(X_{jm1} - b_1^*) + a_{jm2}(X_{jm2} - b_2^*)}{S_{jm}} a_{jm1} Y\lambda_c\right\} = 0\,, \qquad (B.2)
$$

$$
\sum_{j=0}^{1}\sum_{m=0}^{1} E\left\{\frac{a_{jm1}(X_{jm1} - b_1^*) + a_{jm2}(X_{jm2} - b_2^*)}{S_{jm}} a_{jm2} Y\lambda_c\right\} = 0\,. \qquad (B.3)
$$

Solving these equations, we obtain that $(b_1^*,\ b_2^*) = HG^{-1}$. This gives the efficient score function and the information bound for θ given in Section 3.3.

B.2 Proof of Equation (3.5)

Define

$$
\begin{aligned}
U_0(\theta_0, O_{01}, O_{02}) \quad &= \quad \frac{1}{\sqrt{n}} \sum_{i=1}^{n}\sum_{j=0}^{1}\sum_{m=0}^{1} \int \frac{(X_{jmi} - HG^{-1})A_{jmi}}{S_{jmi}} dN_{jmi}(t) \\
&= \quad \frac{1}{\sqrt{n}} \sum_{i=1}^{n}\sum_{j=0}^{1}\sum_{m=0}^{1} \int \frac{(X_{jmi} - HG^{-1})A_{jmi}}{S_{jmi}} dM_{jmi}(t)\,.
\end{aligned}
$$

Then it is easy to show that

$$U(\theta_0, O_{01}, O_{02}) \;=\; U_0(\theta_0, O_{01}, O_{02}) + o_p(1)\,. \qquad (B.4)$$

Thus to show (3.5), it is sufficient to show that

$$U(\theta_0, \hat{O}_{01}, \hat{O}_{02}) \;=\; U_0(\theta_0, O_{01}, O_{02}) + o_p(1)\,. \qquad (B.5)$$

For (A.5), note that we can rewrite $U(\theta_0, \hat{O}_{01}, \hat{O}_{02})$ as

$$\begin{aligned} U(\theta_0, \hat{O}_{01}, \hat{O}_{02}) \;&=\; \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=0}^{1} \sum_{m=0}^{1} \int \frac{(\hat{X}_{jmi} - \hat{H}_n \hat{G}_n^{-1})\hat{A}_{jmi}}{\hat{S}_{jmi}} dM_{jmi} \\ &\quad + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=0}^{1} \sum_{m=0}^{1} \int \frac{(\hat{X}_{jmi} - \hat{H}_n \hat{G}_n^{-1})\hat{A}_{jmi}}{\hat{S}_{jmi}} S_{jmi} Y_i \lambda_c dt \\ &=\; V_{1n}(\theta_0, \hat{O}_{01}, \hat{O}_{02}) \;+\; V_{2n}(\theta_0, \hat{O}_{01}, \hat{O}_{02})\,. \qquad (B.6) \end{aligned}$$

Let $h_{jmi}(O_{01}, O_{02}) = (X_{jmi} - HG^{-1})A_{jmi}/S_{jmi}$,

$$h_{jmiu} \;=\; \frac{\partial h_{jmi}(u,v)}{\partial u}|_{u=O_{01}, v=O_{02}}\;,\; h_{jmiv} \;=\; \frac{\partial h_{jmi}(u,v)}{\partial v}|_{u=O_{01}, v=O_{02}}\,.$$

Applying the Taylor series expansion of $V_{1n}(\theta_0, \hat{O}_{01}, \hat{O}_{02})$ around (O_{01}, O_{02}) and noting the bounded property of the second derivatives function of h_{jmi}, we have that

$$V_{1n} = U_0(\theta_0, O_{01}, O_{02}) + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=0}^{1} \sum_{m=0}^{1} \int \Big[h_{jmiu}(\hat{O}_{01} - O_{01}) + h_{jmiv}(\hat{O}_{02} - O_{02}) \Big] \, dM_{jmi}(t) + o_p(1)$$

Note that both $\hat{O}_{0k}$ and O_{0k} are independent of i. Then the application of Lemma A.1 of Lin and Ying (2001) yields

$$V_{1n}(\theta_0, \hat{O}_{01}, \hat{O}_{02}) \; = \; U_0(\theta_0, O_{01}, O_{02}) \; + \; o_p(1) \, . \qquad (B.7)$$

For V_{2n}, again using the Taylor series expansion around (O_{01}, O_{02}) and noting that $\sum_{j,m} \hat{X}_{jmi} \hat{A}_{jmi} = \partial \sum_{j,m} \hat{S}_{jmi} / \partial \theta = 0$ and $\sum_{j,m} \hat{A}_{jmi} = 0$, we obtain

$$\begin{aligned} V_{2n} \quad &= \quad \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=0}^{1} \sum_{m=0}^{1} \int h_{jmi} \Big[a_{jm1i}(\hat{O}_{01} - O_{01}) + a_{jm2i}(\hat{O}_{02} - O_{02}) \Big] Y_i \lambda_c dt + o_p(1) \\ &= \quad \sum_{k=1}^{2} \int (\hat{O}_{0k} - O_{0k}) \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=0}^{1} \sum_{m=0}^{1} h_{jmi} a_{jmki} Y_i \lambda_c dt + o_p(1) \, . \end{aligned}$$

It then follows immediately from (B.2) and (B.3) that $V_{2n}(\theta_0, O_{01}, O_{02}) = o_p(1)$. This together with (B.7) proves (B.5) and thus (3.5).

BIBLIOGRAPHY

Anderson, G. L. and Fleming, T. R. (1995). Model misspecification in proportional hazards regression. *Biometrika*, 82, 527-541.

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

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T. and Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics*, 26, 641-647.

Bickel, P., Klaassen, C., Ritov, Y. and J. Wellner (1993). Efficient and adaptive estimation for semiparametric models. The Johns Hopkins University Press.

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

Chen, H. (2001). Weighted semiparametric likelihood method for fitting a proportional odds regression model to data from the case-cohort design. *Journal of the American Statistical Association*, 96, 1446-1457.

Chen, H. and Little, R. J. A. (1999). Proportional hazards regression with missing covariates. *Journal of the American Statistical Association*, 94, 896-908.

Chen, K., Jin, Z. and Ying, Z. (2002). Semiparametric analysis of transformation models with censored data. *Biometrika*, 89, 659-668.

Cheng, S. C., Wei, L. J. and Ying, Z. (1995). Analysis of transformation models with censored data. *Biometrika*, 82, 835-845.

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65, 141-151.

Cox, D. R. (1969). Some sampling problems in technology. *In new developments in survey sampling*. N. L. Johnson and H. Smith, Jr., eds. 506-527. Wiley, New York.

Cox, D. R. (1972). Regression models and life-table (with discussion). *Journal of Royal Statistical Society B*, 33, 187-220.

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

Cox, D. R. and Oakes (1984). Analysis of Survival Data. Chapman and Hall, New York.

Ding, A. A. and Wang, W. (2004). Testing independence for bivariate current status data. *Journal of American Statistical Association*, 99, 145-55.

Dunson, D. and Dinse, G. (2002). Bayesian models for multivariate current status data with informative censoring. *Biometrics*, 58, 79-88.

Farrington, C. P. (2000). Residuals for proportional hazards models with interval-censored survival data. *Biometrics*, 56, 473-482.

Fine, J. P., Ying, Z. and Wei, L. J. (1998). On the linear transformation model for censored data. *Biometrika*, 85, 980-986.

Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics*, 42, 845-854.

Finkelstein, D. M. and Wolfe, R. A. (1985). A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics*, 41, 933-945.

Freireich, E. O. et al. (1963). The effect of 6-mercaptopurine on the duration of steroid induced remission in acute leukemia. *Blood*, 21, 699-716.

Gehan, E. A. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika*, 52, 203-223.

Genest, C. and MacKay, J. (1986). The joy of copulas: Bivariate distributions with uniform marginals. *The American statistician*, 4, 280-283.

Gilbert, P. B., Lele, S. R. and Vardi, Y (1999). Maximum likelihood estimation in semi-parametric selection bias models with application to AIDS vaccine trials. *Biometrika*, 86, 27-43.

Gill, R. D., Vardi, Y. and Wellner, J. A. (1988). Large sample theory of empirical distributions in biased sampling models. *Annals of Statistics*, 16, 1069-1112.

Goggins, W. B. and Finkelstein, D. M. (2000). A proportional hazards model for multivariate interval-censored failure time data. *Biometrics*, 56, 940-943.

Groeneboom, P. (1987). Asymptotics for interval censored observations. *Technical Report.* 87-18, Department of Mathematics, University of Amsterdam.

Groeneboom, P. and Wellner, J. A. (1992). *Information Bounds and Non-Parametric Maximum Likelihood Estimation.* Birkhauser, Boston.

Guo, S. W. and Lin, D. Y. (1994). Regression analysis of multivariate grouped survival data. *Biometrics* 50, 632-639.

Hoel, D. G. and Walburg, H. E. (1972). Statistical analysis of survival experiments. *Journal of National Cancer Institute*, 49, 361-372.

Hougaard, P. (2000). Analysis of Multivariate Survival Data. Springer-Verlag.

Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. *Annals of Statistics*, 24, 540-568.

Huang, J. and Rossini, J. A. (1997). Sieve estimation for the proportional odds model with interval-censoring. *Journal of the American Statistical Association* 92, 960-967.

Huang, X. and Wolfe, R. A. (2002). A frailty model for informative censoring. *Biometrics*, 58, 510-520.

Jewell, N. P., Malani, H. M. and Vittinghoff, E. (1994). Nonparametric estimation for a form of doubly censored data, with application to two problems in AIDS. *Journal of the American Statistical Association*, 89, 7-18.

Jewell, N. P., van der Laan, M and Lei, X. (2005). Bivariate current status data with univariate monitoring times. *Biometrika*, 92, 847-862.

Jin, Z., Lin, D. Y., Wei, L. J. and Ying, Z. (2003). Rank-based inference for the accelerated failure time model. *Biometrika* 90, 341-353.

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

Keiding, N. (1991). Age-specific incidence and prevalence: a statistical perspective. *Journal of Royal Statistical Society B*, 154, 371-412.

Kim, M. Y., De Gruttola, V. and Lagakos, S. W. (1993). Analyzing doubly censored data with covariates, with application to AIDS. *Biometrics* 49, 13-22.

Kong, L., Cai, J. and Sen, P. K. (2004). Weighted estimating equations for semiparametric transformation models with censored data from a case-cohort design. *Biometrika*, 91, 305-319.

Lin, D. Y. and Ying, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika*, 81, 61-71.

Lin, D. Y. and Zeng, D. (2009). Proper analysis of secondary phenotype data in case-control association studies. *Genetic Epidemiology*, 32, 256-265.

Lu, W. and Ying, Z. (2004). On semiparametric transformation cure models. *Biometrika*, 91, 331-343.

Martinussen, T. and Scheike, T. H. (2002). Efficient estimation in additive hazards regression with current status data. *Biometrika*, 89, 649-658.

Murphy, S. A., Rossini, A. J. and Van der Vaart, A. W. (1997). Maximum likelihood estimation in the proportional odds model. *Journal of the American Statistical Association*, 92, 968-976.

Oakes, D. (1982). A model for association in bivariate survival data. *Journal of the Royal Statistical Society Series B*, 44, 414-422.

Oakes, D. (1989). Bivariate survival models induced by frailties. *Journal of the American Statistical Association*, 84, 487-493.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75, 237-249.

Owen, A. B. (1990). Empirical likelihood ratio confidence regions. *Annals of Statistics*, 18, 90-120.

Owen, A. B. (1991). Empirical likelihood for linear models. *Annals of Statistics*, 19, 1725-1747.

Pan, X. R. and Zhou, M. (2002). Empirical likelihood ratio in terms of cumulative hazard function for censored data. *Journal of Multivariate Analysis*, 80, 166-188.

Patil, G. P. and Rao, C. R. (1978). Weighted distributions and size-biased sampling with applications to wildlife populations and human families. *Biometrics*, 34, 179-189.

Pearl, R. (1938). Tobacco smoking and longevity. *Science*, New Series, 87, 216-217.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, 66, 403-411.

Qin, J. (1993). Empirical likelihood in biased sample problems. *Annals of Statistics*, 21, 1182-1196.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *Annals of Statistics*, 22, 300-325.

Rabinowitz, D., Betensky, R. A. and Tsiatis, A. A. (2000). Using conditional logistic regression to fit proportional odds models to interval censored data. *Biometrics*, 56, 511-518.

Rossini, A. and Tsiatis, A. A. (1996). A semiparametric proportional odds regression model for the analysis of current status data. *Journal of the American Statistical*

Association, 91, 713-721.

Struthers, C. A. and Kalbfleisch, J. D. (1986). Misspecified proportional hazard models. *Biometrika*, 73, 363-369.

Sun, J. (1999). A nonparametric test for current status data with unequal censoring. *Journal of the Royal Statistical Society Series B*, 73, 363-369.

Sun, J. (2006). The statistical analysis of interval-censored failure time data. Springer, New York.

Sun, J. and Kalbfleisch, J. D. (1996). Nonparametric tests of tumor prevalence data. *Biometrics*, 52, 726-731.

Sun, J. and Sun, L. (2005). Semiparametric linear transformation models for current status data. *Canadian Journal of Statistics*, 33, 85-96.

Sun, J. and Woodroofe, M. B. (1997). Semi-parametric estimates under biased sampling. *Statistica Sinica*, 7, 545-576.

Thomas, D. R. and Grunkemeier, G. L. (1975). Confidence interval estimation of survival probabilities for censored data. *Journal of the American Statistical Association*, 70, 865-871.

Vardi, Y. (1982). Nonparametric estimation in the presence of length bias. *Annals of Statistics*, 10, 616-620.

Vardi, Y. (1985). Empirical distributions in selection bias models. *Annals of Statistics*, 13, 178-203.

Wang, W. (2003). Estimating the association parameter for copula models under dependent censoring. *Journal of the Royal Statistical Society Series B*, 65, 257-273.

Wang, W. and Ding, A. A. (2000). On assessing the association for bivariate current status data. *Biometrika*, 87, 879-893.

Wang, L., Sun, J. and Tong, X. (2008). Efficient estimation for bivariate current status data. *Lifetime Data Analysis*, 14, 134-153.

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

Yang, S. and Prentice, R. L. (1999). Semiparametric inference in the proportional odds regression model. *Journal of the American Statistical Association*, 94, 125-136.

Younes, N. and Lachin, J. (1997). Link-based models for survival data with interval and continuous time censoring. *Biometrics*, 53, 1817-1828.

Zhang, Z. G., Sun, J. and Sun, L. (2005). Statistical analysis of current status data with informative observation times. *Statistics in Medicine*, 24, 1399-1407.

Zhang, Z. G., Sun, L., Zhao, X. and Sun, J. (2005). Regression analysis of interval-censored failure time data with linear transformation models. *The Canadian Journal of*

Statistics, 33, 61-70.

Zhou, H., Weaver M. A., Qin, J., Longnecker, M. P., Wang, M. C. (2002) A semiparametric empirical likelihood method for data from an outcome-dependent sampling scheme with a continuous outcome. *Biometrics*, 58, 413-421.

Table 1.1: Remission times in weeks for acute leukemia patients

Group	Survival times in weeks
6-MP	$6, 6, 6, 6^*, 7, 9^*, 10, 10^*, 11^*, 13, 16, 17^*, 19^*, 20^*, 22, 23$
	$25^*, 32^*, 32^*, 34^*, 35^*$
Palcebo	$1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23$

Note: the numbers with '*' are censoring times or censored remission times

Table 1.2: Death times in days for 144 male RFM mice with lung tumors

Group	Tumor status	Death times
CE	With tumor	381, 477, 485, 515, 539, 563, 565, 582, 603, 616, 624, 650 651, 656, 659, 672, 679, 698, 702, 709, 723, 731, 775, 779 795, 811, 839
	No tumor	45, 198, 215, 217, 257, 262, 266, 371, 431, 447, 454, 459 475, 479, 484, 500, 502, 503, 505, 508, 516, 531, 541, 553 556, 570, 572, 575, 577, 585, 588, 594, 600, 601, 608, 614 616, 632, 632, 638, 642, 642, 642, 644, 644, 647, 647, 653 659, 660, 662, 663, 667, 667, 673, 673, 677, 689, 693, 718 720, 721, 728, 760, 762, 773, 777, 815, 886
GE	With tumor	546, 609, 692, 692, 710, 752, 773, 781, 782, 789, 808, 810 814, 842, 846, 851, 871, 873, 876, 888, 888, 890, 894, 896 911, 913, 914, 914, 916, 921, 921, 926, 936, 945, 1008
	No tumor	412, 524, 647, 648, 695, 785, 814, 817, 851, 880, 913, 942 986

Table 1.3: Intervals (in months) of cosmetic deterioration (retraction) for early breast cancer patients

Radiotherapy Alone			Radio- and Chemotherapy		
$(45, \infty)$	$(25, 37]$	$(37, \infty)$	$(8, 12]$	$(0, 5]$	$(30, 34]$
$(6, 10]$	$(46, \infty)$	$(0, 5]$	$(0, 22]$	$(5, 8]$	$(13, \infty)$
$(0, 7]$	$(26, 40]$	$(18, \infty)$	$(24, 31]$	$(12, 20]$	$(10, 17]$
$(46, \infty)$	$(46, \infty)$	$(24, \infty)$	$(17, 27]$	$(11, \infty)$	$(8, 21]$
$(46, \infty)$	$(27, 34]$	$(36, \infty)$	$(17, 23]$	$(33, 40]$	$(4, 9]$
$(7, 16]$	$(36, 44]$	$(5, 11]$	$(24, 30]$	$(31, \infty)$	$(11, \infty)$
$(17, \infty)$	$(46, \infty)$	$(19, 35]$	$(16, 24]$	$(13, 39]$	$(14, 19]$
$(7, 14]$	$(36, 48]$	$(17, 25)$	$(13, \infty)$	$(19, 32]$	$(4, 8]$
$(37, 44]$	$(37, \infty)$	$(24, \infty)$	$(11, 13]$	$(34, \infty)$	$(34, \infty)$
$(0, 8]$	$(40, \infty)$	$(32, \infty)$	$(16, 20]$	$(13, \infty)$	$(30, 36]$
$(4, 11]$	$(17, 25]$	$(33, \infty)$	$(18, 25]$	$(16, 24]$	$(18, 24]$
$(15, \infty)$	$(46, \infty)$	$(19, 26]$	$(17, 26]$	$(35, \infty)$	$(16, 60]$
$(11, 15]$	$(11, 18]$	$(37, \infty)$	$(32, \infty)$	$(15, 22]$	$(35, 39]$
$(22, \infty)$	$(38, \infty)$	$(34, \infty)$	$(23, \infty)$	$(11, 17]$	$(21, \infty)$
$(46, \infty)$	$(5, 12]$	$(36, \infty)$	$(44, 48]$	$(22, 32]$	$(11, 20]$
$(46, \infty)$			$(14, 17]$	$(10, 35]$	$(48, \infty)$

Note: A right endpoint ∞ indicates observation is right-censored

Table 2.1: Estimation of θ_0 based on simulated data

$n = 100$									
	$r = 0$			$r = 0.5$			$r = 1$		
θ_0	-1	0	1	-1	0	1	-1	0	1
Bias	0.0549	0.0183	0.0201	0.0858	0.0805	0.0770	0.0902	0.1553	0.1130
SSD	0.2821	0.2436	0.3258	0.3258	0.3050	0.3354	0.3688	0.3694	0.4131
ESE	0.3338	0.2864	0.3311	0.4009	0.3667	0.3934	0.4606	0.4347	0.4634
$n = 200$									
	$r = 0$			$r = 0.5$			$r = 1$		
θ_0	-1	0	1	-1	0	1	-1	0	1
Bias	0.0479	0.0168	0.0273	0.0503	0.0603	0.0574	0.0506	0.0636	0.0761
SSD	0.2030	0.1790	0.2007	0.2327	0.2016	0.2339	0.2519	0.2676	0.2660
ESE	0.2331	0.2002	0.2253	0.2812	0.2562	0.2736	0.3228	0.3037	0.3223

Table 3.1: Bias, ESR, SSD and CP of the estimation of β and α with both C and α independent of X based on 1000 simulation with sample size n=200

Parameter	True value	Bias	ESR	SSD	CP
α	2	0.0690	0.4082	0.4283	0.949
β	0.5	0.0020	0.2659	0.2895	0.943
α	2	0.0625	0.3767	0.4028	0.940
β	0	0.0015	0.2588	0.2585	0.953
α	2	0.0591	0.3579	0.3776	0.960
β	-0.5	-0.0026	0.2577	0.2585	0.967

Table 3.2: Bias, ESR, SSD and CP of the estimation of β and γ with both C and α dependent of X based on 1000 simulation with sample size n=200

Parameter	True value	Bias	ASE	SE	CP (95%)
γ	1	0.0445	0.5679	0.5740	0.955
β	0.5	0.0260	0.3174	0.2751	0.974
γ	1	0. 0302	0.4769	0.4924	0.960
β	0	0.0168	0.3051	0.2516	0.942
γ	1	0.0537	0.4080	0.4882	0.947
β	-0.5	0.0255	0.2883	0.2314	0.949

Table 4.1: Simulation results when $\mu = \mu_1$, $\Sigma = \Sigma_1$, $c = -1$ and $d = 1$

n_1	n_2	n_3	parameter	Bias	ASE	SE	CP (95%)
30	30	100	$\lambda = 0$	0.0013	0.0371	0.0384	0.964
			$\theta_1 = 0.1587$	0.0012	0.0287	0.0293	0.943
			$\theta_2 = 0.1587$	-0.0007	0.0296	0.0292	0.935
50	50	200	$\lambda = 0$	0.0010	0.0290	0.0286	0.944
			$\theta_1 = 0.1587$	-4.2438×10^{-4}	0.0200	0.0208	0.956
			$\theta_2 = 0.1587$	7.7653×10^{-4}	0.0208	0.0207	0.941
100	100	400	$\lambda = 0$	-3.2816×10^{-4}	0.0196	0.0201	0.957
			$\theta_1 = 0.1587$	-1.4522×10^{-4}	0.0152	0.0146	0.948
			$\theta_2 = 0.1587$	-4.9566×10^{-5}	0.0140	0.0147	0.953

† ASE, average of the estimated standard errors; SE, sampling standard errors

Table 4.2: Simulation results when $\mu = \mu_1$, $\Sigma = \Sigma_1$, $c = -1.5$ and $d = 1$

n_1	n_2	n_3	parameter	Bias	ASE	SE	CP (95%)
60	30	100	$\lambda = 0$	-0.0015	0.0320	0.0310	0.960
			$\theta_1 = 0.0668$	-0.0004	0.0221	0.0217	0.932
			$\theta_2 = 0.1587$	-0.0006	0.0284	0.0281	0.937
100	50	200	$\lambda = 0$	0.0005	0.0243	0.0245	0.946
			$\theta_1 = 0.0668$	0.0001	0.0156	0.0157	0.937
			$\theta_2 = 0.1587$	-0.0004	0.0205	0.0209	0.945
200	200	400	$\lambda = 0$	1.2971×10^{-4}	0.0150	0.0148	0.953
			$\theta_1 = 0.0668$	1.7809×10^{-4}	0.0111	0.0108	0.959
			$\theta_2 = 0.1587$	3.8471×10^{-5}	0.0147	0.0152	0.947

Table 4.3: Simulation results when $\mu = \mu_2$, $\Sigma = \Sigma_2$, $c = 2$ and $d = 4$

n_1	n_2	n_3	parameter	Bias	ASE	SE	CP (95%)
30	30	100	$\lambda = 0$	-0.0013	0.0467	0.0478	0.956
			$\theta_1 = 0.1587$	0.0022	0.0321	0.0317	0.935
			$\theta_2 = 0.1587$	3.5244×10^{-4}	0.0314	0.0319	0.936
50	50	200	$\lambda = 0$	0.0014	0.0359	0.0357	0.939
			$\theta_1 = 0.1587$	-7.1979×10^{-4}	0.0221	0.0224	0.945
			$\theta_2 = 0.1587$	0.0013	0.0223	0.0224	0.940
200	200	400	$\lambda = 0$	-5.9395×10^{-4}	0.0245	0.0252	0.960
			$\theta_1 = 0.1587$	-2.7671×10^{-4}	0.0159	0.0158	0.940
			$\theta_2 = 0.1587$	3.1067×10^{-5}	0.0160	0.0159	0.944

Table 4.4: Simulation results for β when $\mu = \mu_1$, $\Sigma = \Sigma_2$, $c = -1$ and $d = 1$

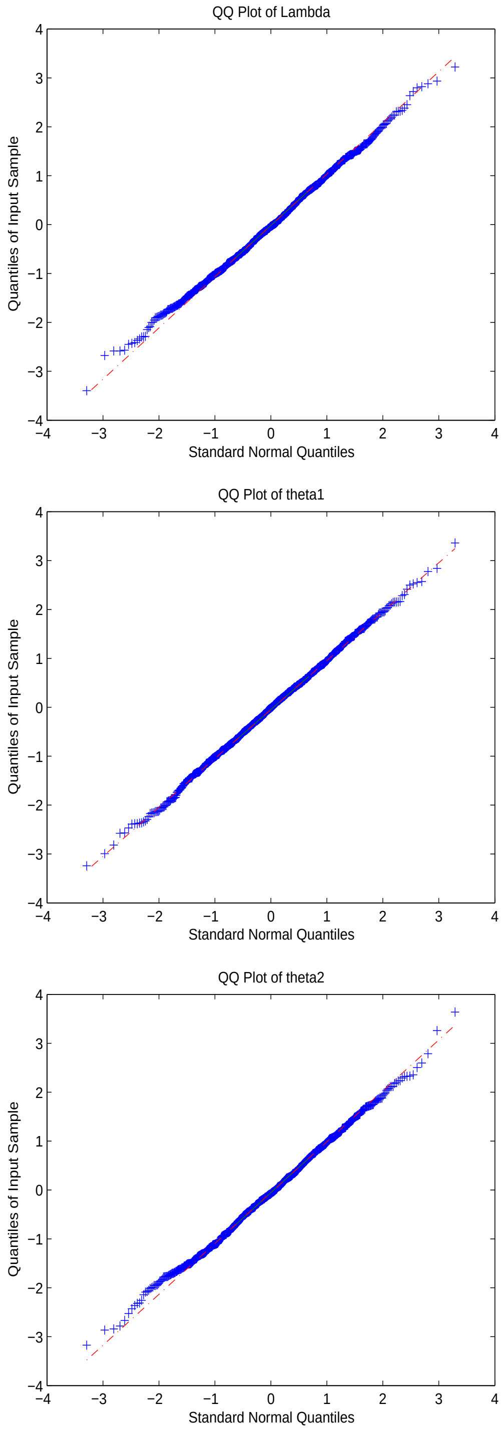
90% CI

n_1	n_2	n_3	Avg Mean	Avg Length	ECP
10	10	30	9.6026×10^{-4}	0.7876	88.7%
30	30	50	-7.4708×10^{-4}	0.6621	89.1%
30	30	100	-0.0073	0.4338	90.3%

95% CI

n_1	n_2	n_3	Avg Mean	Avg Length	ECP
10	10	30	0.0048	0.9302	94.5%
30	30	50	-0.0014	0.7744	93.6%
30	30	100	-0.0012	0.5150	95.2%

Figure 4.1: Q-Q plot for λ and θ

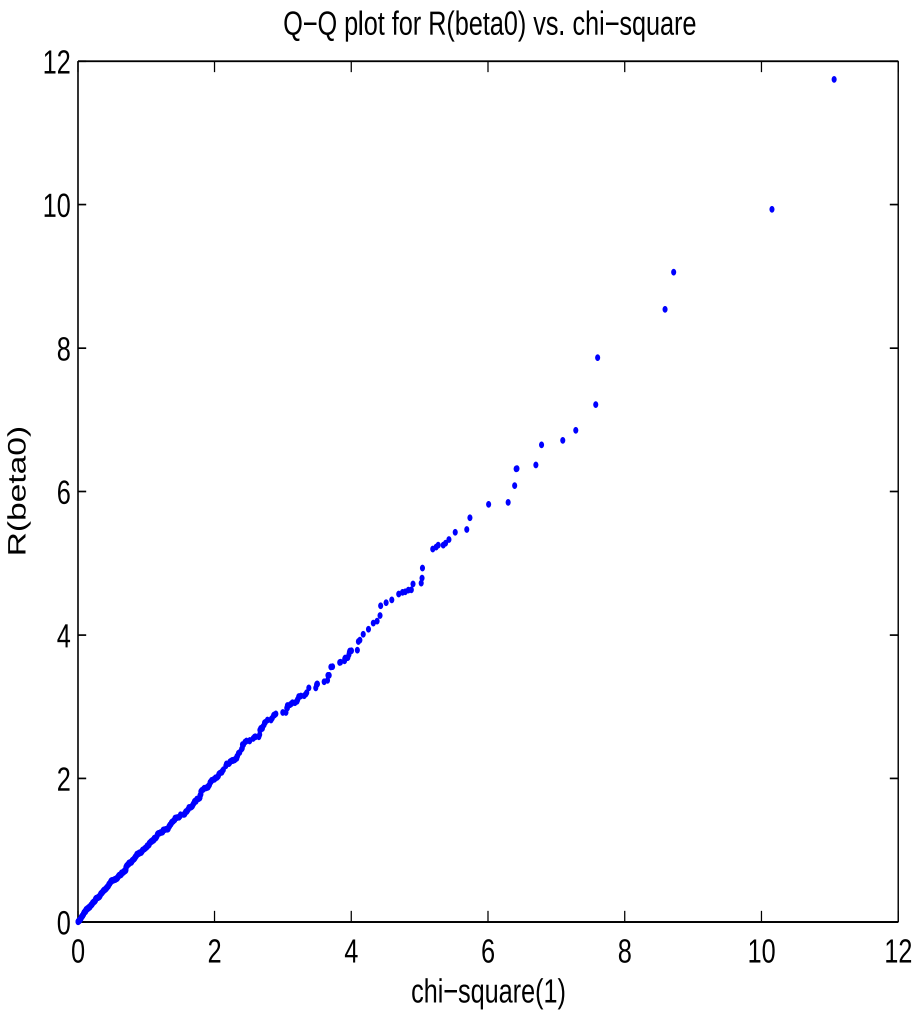


QQ Plot of Lambda

QQ Plot of theta1

QQ Plot of theta2

Figure 4.2: Q-Q plot for $R(\beta_0)$ vs. standard $\chi^2_{(1)}$



VITA

Bin Zhang was born on November 9, 1979 in Baoding, Hebei, China. He received his B. A. in Mathematics from the University of Science and Technology of China (USTC) in 2002. After three years' teaching and researching in Mathematics at USTC, he joined the Department of Statistics at the University of Missouri. He will receive his Ph.D. in Statistics in summer 2009. As of August 2009, he will work as an assistant professor in the Department of Biostatistics at the University of Alabama-Birmingham.