STATISTICAL ANALYSIS OF MULTIVARIATE INTERVAL-CENSORED FAILURE TIME DATA

A Dissertation Presented

to

the Faculty of the Graduate School

University of Missouri-Columbia

In Partial Fulfillment

Of the Requirements for the Degree

Doctor of Philosophy

---

by

LIANMING WANG

Dr. (Tony) Jianguo Sun, Supervisor

AUGUST 2006

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

Statistical Analysis of Multivariate Interval-censored Failure Time Data

presented by Lianming Wang

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. Farroll T. Wright _____

Dr. Athanasios C. Micheas _____

Dr. Allanus Tsoi _____

Dedicated to my parents,

Wang, Chongfa and Liu, Yongfang

ACKNOWLEDGEMENTS

I would like to express my deep and sincere gratitude to my advisor Dr. Tony Sun. His wonderful guidance, generous support, and endless patience enable me to complete this work. His encouragement and instruction lead me to a good start of my research life. I extend my gratitude to the members of my advisory committee: Dr. Nancy Flournoy, Dr. Tim Wright, Dr. Athanasios Micheas and Dr. Allanus Tsoi for their insightful comments and suggestions on my work. Special thanks are due to Liuquan Sun and Xingwei Tong for their helpful academic discussion and help.

I want to thank our department for offering me such a wonderful opportunity studying here. Thank our fantastic faculty for providing many great courses. Thank our great staff Judy and Tracy for endless help. I also want to thank my friends in this department. We have had so much fun together during my graduate life here.

I am deeply indebted to my parents and my brothers for their unselfish love and support throughout my life. Foremost, I am grateful to my beloved wife, Xiaoyan Lin, for making my life so wonderful.

Table of Contents

List of Tables

List of Figures

Statistical Analysis of Multivariate Interval-censored Failure Time Data

Lianming Wang

Dr. (Tony) Jianguo Sun, Dissertation Supervisor

ABSTRACT

Interval-censored failure time data commonly arise in clinical trials and medical studies. In such studies, the failure time of interest is often not exactly observed, but known to fall within some interval. For multivariate interval-censored data, each subject may experience multiple events, each of which are interval-censored. Research interests on such data in this dissertation focus on regression analysis and studying the statistical association between these events.

In Chapter 1, three real-life studies are discussed to illustrate interval-censored data. We review the research in the literature with the focus on nonparametric maximum likelihood estimation, regression analysis of univariate interval-censored data, and estimation of the association parameter for multivariate data.

Chapter 2 considers regression analysis of correlated interval-censored failure time data. We first review the marginal Cox model approach for regression analysis of multivariate interval-censored data, and then construct a goodness-of-fit test based on a discrete score process. Simulation results show that the proposed test works well for finite sample sizes. The proposed method is illustrated using a set of real data from an AIDS study.

Chapter 3 considers estimation of the association of bivariate interval-censored failure time data. We assume that the joint survival function of the two variables follows a Copula model. A two-stage estimation procedure is proposed to estimate the association parameter and the asymptotic properties of the proposed estimator are established. Simulations are conducted to assess the finite sample properties of the proposed approach. The same AIDS

data used in Chapter 2 are analyzed again with the proposed method.

Chapter 4 discusses regression analysis of case-2 interval-censored data using the additive hazards model. In this chapter, we propose an easy procedure to estimate the regression parameter without estimating the baseline hazard function or survival function. We construct two types of counting processes and martingales and develop some estimating equations for the regression parameters. The large sample properties of the proposed estimators are proved. Simulations results suggest that the proposed approach works well for finite sample sizes. Data from a breast cancer study are analyzed to illustrate the proposed method.

In Chapter 5, we study the efficient estimation of regression parameters and association parameter simultaneously for bivariate current status data with univariate censoring. Our proposed method applies to the situations where the two marginal distributions may be different and where the censoring variable may be dependent of covariates. In the estimation procedure, the sieve method is applied to approximate the infinite dimensional baseline hazard function. A set of tumor data from National Toxicology Program is used to illustrate our method.

Some future research directions are addressed in Chapter 6.

Chapter 1

Introduction

In survival analysis, failure time can be broadly defined as the time to the occurrence of a certain event. Examples of failure times include the lifetimes of machine components, duration of strokes, time to HIV infections, and duration of the first marriage. One important feature of survival data is that each subject can only have one failure for the defined event.

Another feature of survival data is incomplete observation of failure time due to various reasons including dropouts of study or limited follow-ups. For example, it is common that not all patients survive at the end of a clinic study, not all people get divorced in a sociology study, and nonlethal tumors can not be observed until the rats die or are sacrificed in a cross-sectional study. Such incomplete observation of failure time is called censoring (Cox and Oakes 1984). A censoring time is also a event time and is always observable. A censoring time for a subject can be an examination (observation) time, dropout time for the subject, or the time when the study ends. In some studies, there may be more than one censoring times for each subject.

There are various types of censoring mechanisms depending on the relationship of the failure time and the censoring time(s), including left-censoring, right-censoring, and interval-censoring. Suppose each subject is observed or examined multiple times in the study. Left censoring occurs when a subject has already experienced the failure at the first censoring time. Right censoring occurs when a subject has not experienced the failure at the last censoring time. Interval-censoring occurs when the failure time happens between two

adjacent observation times.

Let T_i denote the failure time of interest for subject i in the study. Interval-censoring means that T_i is not observed exactly but only known to fall within some interval $(L_i, R_i]$, where L_i and R_i can be regarded as two censoring times for subject i. The observed interval $(L_i, R_i]$ for subject i has one of the follow forms: $(0, u]$, $(u, v]$, and $(v, +\infty)$, where $v > u > 0$, corresponding to left-censoring, interval-censoring, and right-censoring. Thus, interval-censored data have a mixed data structure. If $L_i = 0$ for all i, we have left-censored data; if $R_i = +\infty$ for all i, we have right-censored data; if either $L_i = 0$ or $R_i = +\infty$ for all i, we have current status data, which is also called case 1 interval-censored data. The latter type of data usually arises when each subject is observed only once in the study. The general interval-censored data are often called case 2 interval-censored data.

In most research, people assume that failure time and censoring time(s) are independent, which is referred to as noninformative censoring. Without such assumptions, it is hard to make inference. We assume noninformative censoring throughout this thesis unless it is mentioned otherwise.

This chapter is organized as follows. Section 1.1 describes three examples of interval-censored data that motivated this research. Section 1.2 reviews nonparametric maximum likelihood estimation for interval-censored data. In Section 1.3, we discuss regression analysis of interval-censored data with commonly used semiparametric regression models. In Section 1.4, we discuss two main research interests and existing approaches for the analysis of multivariate interval-censored data. Finally the outline of this thesis is given in Section 1.5.

1.1 Examples of Interval-censored Data

Interval-censored data commonly arise from clinic trials and medical studies. In such studies, subjects are usually not under continuous observation, and they undergo periodical

examinations instead. In consequence, the failure time can not be observed exactly but only known to be within some intervals. In this section, we describe three examples of interval-censored data that motivated this research.

1.1.1 Breast cancer data

Breast cancer data contain 94 early breast cancer patients in two treatment groups, radiotherapy alone and radiation therapy together with adjuvant chemotherapy. Among the patients, 46 received radiotherapy and 48 received combined therapy aforementioned. In this study, patients were examined periodically and actual examination times differ from patient to patient since some of them missed their visits. One objective of this study is to detect whether chemotherapy changes the rate of deteriorations of the cosmetic state. To do this, we can compare the treatments with respect to time until the appearance of breast retraction, a response that has a negative impact on overall cosmesis appearance. Treating the appearance of breast retraction as failure of interest, we have only interval-censored data, which are presented in Table 1.1. References that discussed this data set include Finkelstein et al. (2002) and Goggins and Finkelstein (2000).

1.1.2 Hemophilia data

A multi-center prospective study was conducted in 1980's to investigate HIV-1 infection rate among people with hemophilia (Kroner et al., 1994). The patients were at risk of HIV-1 infection from blood products such as factor VIII and factor IX made from donors' plasma. In this study, only case 2 interval-censored data were observed for patients' HIV-1 infection times. The patients were categorized into one of four groups according to the average annual dose of the blood products they received: high-, median-, low-, or none-dose group. The goal of this study is to compare the HIV-1 infection rates between treatment groups. More details about this study can be found in Kroner et al. (1994) and Goedert et al. (1989).

1.1.3 ACTG 181 data

The ACTG 181 data come from an AIDS observational study conducted by the AIDS Clinical Trials Group (ACTG). In this study, patients were scheduled to provide blood and urine samples at clinic visits every 12 weeks and every 4 weeks, respectively. Urine samples and blood samples were tested in order to detect the presence of the cytomeglovirus (CMV) virus. Since CMV shedding is not accompanied by any symptoms and only detectable in the laboratory test, it can not be observed immediately when it exists and only possibly observed at scheduled clinic visits. In the study, not all subjects made each clinic visit, and many people visited sometime after the scheduled time. Some patients missed several visits and returned with changed CMV shedding status, which resulted in interval-censored data. Some patients were already shedding when they entered the study, which gave left-censored data. Some patients had not started shedding by the time the study had ended, which resulted in right-censored data. The CD4 count was recorded at the entry time. There are 204 subjects in the study. For the blood shedding, 7 subjects are left-censored , 23 interval-censored, and 174 right-censored. For the urine shedding, 49 subjects are left-censored subjects, 67 interval-censored, and 88 right-censored. More details about this study can be found in Finkelstein et al. (2002) and Goggins and Finkelstein (2000).

In this study, the investigators were interested in determining whether the stage of HIV disease at study entry was predictive of an increased hazard for CMV shedding in either blood or urine. The stage of HIV is categorized by a CD4 count below a certain threshold. Since both blood shedding and urine shedding contribute to CMV shedding, we need to model blood shedding and urine shedding jointly and consider the bivariate regression in term of the HIV status categorized by CD4 count. Goggins and Finkelstein (2000) studied this problem by using the marginal Cox model approach.

Another question of interest about this study is to consider the correlation between blood shedding and urine shedding. We will study this problem in terms of the association

parameter in Chapter 3.

1.1.4 NPT tumor data

This data set comes from a part of an animal tumorigenicity experiment conducted by National Toxicology Program (NTP). It is a 2-year rodent carcinogenicity study of chloroprene, in which subjects were F344/N rats and B6C3F$_1$ mice with both sexes. The experiment was described and summarized in Dunson and Dinse (2002). The experiment contained a control group with no chloroprene and three dose groups with 50 rodents in each group. Rodents in the dose groups were exposed to chloroprene at the concentration of 12.8, 32, and 80 ppm, respectively, 6 hours per day, 5 days per week for up to 2 years. The occurrence of tumors was determined through a pathologic examination when the rodents died. Some rodents died during the study. Those rodents who did not die at the end of the 2-year study were sacrificed regardless of health condition. Dunson and Dinse (2002) summarized the data for male rats in the control group and 80mmp dose group concentrating on adrenal and lung tumor only. Table 1.2 shows the summarized data in month by Dunson and Dinse (2002). Since each rat was examined for tumor at the death time, the only information available is whether the rat had suffered adrenal tumor or lung tumor at that time, that is, the onset time of adrenal tumor and lung tumor were either left-censored or right-censored by the death time. Thus, this is bivariate current status data with univariate censoring. In Chapter 5, we study the dose effects and association of the two types of tumors simultaneously.

1.2 Nonparametric Maximum Likelihood Estimation

In the nonparametric setting, the primary interest is to estimate the survival function or cumulative distribution function (CDF) F of the failure time event T, that is, to find nonparametric maximum likelihood estimate (NPMLE) given observed data.

1.2.1 Case 1 interval-censored or current status data

Let $\{(X_i, \delta_i), i = 1, \cdots, n\}$ denote the observed data, where X_i is the censoring time for subject i, and δ_i indicates by 1 if subject i has already experienced the failure at X_i. Let $X_{(i)}$ be the i^{th} order statistic of $(X_1, \cdots, X_n)$ and let $\delta_{(i)}$ denote the corresponding indicator, i.e., if $X_{(i)} = X_j$, then $\delta_{(i)} = \delta_j$. Then the NPMLE $\hat{F}$ is the maximizer of the following log-likelihood function

$$l(\tilde{\boldsymbol{x}}) = \sum_{i=1}^{n} \{\delta_{(i)} \log(x_i) + (1 - \delta_{(i)}) \log(1 - x_i)\}$$

under the condition $0 \leq x_1 \leq \cdots \leq x_n \leq 1$. The NPMLE $\hat{F}$ can be obtained using either the self-consistency algorithm or the greatest convex minorant algorithm described in Groeneboom and Wellner (1992) and Robertson et al. (1988). $\hat{F}$ can also be represented by the max-min formula:

$$\hat{F}(X_{(i)}) = \max_{l \leq i} \min_{k \geq i} \frac{\sum_{j=l}^{k} \delta_{(j)}}{k - l + 1}.$$

The distribution of $\hat{F}$ was derived by Groeneboom and Wellner (1992) and $\hat{F}$ was shown to have $n^{1/3}$- convergence rate instead of $n^{1/2}$. However, for the smooth functional $\mu(F)$ like mean $\mu(F) = \int t dF(t)$, the NPMLE $\mu(\hat{F})$ converges to a normal distribution with $n^{1/2}$ convergence rate under some extra conditions (Groeneboom and Wellner, 1992).

1.2.2 Case 2 interval-censored data

Suppose $(L_i, R_i]$ is the observed interval for subject i to contain T_i. Let $\{s_j, j = 0, \cdots, m + 1\}$ denote the distinct elements of $\{0, \{L_i, R_i\}_{i=1}^{n}, \infty\}$ in ascending order. Let α_{ij} be the indicator of the event $s_j \, \epsilon \, (L_i, R_i]$ and $p_j = F(s_j) - F(s_{j-1})$, where $j = 1, \ldots, m+1$. Under this setting, to find the NPMLE of F is equivalent to maximizing the following likelihood $L(p)$ with respect to p with constraints $\sum_{j=1}^{m+1} p_j = 1$, and $p_j \geq 0$, $j = 1, \cdots, m +$

1 :

$$L(p) = \prod_{i=1}^{n} \{F(R_i) - F(L_i)\} = \prod_{i=1}^{n} (\sum_{j=1}^{m+1} \alpha_{ij} p_j).$$

For case 2 interval-censored data, there is no close form for the NPMLE . Turnbull's estimator is commonly used in this situation and can be obtained through a self-consistency algorithm (Turnbull,1976). Turnbull's estimator is easy to implement, but has a slow convergence rate. Another problem to note is that a self-consistent estimator may not be the NPMLE. Gentleman and Geyer (1994) showed that the Kuhn-Tucker conditions are necessary and sufficient conditions for a self-consistent estimator to be the NPMLE by using standard convex optimization techniques. Another method to obtain the NPMLE is to apply the convex minorant algorithm introduced by Groeneboom and Wellner (1992), which converges faster than the self-consistency algorithm. Since case 2 interval-censored data contains more information about failure time than does case 1 data, one may expect that the NPMLE has a faster convergence rate than $n^{1/3}$ for case 2 interval-censored data. However, the asymptotic distribution of the NPMLE is not established yet for case 2 interval-censored data although its consistency is already known. It is conjectured that the NPMLE has a convergence $(n \log n)^{1/3}$ when the two censoring times are bounded away (Groeneboom and Wellner 1992; Geskus and Groenboom 1999).

1.3 Regression Models

In regression analysis, the prime interest is to estimate covariate effects such as treatment, age, sex, income on the failure time. The baseline survival or hazard function is of secondary interest or is treated as a nuisance parameter with infinite dimension. In the following, we use Z to denote covariates and describe several regression models commonly used in survival analysis.

1.3.1 The proportional hazards model

The proportional hazards model, also termed as Cox model since it was first proposed by Cox (1972), specifies that covariates have a multiplicative effect on the hazard function, i.e.,

$$\lambda(t) = \lambda_0(t) \exp(\beta' Z),$$

where $\lambda_0(t)$ denotes the unknown baseline hazard function.

The proportional hazards model has been widely used in survival analysis mostly due to the existence of the partial likelihood function for right-censored data under this model (Cox, 1975). Based on the partial likelihood function, one can estimate the regression parameters without estimating the unspecified baseline hazard function. Anderson and Gill (1982) gave a simple and elegant proof for the asymptotic properties of regression parameter estimators using counting process and martingale theory.

Many methods have been proposed to analyze interval-censored data by using proportional hazards model, including full likelihood approach and marginal likelihood approach.

Full likelihood approach requires maximization of the full likelihood over the regression parameters and the baseline hazard or survival function simultaneously. Finkelstein (1986) first studied this approach for discrete failure time. Huang (1996) also studied this approach for current status data and showed the MLE of regression coefficients is consistent and efficient and has asymptotic normal distribution with $n^{1/2}$-convergence rate. Huang and Wellner (1996) proved the same results for case 2 interval-censored data under some conditions.

One way to avoid estimating the baseline hazard function is to use the marginal likelihood approach. This approach defines a marginal likelihood as the summation of the probabilities of the rank of the $T_i's$ that are consistent with the observed interval-censored data (Satten, 1996). However, this approach needs great computation effort because it does not have a simple and manageable form. Moreover, little is known about the asymptotic properties of the estimators obtained. As an alternative, Satten et al. (1998) and Pan (2000)

proposed imputation methods, in which right-censored data were generated and imputed based on the observed interval-censored data.

Many other people studied the proportional hazards model for interval-censored data. Among them, Kooperberg and Stone (1992) and Rosenberg (1995) gave spline-based estimators. More recently, Bebchuk and Betensky (2000) proposed multiple imputation approach, Betensky et al. (2002) explored a local likelihood method, and Cai and Betensky (2003) considered piecewise linear penalized spline.

1.3.2 The additive hazards model

The additive hazards model specifies the hazard function as a sum of the baseline hazard function and a regression function of covariates, i.e.,

$$\lambda(t|Z) = \lambda_0(t) + \beta' Z,$$

where $\lambda_0(t)$ is a completely unspecified baseline hazard function.

The additive hazards model has drawn attention since Lin and Ying (1994) proposed an easy-implemented procedure for right-censored data, in which the regression coefficient estimator has a close form.

There are a few papers discussing the additive hazards model for current status data. Lin et al. (1998) proposed an easy procedure to estimate the regression parameters without estimating any nuisance parameters, Martinussen and Scheike (2002) explored efficient estimation under the same setting, and Zhang et al. (2005) studied informative censoring using the additive hazards model.

Both Lin et al. (1998) and Zhang et al. (2005) assumed that the failure time follows an additive model and the censoring time follows a proportional hazards model. The advantage of this setting is that there exists a counting process based on the failure time and the censoring time that has an intensity function of multiplicative form, thus partial likelihood

score function can be used directly as with right-censored data.

We generalize this setting to case 2 interval-censored data by assuming that both of the censoring times follow the proportional hazards model in Chapter 4. In our proposed method, there is no need estimating any nuisance parameters and asymptotic properties are established.

1.3.3 The accelerated failure time model

The accelerated failure time model is also widely used in survival analysis. This model relates the covariates linearly to the logarithm of the survival time T with the following form:

$$\log(T) = \beta' Z + \epsilon,$$

where the distribution of ϵ is completely unknown.

There are many papers in literature analyzing right-censored data with the accelerated failure time model, but not many for interval-censored data. For interval-censored data, Rabinowitz et al. (1995) proposed a class of score statistics and Betensky et al. (2001) proposed an estimating equation by treating the examination times from the same subject as if they were from different subjects. Both methods need great computation effort in the estimation procedure since estimation of the distribution of ϵ is needed. Huang and Wellner (1996) discussed this model for both current status data and case 2 interval-censored data. More recently, Tian and Cai (2004) proposed a MCMC based approach with accelerated failure time model in a technique report.

1.3.4 The proportional odds model

An important alternative to the proportional hazards model is the proportional odds model, which assumes that

$$\log\{F(t|Z)/S(t|Z)\} = \lambda_0(t) + \beta' Z$$

where $F(t|Z)$ and $S(t|Z)$ are the distribution and survival functions of failure time T given covariate Z, respectively, and $\lambda_0(t)$ is the baseline function referred to as the baseline log odds.

Rossini and Tsiatis (1996) studied the current status data with the proportional odds model by approximating the baseline log-odds function with a step function. For interval-censored data, Huang and Rossini (1997) studied sieve estimation and showed the estimate of regression parameter converges at $n^{1/2}$ rate, and Rabinowitz et al. (2000) considered conditional logistic regression. For the asymptotic properties of regression coefficient estimators, see the review paper by Huang and Wellner (1996). Zhang, et al. (2005) considered a class of linear transformation models that included the proportional odds model as a special case and proposed a method that does not need to estimate the baseline log odds $\lambda_0(t)$.

1.4 Statistical Analysis of Multivariate Interval-censored Data

Multivariate failure time data commonly occur in medical studies, in which there are at least two failure time events of interest and these events are correlated. Multivariate interval-censored failure time data can occur when the outcomes are not directly observable but are detected from periodic clinical examination or laboratory tests, the occurrences of bacterial and viral infections.

The ACTG 181 study mentioned above provides an example of bivariate interval-censored data, in which blood shedding and urine shedding are both of interest. In this example, two questions are usually of interest: (1) Do covariates have significant effects on occurrences of the blood or urine shedding? (2) Are the two events correlated? If so, in what degree are these two failure time events correlated? To answer the first question, we need to perform regression analysis, while for the second question, we need to test or measure the association of dependence in term of parameter such as Pearson's correlation ρ and Kendall's τ, among others.

1.4.1 Regression analysis

Two approaches are commonly used in modelling multivariate data: marginal approach and random effect models. In the literature, both approaches have been applied to the proportional hazards model. In the following, we will discuss these two approaches in terms of the proportional hazards model.

Marginal approach

The marginal proportional hazards model approach assumes that each of the correlated failure times follows a proportional hazards model. It deals with the marginal distributions directly and ignores the dependence structure between the failure times completely. For regression analysis, marginal approach focuses on the robust estimation of regression coefficients.

Marginal proportional hazards model approach has been widely used for regression analysis of correlated failure time data. For example, Wei et al. (1989) first discussed this method for correlated right-censored failure time and proposed the partial likelihood estimators for regression parameters by using a working independence assumption. Cai and Prentice (1995) investigated the same idea and developed some estimating equation methods for inference about regression parameters. More recently, Goggins and Finkelstein (2000) and Kim and Xue (2002) considered the maximum likelihood approach for regression analysis of multivariate interval-censored failure time data. Bogaerts et al. (2002) applied generalized estimating equation approach with accelerated failure time model to multivariate interval-censored data.

Random effect model approach

Random effect model approach assumes conditional independence between different failure times by introducing a common latent random variable to the marginal hazard functions. The latent random variable is also called frailty, which reflects the dependence of

failure times. Thus, we can study the dependence through the estimation of the frailty.

Random effect models have been widely applied to analyze multivariate data. Among others, Yue and Chan (1997) proposed a dynamic frailty model with gamma frailty for serially correlated right-censored data. Huang and Wolfe (2002) studied informative censoring for clustered right-censored data by frailty models.

1.4.2 Estimation of association parameter

When the estimation of association is of prime interest, we focus on estimating association parameter and treat marginal distributions as nuisance functions. Thus, parametric, semiparametric, and even nonparametric method can be used to estimate the marginal distributions if needed.

The so-called two-stage procedure is as follows: first, estimate the marginal distributions or survival functions; secondly, estimate the association parameter by maximizing the pseudo-likelihood obtained by plugging in the estimated marginal survival functions. Shih and Louis (1995) first studied this two-stage procedure to estimate the association parameter for right-censored data. Wang and Ding (2000) consider the same procedure for current status data. In Chapter 4, we generalize this procedure to interval-censored data and asymptotic properties of estimated parameters are established.

There are many other methods studying the dependence of bivariate survival data. Among others, Shih and Louis (1996) proposed two test statistics based on martingale residuals for testing the independence for bivariate right-censored data. Hsu and Prentice (1996) studied the dependency of bivariate right-censord failure times with the correlation coefficient between cumulative hazard functions. Betensky and Finkelstein (1999) studied an extension of Kendall's coefficient of concordance and applied it to interval-censored data.

In Chapter 5, we estimate the association parameter for bivariate current status data under univariate censoring when covariates exist. We proposed to estimate regression coefficient and association parameter simultaneously and the estimates are efficient.

1.5 Outline

The reminder of this thesis is organized as follows. In Chapter 2, we propose a goodness-of-fit test for marginal Cox model for correlated interval-censored failure time data. We first review the marginal Cox model approach for interval-censored data, and then construct a goodness-of-fit test based on a discrete score process. Simulation results show that the proposed test works well for finite sample sizes. The proposed method is illustrated by ACTG 181 data mentioned before.

Chapter 3 considers estimation of the association of bivariate interval-censored failure time data. We assume that the joint survival function of the two variables follows a Copula model. A two-stage estimation procedure is proposed to estimate the association parameter and the asymptotic properties of the proposed estimator are established. Simulations are conducted to assess the finite sample properties of the proposed approach. The ACTG 181 data are analyzed with the proposed method.

Chapter 4 discusses regression analysis of interval-censored data using the additive hazards model. In this chapter, we develop some estimating equations for regression parameters without involving any baseline hazard functions and the large sample properties of the proposed estimators are established. Simulations results suggest that the proposed approach works well for finite sample size. Breast cancer data are analyzed to illustrate the proposed method.

In Chapter 5, we study the efficient estimation of regression parameters and association parameter simultaneously for bivariate current status data with univariate censoring. In the estimation procedure, we use the sieve method to approximate the infinite dimensional baseline function by a step function with finite number of jump points. Full likelihood method is used for efficiency comparison. The tumor data conducted by National Toxicology Program are analyzed to illustrate our method.

In Chapter 6, several directions for future research are addressed.

Chapter 2

A Goodness-of-fit Test for the Marginal Cox Model for Correlated Interval-censored Failure Time Data

2.1 Introduction

This chapter discusses regression analysis of correlated interval-censored failure time data with the most commonly used approach in this situation: marginal proportional hazards model approach. We propose a goodness-of-fit test to assess model adequacy for this approach.

As the most commonly used model for multivariate regression analysis, marginal Cox model has been applied to interval-censored data. Among others, Goggins and Finkelstein (2000) and Kim and Xue (2002) considered the maximum likelihood approach for regression analysis of correlated interval-censored failure time data.

It is well-known that the assessment of the adequacy of an assumed model is often critical for the validity of statistical inference. To check the appropriateness of the Cox model, a number of methods have been proposed for univariate and correlated right-censored failure time data. For example, Wei (1984) discussed the testing of the two-sample continuous Cox model for univariate right-censored failure time data and proposed to base the test on the score process given by the partial likelihood. Klein and Moeschberger (2003) gave relatively complete discussion about commonly used goodness-of-fit test procedures for univariate right-censored data. Spiekerman and Lin (1996) discussed the assessment of the

marginal Cox model for correlated right-censored data and developed a class of numerical and graphical procedures using martingale-based residuals. However, it seems that there is no approach available for correlated interval-censored failure time data. Note that because of the difference between censoring structures, the development of both inference and model-checking procedures for correlated interval-censored data is much more difficult than for correlated right-censored data.

The remainder of this chapter is organized as follows. We begin in Section 2.2 with introducing the notation and then discuss the marginal Cox model approach for correlated interval-censored failure time data. Section 2.3 considers the goodness-of-fit test of the marginal Cox model and a test procedure is presented. The proposed procedure consists of constructing a score process (Wei, 1984) and approximating the null distribution of the test statistic through simulation. Section 2.4 presents some numerical results obtained from a simulation study and by applying the proposed methodology to the set of correlated AIDS interval-censored data discussed in Chapter 1. Some concluding remarks are given in Section 2.5. In the following, as Goggins and Finkelstein (2000) and Kim and Xue (2002), we will focus on the discrete marginal Cox model.

2.2 The Marginal Cox Model

Consider a survival study that consists of N independent subjects and in which each subject experiences K correlated events or types of failures. Let T_k denote the time to the kth event and Z a vector of covariates that may affect T_k, $k = 1, ..., K$. Also let $0 = s_0 < s_1 < ... < s_m < s_{m+1} = \infty$ denote all possible values that the T_k's take. Sometimes it may be more reasonable to assume that the s_j's are all possible time points at which subjects are observed. This is the situation for most of interval-censored data arising from periodic follow-up studies and in this case, the s_j's could be days, months or years, while T_k could be a continuous or discrete variable.

Suppose that the failure times T_k's follow the discrete marginal Cox model given by

$$S_{jk}(Z) = Pr\{T_k > s_j|Z\} = (\lambda_1 \cdots \lambda_j)^{\exp(\beta' Z)} \tag{2.1}$$

for the marginal survival function of T_k given Z (Wei et al., 1989; Goggins and Finkelstein, 2000). In the above, β is the p-dimensional vector of regression parameters and $\lambda_j = Pr\{T_k > s_j|T_k > s_{j-1}, Z = 0\}$, $k = 1, ..., K$, $j = 1, ..., m$. Note that model (2.1) assumes that the K events share the same baseline survival function. An alternative is to assume that different types of failures have different baseline survival functions. As pointed out in Spiekerman and Lin (1996), from the model-checking point of view, the development of a goodness-of-fit procedure for model (2.1) is more delicate and interesting than for the alternative model and the resultant methodologies are similar for both models. Thus we will focus on model (2.1) in the following.

Suppose that one observes only interval-censored failure time data given by $\{(A_{ik}, Z_i)$, $i = 1, ..., N$, $k = 1, ..., K\}$, where $A_{ik} = [L_{ik}, R_{ik}]$ is the interval within which the kth failure of the ith subject is observed to occur and Z_i denotes the vector of covariates associated with subject i. Assume that $\{L_{ik}\,, R_{ik}\} \subseteq \{s_j\}_{j=0}^{m+1}$ and define $\alpha_{ijk} = 1$ if A_{ik} contains s_j and $\alpha_{ijk} = 0$ otherwise, $j = 1, ..., m+1$, $k = 1, ..., K$, $i = 1, ..., N$. Let $\gamma_j = \log(-\log \lambda_j)$, $\gamma' = (\gamma_1, ..., \gamma_m)$ and $\theta' = (\beta', \gamma')$. Then the log-likelihood contribution from the kth type of failure of the ith subject is given by

$$l_{ik}(\theta) = \log \sum_{j=1}^{m+1} \alpha_{jik} \left([1 - \exp\{-\exp(\beta' Z_i + \gamma_j)\}] \prod_{l=1}^{j-1} \exp\{-\exp(\beta' Z_i + \gamma_l)\} \right) , \tag{2.2}$$

where $\gamma_{m+1} = \infty$. Note that the reparameterization by the γ_j's removes the range restriction on the λ_j's.

For inference about θ, both Goggins and Finkelstein (2000) and Kim and Xue (2002)

suggested to base the log-likelihood function

$$l(\theta) = \sum_{i=1}^{N} \sum_{k=1}^{K} l_{ik}(\theta)$$

obtained under the working independence assumption that the K types of failures are independent even it is not true in reality. The point estimate of θ is obtained by maximizing this log-likelihood and the covariance matrix is adjusted by taking into account the correlation later. The same idea was used by, among others, Wei et al. (1989) for the analysis of multivariate right-censored data. Let $\hat{\theta}$ denote the solution to $\partial l(\theta)/\partial\theta = 0$ and $I(\theta) = -N^{-1}\partial^2 l(\theta)/\partial\theta\partial\theta'$. Then it can be shown (Kim and Xue, 2002) that $N^{1/2}(\hat{\theta} - \theta)$ has an asymptotic normal distribution with mean zero and covariance matrix that can be consistently estimated by $I^{-1}(\hat{\theta})D(\hat{\theta})I^{-1}(\hat{\theta})$, where

$$D(\theta) = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \sum_{l=1}^{K} \frac{\partial l_{ik}(\theta)}{\partial\theta} \frac{\partial l_{il}(\theta)}{\partial\theta'} \ . \tag{2.3}$$

In the next section, we will consider the assessment of the marginal model (2.1).

2.3 A Goodness-of-fit Test

To develop a test procedure for checking the adequacy of model (2.1), following the idea discussed in Wei (1984), we will construct a score process. For this purpose, we need to define the observed data if all subjects are observed only up to time s_r, $r = 1, ..., m$. One natural way is to define the observed data up to time s_r by $\{(A_{ik}^{(r)}, Z_i), i = 1, ..., N, k = 1, ..., K\}$, where

$$A_{ik}^{(r)} = \begin{cases} [L_{ik}, R_{ik}] & \text{if } R_{ik} \le s_r, \\ [L_{ik}, \infty) & \text{if } L_{ik} \le s_r < R_{ik}, \\ [s_r, \infty) & \text{if } L_{ik} > s_r \ , \end{cases} \tag{2.4}$$

which is a simple generalization of the definition that one would use for continuous right-censored data.

The above definition, however, does not seem to make use of the discrete and finite nature of the problem. Corresponding to this, one could alternatively define the observed data up to s_r as $\{(A_{ik}^{(r)}, Z_i), i = 1, ..., N, k = 1, ..., K\}$, where for $R_{ik} \leq s_r$, $A_{ik}^{(r)}$ is the same as in (2.4) and for $L_{ik} \leq s_r < R_{ik}$, $A_{ik}^{(r)} = [L_{ik}, s_r]$ or $[s_r, \tau)$ determined by a random number from the Bernoulli distribution with the probabilities proportional to the lengths of the two intervals, where τ denotes the largest follow-up time. This alternative definition is used in the following numerical studies and more comments on this are given in Section 2.5.

Let $l_{ik}^{(r)}(\theta)$ denote the log-likelihood contribution given in (2.2) with replacing A_{ik} by $A_{ik}^{(r)}$ and define

$$S_r(\theta) \;=\; N^{-1/2} \sum_{i=1}^{N} \sum_{k=1}^{K} \frac{\partial l_{ik}^{(r)}(\theta)}{\partial \beta}$$

and

$$S(\theta) \;=\; (S_1'(\theta), ..., S_{m-1}'(\theta))' \; .$$

Note that $S_r(\theta)$ can be viewed as the score function for β at time s_r under the working independence assumption and its distribution can be asymptotically approximated by a normal distribution with mean zero if model (2.1) is correct. Thus it is natural to base the goodness-of-fit test on $S(\hat{\theta})$ with large values of $\|S(\hat{\theta})\|$ indicating the invalidity of model (2.1). Also it will be seen below that $S(\hat{\theta})$ is asymptotically equivalent to a score test statistic for the hypothesis that model (2.1) is a special case of the Cox model with time-varying coefficients.

To investigate the distribution of $S(\hat{\theta})$, using the Taylor series expansion twice, we obtain

$$N^{1/2}(\hat{\theta} \,-\, \theta) \;=\; I^{-1}(\theta) \left\{ N^{-1/2} \, \frac{\partial l(\theta)}{\partial \theta} \right\} \,+\, o_p(1)$$

and

$$S_r(\hat{\theta}) \;=\; S_r(\theta) \,-\, N^{-1/2} \, I_{1r}(\theta) I^{-1}(\theta) \frac{\partial l(\theta)}{\partial \theta} \,+\, o_p(1) \; ,$$

where

$$I_{1r}(\theta) \;=\; -N^{-1}\sum_{i=1}^{N}\sum_{k=1}^{K}\frac{\partial^2 l_{ik}^{(r)}(\theta)}{\partial\beta\partial\theta'}\;,\; r=1,...,m-1\;.$$

This together with the multivariate central limit theorem suggests that the distribution of $S(\hat{\theta})$ can be approximated by a normal distribution with mean zero and covariance matrix that can be consistently estimated by

$$\hat{B} \;=\; \frac{1}{N}\sum_{i=1}^{N}B_i(\hat{\theta})B_i'(\hat{\theta}) \;-\; \left\{\frac{1}{N}\sum_{i=1}^{N}B_i(\hat{\theta})\right\}\times\left\{\frac{1}{N}\sum_{i=1}^{N}B_i(\hat{\theta})\right\}'\;,$$

where

$$B_i(\theta) = \begin{pmatrix} \sum_{k=1}^{K}\left\{\frac{\partial l_{ik}^{(1)}(\theta)}{\partial\beta} - I_{11}(\theta)I^{-1}(\theta)\frac{\partial l_{ik}(\theta)}{\partial\theta}\right\} \\ . \\ . \\ . \\ \sum_{k=1}^{K}\left\{\frac{\partial l_{ik}^{(m-1)}(\theta)}{\partial\beta} - I_{1,m-1}(\theta)I^{-1}(\theta)\frac{\partial l_{ik}(\theta)}{\partial\theta}\right\} \end{pmatrix}\;.$$

Hence it is natural to assess the adequacy of model (2.1) using the statistic

$$U \;=\; S'(\hat{\theta})\,\hat{B}^{-1}\,S(\hat{\theta})$$

based on the χ^2 distribution with the degrees of freedom $p(m-1)$.

Note that the statistic U involves the inverse of $\hat{B}$, the determination of which can be difficult due to its high dimension. To overcome this, we propose to use the standardized $S(\hat{\theta})$. Specifically, let v be the vector consisting of the square roots of the first p diagonal elements of $D(\hat{\theta})$ defined in (2.3), the estimates of the asymptotic variances of the components of $S_m(\theta)$. Also let $\hat{V}$ be the $p(m-1)\times p(m-1)$ diagonal matrix with the diagonal elements given by repeating the vector v $m-1$ times. Then one can carry out the goodness-of-fit test for model (2.1) by using the statistic

$$W \;=\; \|\hat{V}^{-1}S(\hat{\theta})\| \;=\; \left\{[\hat{V}^{-1}S(\hat{\theta})]'[\hat{V}^{-1}S(\hat{\theta})]\right\}^{1/2}\;,$$

the Euclidean norm of the standardized $S(\hat{\theta})$. By the law of large numbers, the distribution of W can be approximated by $\|N(0, V^{-1}BV^{-1})\|$, where V and B are the limits of $\hat{V}$ and $\hat{B}$, respectively.

To implement the goodness-of-fit test based on W, one needs to know the properties of $\|N(0, \hat{V}^{-1}\hat{B}\hat{V}^{-1})\|$ under model (2.1). To this end, following the idea used in Spiekerman and Lin (1996) among others, we propose to approximate W through simulation as follows. Let $(G_1, ..., G_N)$ be independent standard normal random variables independent of observed data and define

$$\tilde{S}(\hat{\theta}) = N^{-1/2} \sum_{i=1}^{N} G_i \left\{ B_i(\hat{\theta}) - N^{-1} \sum_{i=1}^{N} B_i(\hat{\theta}) \right\} .$$

Then it can be shown by following Spiekerman and Lin (1996) that the distribution of $\|\hat{V}^{-1}\tilde{S}(\hat{\theta})\|$ is given by $\|N(0, \hat{V}^{-1}\hat{B}\hat{V}^{-1})\|$ conditional on the observed data. This suggests that the null distribution of W or $\|N(0, \hat{V}^{-1}\hat{B}\hat{V}^{-1})\|$ can be approximated by the sampling distribution of $\|\hat{V}^{-1}\tilde{S}(\hat{\theta})\|$. Specifically, let w_0 denote the observed value of W and M an integer. To determine the p-value for testing the adequacy of model (2.1),

Step 1. For each j $(1 \le j \le M)$, generate an independent sample $(G_1^{(j)}, ..., G_N^{(j)})$ of size N from the standard normal distribution.

Step 2. Calculate $\tilde{S}^{(j)} = \tilde{S}(\hat{\theta})$ with replacing $(G_1, ..., G_N)$ by $(G_1^{(j)}, ..., G_N^{(j)})$, $j = 1, ..., M$.

Step 3. Calculate the p-value as $\left(\sum_{j=1}^{M} I(\|\hat{V}^{-1} \tilde{S}^{(j)}\| \ge w_0) + 1 \right) /(M + 1)$.

2.4 Numerical Results

First we report some results obtained from a simulation study with the set-up similar to the AIDS example discussed before for evaluating the proposed methodology. In the study, we assumed that $m = 11$ and $s_j = j$ and generated Z from a Bernoulli distribution with success probability 0.5. For the generation of failure times, we used the Gumbel's bivariate

exponential distribution given by

$$F(t_1, t_2) \;=\; F_1(t_1)\, F_2(t_2)\, [\, 1 \,+\, c\, \{1 \,-\, F_1(t_1)\}\, \{1 \,-\, F_2(t_2)\}\,]\,,$$

for the joint distribution of (T_1, T_2), where $F_1(t_1)$ and $F_2(t_2)$ denote the marginal distributions of T_1 and T_2, respectively, and c is a known constant. Under the above model, the correlation of the two failure times is $c\,/4$. For censoring intervals, to mimic the common structure of follow-up studies, it was assumed that every subject was supposed to be observed at s_j with the probability of missing an observation being $\phi \;=\; 0.2$. The end points L_{ik} and R_{ik} of censoring intervals were then defined as the last real observation time point before T_{ik} and the first real observation time point after T_{ik}, respectively, $k \;=\; 1, 2$. The results given below are based on $N \;=\; 200$ and $M \;=\; 1000$ with 1000 replications.

For assessing the size of the presented procedure based on statistic W, we used the exponential distribution with the hazard function $\lambda_0 \exp(\beta Z)$ for both F_1 and F_2, where $\lambda_0 \;=\; 0.1$. Table 2.1 presents the empirical sizes obtained from simulated interval-censored failure time data for the situations where $\beta \;=\; 0, 0.1, 0.25$ or 0.5 and $c \;=\; 0$ or 1, giving the correlation of T_1 and T_2 being 0 or 0.25, respectively. The results indicate that the procedure seems to have the right size. To investigate the power of the procedure, we again took F_1 and F_2 to the exponential distribution, but with different hazard functions. Specifically, we considered three nonproportional hazard functions: $\lambda_1(t|Z) \;=\; \lambda_0 \,+\, \beta\, Z$, $\lambda_2(t|Z) \;=\; \lambda_0\, t + \beta\, Z$ and $\lambda_3(t|Z) \;=\; \lambda_0 \exp(\beta Z\, t)$. Table 2.2 gives the estimated power under the three hazard functions and for $\beta \;=\; 0.1, 0.25$ or 0.5, respectively. It can be seen from Table 2.2 that the procedure possesses reasonably good power for the situations considered and has greater power under λ_3 than λ_1 and λ_2 as expected.

We plot the standardized statistic process S_r to compare its behavior under null hypothesis and alternative hypothesis. Figure 2.1 shows randomly selected 50 standardized test processes under null hypothesis (top one) and alternative hypothesis (bottom one). In

both cases, the true value $\hat{\beta}$ is 0.25 and the alternative hypothesis in Figure 2.1 has hazard function $\lambda_2(t|Z) = \lambda_0 t + \beta Z$. As we can see here, the test processes are centered around 0 under null hypothesis, while they deviate 0 clearly under alternative hypothesis. This difference contributes to the power of the proposed test. Another point is that under null hypothesis the standardized process tends to have a smaller variance as r increases. This pattern is more clear when β is larger.

Now we apply the proposed goodness-of-fit test to the set of bivariate interval-censored AIDS data mentioned above and discussed in Finkelstein et al. (2002) and Goggins and Finkelstein (2000). The data set considered here consists of 204 patients whose CMV shedding times in blood (T_1) and urine (T_2) are left-, right-, or interval-censored with month as the time unit. One of the objectives of the study was to determine whether the stage of HIV disease characterized by the CD4 count was predictive of an increased hazard for CMV shedding in either the blood or urine.

Following Goggins and Finkelstein (2000), define the covariate Z to be 1 if the baseline CD4 was less than 75 (in late stage of the HIV disease) or 0 otherwise. To study the covariate effect on CMV shedding in either the blood or urine, Goggins and Finkelstein (2000) suggested to use the marginal Cox model approach. Corresponding to this, we first considered the marginal Cox model (2.1). To check its appropriateness, we applied the goodness-of-fit test procedure based on W given in the previous section and obtained a p-value of almost zero. This indicates that model (2.1) may not be appropriate for the problem. To see this graphically, Figure 2.2 presents the separate nonparametric maximum likelihood estimators (NON) of the marginal survival functions based on univariate interval-censored data for CMV shedding in urine for patients with $Z = 1$ and 0, respectively, along with the maximum likelihood estimators (Cox) of the same functions given by $l(\theta)$. It suggests that model (2.1) does not seem to fit the data well. The similar plots were obtained for CMV shedding in blood.

Note that model (2.1) assumes that CMV sheddings in the blood and urine have the same baseline survival function. By relaxing this to allow different baseline survival functions, the application of the proposed goodness-of-fit test procedure based on W gave a p-value of 0.169, indicating that the marginal Cox model with different baseline hazard functions for blood and urine is reasonable. Corresponding to Figure 2.1, we obtained the same estimators under the new marginal model and presented the estimators in Figure 2.3, which suggests that now the fit is good. Under the new model, we obtained $\hat{\beta} = 0.9503$ with the estimated standard error being 0.1932. This suggests that the patients with lower baseline CD4 counts had higher risk of CMV shedding in the blood or urine than those with higher baseline CD4 counts. The conclusion is similar to that given in Goggins and Finkelstein (2000).

2.5 Concluding Remarks

A goodness-of-fit test procedure has been presented for assessing the validity of the marginal Cox model for correlated interval-censored failure time data. The basic idea behind the approach is the construction of a discrete score process and the simulation results suggest that the method works reasonably well for the situations considered. Note that for right-censored failure time data, a common approach for checking the Cox model is to employ the martingale-based residuals. In the presence of interval-censoring, however, the counting process formulation does not seem to be helpful anymore and it is not clear how the residual-based approaches for right-censored data can be generalized to interval-censored data.

An important step in the development of the presented test procedure is the definition of the observed interval-censored data up to a given time point. In addition to the two approaches discussed in Section 2.3, we also investigated several other methods and performed simulation studies on them. Simulation results suggest that the second method given in Section 2.3 and used in the numerical study seems always to perform better than others. A possible reason for this is that it takes advantage of the finite interval structure of the data.

It would be helpful to conduct a thorough and theoretical study on this. Another issue that needs to be investigated rigorously is the asymptotic properties of the proposed method.

To look at the statistic $S(\hat{\theta})$ from another point of view, assume that the underlying survival times T_k's are continuous with their hazard functions given by the following more general marginal Cox model

$$\lambda_k(t|Z_k) = \lambda_0(t) \exp\{\beta(t)'Z_k\} \tag{2.5}$$

instead of model (2.1), $k = 1, ..., K$. In the above, as β, $\beta(t)$ is the p-dimensional vector of regression parameters, but unlike β, it may be time-dependent. Model (2.1) can be obtained from model (2.5) if each subject is observed only at the s_j's and we define $\lambda_j = \exp\{-\int_{s_{j-1}}^{s_j} \lambda_0(t)dt\}$, $j = 1, ..., m$. Assume that $\beta(t)$ is constant within each interval $(s_{j-1}, s_j]$ and let β_j denote the value of $\beta(t)$ within $(s_{j-1}, s_j]$. Then the checking of model (2.1) is equivalent to testing $\beta_1 = ... = \beta_m = \beta$ under model (2.5), which can be naturally carried out based on the partial score function

$$\Big(S_1'(\hat{\theta}), \{S_2(\hat{\theta}) - S_1(\hat{\theta})\}', ..., \{S_m(\hat{\theta}) - S_{m-1}(\hat{\theta})\}'\Big)'$$

obtained under the working independence assumption and evaluated at $\gamma = \hat{\gamma}$ and $\beta_1 = ... = \beta_m = \hat{\beta}$. It is easy to see that $S(\hat{\theta})$ is asymptotically equivalent to the above score function since $S_m(\hat{\theta}) = 0$.

Chapter 3

Estimation of the Association Parameter for Bivariate Interval-censored Failure Time Data

3.1 Introduction

In this chapter, we study estimation of the association parameter for bivariate interval-censored data. Estimation of the dependence or association of correlated failure time events is a very important topic in multivariate data analysis. Many methods have been proposed to deal with this problem (Shih and Louis, 1995; Hsu and Prentice, 1996; Wang and Ding, 2000; Betensky and Finkelstein 1999). Among them, a two-stage procedure was proposed to estimate association parameter for bivariate right-censored and current status data by Shih and Louis (1995) and Wang and Ding (2000), respectively. In this chapter, we generalize their ideas and propose a two-stage estimation procedure for the same purpose for bivariate interval-censored data. The asymptotic properties of the proposed estimator are established and the simulation results suggest that the proposed estimation procedure works well for practical situations.

In the following, we study the estimation of the association of two correlated continuous survival variables based on interval-censored data. Let T_1 and T_2 be two failure times of interest with respective marginal survival functions $S_1(t)$ and $S_2(t)$ and joint survival function $S(t_1, t_2) = P(T_1 > t_1, T_2 > t_2)$. To estimate the association between T_1 and T_2, as most

authors, we will focus on the situation where (T_1, T_2) follow a copula model given by

$$S(t_1, t_2) \;=\; C_\alpha(S_1(t_1), S_2(t_2)) \;, \tag{3.1}$$

where C_α is a genuine survival function on the unit square and $\alpha \in R$ is a global association parameter. One attractive feature of model (3.1) is its modeling flexibility since it includes as special cases many useful bivariate failure time models such as the Archimedean copula family

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

where $0 \le \phi_\alpha \le 1$, $\phi_\alpha(0) = 1$, $\phi'_\alpha < 0$, $\phi''_\alpha > 0$. Here $\phi'_\alpha(u) = d\phi_\alpha(u)/du$ and $\phi''_\alpha(u) = d\phi'_\alpha(u)/du$. Taking $\phi_\alpha(u) = (1 + u)^{1/(1-\alpha)}$, the Laplace transformation of a gamma distribution, we have

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

which is commonly referred to as the Clayton family (Clayton, 1978). Another attractive feature of copula models is that marginal distributions do not depend on the choice of the association structure and thus one can model the marginal distributions and the association separately.

Another parameter measuring global association is Kendall's τ, which is defined as

$$\tau \;=\; Pr\{\,(T_{1i} - T_{1j})(T_{2i} - T_{2j}) > 0\} \;-\; Pr\{\,(T_{1i} - T_{1j})(T_{2i} - T_{2j}) < 0\}$$

for i.i.d. replicates (T_{1i}, T_{2i}) and (T_{1j}, T_{2j}) of (T_1, T_2). When the marginal distributions are uniform, τ can be evaluated as follows:

$$\tau \;=\; 4 \int_0^1 \int_0^1 S(u, v)\, f(u, v)\, du\, dv \;-\; 1 \,,$$

where $S(\cdot\,,\,\cdot)$ and $f(\cdot\,,\,\cdot)$ are the joint survival and density function, respectively, and $S(\cdot\,,\,\cdot)$

is given in (3.1). When the marginal distributions are not uniform, the same formula holds with integration range covering the full distribution. Kendall's τ has a nice property that it is unchanged by both linear and nonlinear increasing transformation (Hougaard, 2000).

Both α and τ measure the global association. Under the Clayton model, α is also the ratio of the hazard function of $T_1 = t_1$ given $T_2 = t_2$ to that given $T_2 \geq t_2$ or that of $T_2 = t_2$ given $T_1 = t_1$ against given $T_1 \geq t_1$ (Clayton, 1978). The relationship between α and τ is summarized as $\tau = (\alpha - 1)/(\alpha + 1)$ (Genest and MacKay, 1986; Genest and Rivest, 1993).

Many authors have considered the copula model for bivariate distributions (Clayton, 1978; Genest and Rivest, 1993; Hougaard, 1986). In particular, Shih and Louis (1995) and Wang and Ding (2000) discussed estimation of the association parameter under the model for bivariate right-censored and current status data, respectively. It does not seem, however, that there exists research on the association parameter for bivariate interval-censored data except Betensky and Finkelstein (1999), who considered estimation of the Kendall's τ by using the multiple imputation approach. However, they did not give a theoretical justification for their method.

The remainder of this chapter is organized as follows. In Section 3.2, we will describe the structure of observed data and some assumptions. A two-stage inference procedure for the association parameter α is presented in Section 3.3 and the asymptotic properties of the proposed estimate are established. Section 3.4 discusses estimation of the asymptotic variance of the proposed estimate. Due to the significant difference between censoring structures, the derived asymptotic variance estimate does not have the simple form as that for right-censored or current status data . To address this problem, a bootstrap procedure for variance estimation of the proposed point estimate is also presented. In Section 3.5, we report some simulation results for the assessment of the proposed method in addition to applying the method to the ACTG 181 data. Some concluding remarks are given in Section 3.6.

3.2 Data Structure and Assumptions

Consider a survival study involving two survival variables T_1 and T_2 of interest. Suppose that (T_1, T_2) are not exactly observable except for knowing that they belong to some intervals given by

$$\{U^{(1)}, V^{(1)}, \Delta_1^{(1)} = I(T_1 \leq U^{(1)}), \Delta_2^{(1)} = I(U^{(1)} < T_1 \leq V^{(1)})\}$$

and

$$\{U^{(2)}, V^{(2)}, \Delta_1^{(2)} = I(T_2 \leq U^{(2)}), \Delta_2^{(2)} = I(U^{(2)} < T_2 \leq V^{(2)})\} ,$$

where $(U^{(1)}, V^{(1)})$ and $(U^{(2)}, V^{(2)})$ are random monitoring times for T_1 and T_2, respectively, and $I(.)$ is the indicator function. We assume that (T_1, T_2) are independent of $(U^{(1)}, V^{(1)}, U^{(2)}, V^{(2)})$ but $(U^{(1)}, V^{(1)})$ and $(U^{(2)}, V^{(2)})$ could be dependent. Let $H(\boldsymbol{x})$ denote the joint distribution function of $(U^{(1)}, V^{(1)}, U^{(2)}, V^{(2)})$ and $G_\alpha(\boldsymbol{x}, \delta)$ the distribution function of $(U^{(1)}, V^{(1)}, U^{(2)}, V^{(2)}, \Delta)$, where $x = (x_1, x_2, x_3, x_4)$, $\delta = (\delta_1^{(1)}, \delta_2^{(1)}, \delta_1^{(2)}, \delta_2^{(2)})$ and $\Delta = (\Delta_1^{(1)}, \Delta_2^{(1)}, \Delta_1^{(2)}, \Delta_2^{(2)})$. The density or probability functions of $(U^{(1)}, V^{(1)}, U^{(2)}, V^{(2)})$ and $(U^{(1)}, V^{(1)}, U^{(2)}, V^{(2)}, \Delta)$ will be denoted by $h(\boldsymbol{x})$ and $g_\alpha(\boldsymbol{x}, \delta)$.

Suppose that observed data are n i.i.d. replicates of $(U^{(1)}, V^{(1)}, U^{(2)}, V^{(2)}, \Delta)$ and given by

$$\{U_i^{(1)}, V_i^{(1)}, U_i^{(2)}, V_i^{(2)}, \Delta_i, \quad i = 1, ..., n\} ,$$

where $\Delta_i = (\Delta_{1i}^{(1)}, \Delta_{2i}^{(1)}, \Delta_{1i}^{(2)}, \Delta_{2i}^{(2)})$. Then under model (3.1), the log likelihood function is given by

$$\log L(\alpha, S_1, S_2) = \sum_{i=1}^{n} l(\alpha, S_1, S_2, U_i^{(1)}, V_i^{(1)}, U_i^{(2)}, V_i^{(2)}, \Delta_i) , \tag{3.2}$$

where

$$l(\alpha, S_1, S_2, \boldsymbol{x}, \delta) = \delta_1^{(1)}\delta_1^{(2)} \log S_{11}(\alpha, \boldsymbol{x}) + \delta_1^{(1)}\delta_2^{(2)} \log S_{12}(\alpha, \boldsymbol{x})$$

$$+\delta_1^{(1)}(1 - \delta_1^{(2)} - \delta_2^{(2)}) \log S_{13}(\alpha, \boldsymbol{x}) + \delta_2^{(1)}\delta_1^{(2)} \log S_{21}(\alpha, \boldsymbol{x})$$

$$+\delta_2^{(1)}\delta_2^{(2)} \log S_{22}(\alpha, \boldsymbol{x}) + \delta_2^{(1)}(1 - \delta_1^{(2)} - \delta_2^{(2)}) \log S_{23}(\alpha, \boldsymbol{x})$$

$$+(1-\delta_1^{(1)}-\delta_2^{(1)})\delta_1^{(2)}\log S_{31}(\alpha,\boldsymbol{x})+(1-\delta_1^{(1)}-\delta_2^{(1)})\delta_2^{(2)}\log S_{32}(\alpha,\boldsymbol{x})$$

$$+(1-\delta_1^{(1)}-\delta_2^{(1)})(1-\delta_1^{(2)}-\delta_2^{(2)})\log S_{33}(\alpha,\boldsymbol{x})+\log h(\boldsymbol{x})$$

and

$$S_{11}(\alpha,\boldsymbol{x})=P(T_1\le x_1,T_2\le x_3)=1-S_1(x_1)-S_2(x_3)+C_\alpha(S_1(x_1),S_2(x_3))\,,$$

$$S_{12}(\alpha,\boldsymbol{x})=P(T_1\le x_1,x_3<T_2\le x_4)=S_2(x_3)-S_2(x_4)+C_\alpha(S_1(x_1),S_2(x_4))$$

$$-C_\alpha(S_1(x_1),S_2(x_3))\,,$$

$$S_{13}(\alpha,\boldsymbol{x})=P(T_1\le x_1,T_2>x_4)=S_2(x_4)-C_\alpha(S_1(x_1),S_2(x_4))\,,$$

$$S_{21}(\alpha,\boldsymbol{x})=P(x_1<T_1\le x_2,T_2\le x_3)=S_1(x_1)-S_1(x_2))+C_\alpha(S_1(x_2),S_2(x_3))$$

$$-C_\alpha(S_1(x_1),S_2(x_3))\,,$$

$$S_{22}(\alpha,\boldsymbol{x})=P(x_1<T_1\le x_2,x_3<T_2\le x_4)=C_\alpha(S_1(x_1),S_2(x_3))-C_\alpha(S_1(x_1),S_2(x_4))$$

$$-C_\alpha(S_1(x_2),S_2(x_3))+C_\alpha(S_1(x_2),S_2(x_4))\,,$$

$$S_{23}(\alpha,\boldsymbol{x})=P(x_1<T_1\le x_2,T_2>x_4)=C_\alpha(S_1(x_1),S_2(x_4))-C_\alpha(S_1(x_2),S_2(x_4))\,,$$

$$S_{31}(\alpha,\boldsymbol{x})=P(T_1>x_2,T_2\le x_3)=S_1(x_2)-C_\alpha(S_1(x_2),S_2(x_3))\,,$$

$$S_{32}(\alpha,\boldsymbol{x})=P(T_1>x_2,x_3<T_2\le x_4)=C_\alpha(S_1(x_2),S_2(x_3))-C_\alpha(S_1(x_2),S_2(x_4))\,,$$

$$S_{33}(\alpha,\boldsymbol{x})=P(T_1>x_2,T_2>x_4)=C_\alpha(S_1(x_2),S_2(x_4))\,.$$

When $\delta_2^{(1)}=0$ and $\delta_2^{(2)}=0$, $\log L(\alpha,S_1,S_2)$ reduces to the log-likelihood for bivariate current status data (Wang and Ding, 2000). In the next section, we will discuss estimation of the association parameter α.

3.3 Estimation of the Association Parameter

To estimate α, note that if the marginal survival functions S_1 and S_2 are known, a natural estimator is then given by the maximum likelihood estimator from (3.2). This naturally leads to the following two-stage procedure: first estimate S_1 and S_2 and then estimate α by maximizing the pseudo log likelihood given by $\log L(\alpha, S_1, S_2)$ with S_1 and S_2 replaced by their estimates. The same idea was used by Shih and Louis (1995) and Wang and Ding (2000) among others.

In the first stage, we propose to consider the univariate sample

$$\{U_i^{(r)}, V_i^{(r)}, \Delta_{1i}^{(r)} = I(T_{ri} \leq U_i^{(r)}), \Delta_{2i}^{(r)} = I(U_i^{(r)} < T_{ri} \leq V_i^{(r)}), i = 1, ..., n\}$$

and to estimate S_r by the nonparametric maximum likelihood estimator $\hat{S}_r$ given by maximizing the univariate interval-censored data likelihood

$$L_r = \prod_{i=1}^{n} (1 - S_r(U_i^{(r)}))^{\Delta_{1i}^{(r)}} (S_r(U_i^{(r)}) - S_r(V_i^{(r)}))^{\Delta_{2i}^{(r)}} S_r(V_i^{(r)})^{1-\Delta_{1i}^{(r)}-\Delta_{2i}^{(r)}}$$

given $\{(U_i^{(r)}, V_i^{(r)}), i = 1, ..., n\}$, $r = 1, 2$. Several algorithms for maximizing L_r have been proposed including the self-consistency and iterative convex minorant algorithms given in Turnbull (1976) and Groeneboom and Wellner (1992), respectively. Given $\hat{S}_1$ and $\hat{S}_2$, the association parameter α can be estimated by the solution $\hat{\alpha}$ to the pseudo score equation $U(\alpha, \hat{S}_1, \hat{S}_2, \hat{G}_n) = 0$, where

$$U(\alpha, \hat{S}_1, \hat{S}_2, \hat{G}_n) = \frac{1}{n} \frac{\partial}{\partial \alpha} \log L(\alpha, \hat{S}_1, \hat{S}_2) = \int \frac{\partial}{\partial \alpha} l(\alpha, \hat{S}_1, \hat{S}_2, \boldsymbol{x}, \delta) dG_n(\boldsymbol{x}, \delta) \tag{3.3}$$

with $G_n(\boldsymbol{x}, \delta)$ denoting the empirical estimator of $G_\alpha(\boldsymbol{x}, \delta)$. It can be easily shown that $\hat{\alpha}$ is consistent and the above equation can be solved by a standard root finding method or the Newton-Raphson algorithm.

The asymptotic distribution of $\hat{\alpha}$ depends on some regularity conditions on the im-

posed copula model and the plugged-in estimators $\hat{S}_r$ $(r = 1, 2)$. Since the convergence rate of the nonparametric maximum likelihood estimator for interval-censored data is only $n^{1/3}$ (Groeneboom and Wellner, 1992), the asymptotic expansion of $\hat{S}_r$ $(r = 1, 2)$ are more complex in the present setting than in the case of right-censored data. In the following, using the results given in Groeneboom and Wellner (1992) and Geskus and Groeneboom (1999), we show that, under suitable assumptions, the proposed estimator $\hat{\alpha}$ converges to a normal random variable at the standard rate $n^{1/2}$.

Let α_0 be the true value of α and let $\Psi_r(t)$ be the influence curve of the functional $U(\alpha_0, S_1, S_2, G_{\alpha_0})$ at S_r, obtained by differentiating $U(\alpha_0, (1-\epsilon_1)S_1 + \epsilon_1 S_1, (1-\epsilon_2)S_2 + \epsilon_2 S_2, G_{\alpha_0})$ with respect to ϵ_r $(r = 1, 2)$ and evaluating at $\epsilon_1 = \epsilon_2 = 0$. Moreover, let ϕ_{S_r} denote the solution to the Fredholm integral equation

$$\phi_{S_r}(t) = d_{S_r}(t)\left\{w_r(t) - \int_0^t \frac{\phi_{S_r}(t) - \phi_{S_r}(x)}{S_r(x) - S_r(t)} h_r(x,t)dx + \int_t^\infty \frac{\phi_{S_r}(x) - \phi_{S_r}(t)}{S_r(t) - S_r(x)} h_r(t,x)dx\right\},$$

where $d_{S_r}(t) = S_r(t)(1 - S_r(t))/(h_{r1}(t)S_r(t) + h_{r2}(t)(1 - S_r(t)))$, $w_r(t) = d\Psi_r(t)/dt$, h_r is the density function of $(U^{(r)}, V^{(r)})$, and h_{r1} and h_{r2} are the marginal densities of $U^{(r)}$ and $V^{(r)}$, respectively. Define

$$\Phi_r(u, v, \delta_1^{(r)}, \delta_2^{(r)}) = -\delta_1^{(r)} \frac{\phi_{S_r}(u)}{1 - S_r(u)} - \delta_2^{(r)} \frac{\phi_{S_r}(v) - \phi_{S_r}(u)}{S_r(u) - S_r(v)} + (1 - \delta_1^{(r)} - \delta_2^{(r)}) \frac{\phi_{S_r}(v)}{S_r(v)}.$$

The asymptotic distribution of $\hat{\alpha}$ is then given by the following theorem.

Theorem 3.1. *Under mild regularity conditions, $n^{1/2}(\hat{\alpha} - \alpha_0)$ has an asymptotic normal distribution with mean zero and variance*

$$\sigma^2 = (A(\alpha_0, S_1, S_2, G_{\alpha_0}))^{-2} Var\{B(\alpha_0, S_1, S_2, U^{(1)}, V^{(1)}, U^{(2)}, V^{(2)}, \Delta)\},$$

where

$$A(\alpha, S_1, S_2, G_\alpha) = \int \frac{\partial^2}{\partial \alpha^2} l(\alpha, S_1, S_2, \boldsymbol{x}, \delta) dG_\alpha(\boldsymbol{x}, \delta)$$

and

$$B(\alpha, S_1, S_2, \boldsymbol{x}, \delta) = \frac{\partial}{\partial \alpha} l(\alpha, S_1, S_2, \boldsymbol{x}, \delta) - \Phi_1(x_1, x_2, \delta_1^{(1)}, \delta_2^{(1)}) - \Phi_2(x_3, x_4, \delta_1^{(2)}, \delta_2^{(2)}) \,.$$

The conditions and proof of Theorem 3.1 are attached in the first part of Appendix A. In the next section, we discuss how to estimate the asymptotic variance σ^2.

3.4 Variance Estimation

We first consider the deviation of a consistent estimate of the asymptotic variance σ^2. For $r = 1, 2$, let $0 < t_1^{(r)} < ... < t_{m_r}^{(r)} < \infty$ denote the time points at which $\hat{S}_r$ jumps and $z_j^{(r)} = 1 - \hat{S}_r(t_j^{(r)})$. According to Theorem 3.5 of Groeneboom (1996), $\phi_{\hat{S}_r}$ is absolutely continuous with respect to $\hat{S}_r$ and a step function with jumps at the $t_j^{(r)}$'s. Let $\hat{H}_{rn}$, $\hat{H}_{r1n}$ and $\hat{H}_{r2n}$ denote the empirical distribution functions of $(U^{(r)}, V^{(r)})$, $U^{(r)}$ and $V^{(r)}$, respectively. Define $\tilde{\Psi}_r(t)$ as $\Psi_r(t)$ with $\alpha_0, S_1, S_2, G_{\alpha_0}$ replaced by $\hat{\alpha}, \hat{S}_1, \hat{S}_2, \hat{G}_n$, respectively. Let

$$\Delta_j(h_{rl}) = \int_{t_j^{(r)}}^{t_{j+1}^{(r)}} h_{rl}(t)dt \approx \int_{t_j^{(r)}}^{t_{j+1}^{(r)}} d\hat{H}_{rln}(t), \quad l = 1, 2,$$

$$\Delta_{jk}(h_r) = \int_{u=t_j^{(r)}}^{t_{j+1}^{(r)}} \int_{v=t_k^{(r)}}^{t_{k+1}^{(r)}} h_r(u, v)dudv \approx \int_{u=t_j^{(r)}}^{t_{j+1}^{(r)}} \int_{v=t_k^{(r)}}^{t_{k+1}^{(r)}} d\hat{H}_{rn}(u, v),$$

$$d_j^{(r)} = \frac{z_j^{(r)}(1 - z_j^{(r)})}{\Delta_j(h_{r1})(1 - z_j^{(r)}) + \Delta_j(h_{r2})z_j^{(r)}},$$

$\Delta_j(w_r) = \tilde{\Psi}(t_{j+1}^{(r)}) - \tilde{\Psi}(t_j^{(r)})$, $j, k = 1, ..., m_r$. Moreover, let $y_j^{(r)} = \phi_{\hat{S}_r}(t_j^{(r)})$. Then it can be shown that the vector $y^{(r)} = (y_1^{(r)}, ..., y_{m_r}^{(r)})'$ $(r = 1, 2)$ is the unique solution to the following set of linear equations (e.g. Theorem 3.1 of Geskus and Groeneboom, 1999)

$$\begin{aligned} & y_j^{(r)} \left\{ (d_j^{(r)})^{-1} + \sum_{k<j} \frac{\Delta_{kj}(h_r)}{z_j^{(r)} - z_k^{(r)}} + \sum_{k>j} \frac{\Delta_{jk}(h_r)}{z_k^{(r)} - z_j^{(r)}} \right\} \\ = \quad & \Delta_j(w_r) + \sum_{k<j} \frac{\Delta_{kj}(h_r)}{z_j^{(r)} - z_k^{(r)}} y_k^{(r)} + \sum_{k>j} \frac{\Delta_{jk}(h_r)}{z_k^{(r)} - z_j^{(r)}} y_k^{(r)} \end{aligned}$$

for $j = 1, ..., m_r$.

Define

$$\tilde{\Phi}_r(u, v, \delta_1^{(r)}, \delta_2^{(r)}) = -\delta_1^{(r)} \frac{\phi_{\hat{S}_r}(u)}{1 - \hat{S}_r(u)} - \delta_2^{(r)} \frac{\phi_{\hat{S}_r}(v) - \phi_{\hat{S}_r}(u)}{\hat{S}_r(u) - \hat{S}_r(v)} + (1 - \delta_1^{(r)} - \Delta_2^{(r)}) \frac{\phi_{\hat{S}_r}(v)}{\hat{S}_r(v)}.$$

Since $\hat{H}_{rn}$, $\hat{H}_{r1n}$, $\hat{H}_{r2n}$, $\hat{\alpha}$ and $\hat{S}_r$ are uniformly consistent estimates (Van der Vaart and Wellner, 1996; Groeneboom and Wellner, 1992), $\tilde{\Phi}_r(u, v, \delta_1^{(r)}, \delta_2^{(r)})$ is uniformly consistent to $\Phi_r(u, v, \delta_1^{(r)}, \delta_2^{(r)})$. Also define

$$\tilde{B}(\boldsymbol{x}, \delta) = \frac{\partial}{\partial \alpha} l(\hat{\alpha}, \hat{S}_1, \hat{S}_2, \boldsymbol{x}, \delta) - \tilde{\Phi}_1(x_1, x_2, \delta_1^{(1)}, \delta_2^{(1)}) - \tilde{\Phi}_2(x_3, x_4, \delta_1^{(2)}, \delta_2^{(2)}).$$

Note that $\tilde{B}$ is continuous in α and S_r. It then follows that

$$Var\{\tilde{B}(U^{(1)}, V^{(1)}, U^{(2)}, V^{(2)}, \Delta)\} \to Var\{B(\alpha_0, S_1, S_2, U^{(1)}, V^{(1)}, U^{(2)}, V^{(2)}, \Delta)\}$$

because of the uniform consistency of $\tilde{\Phi}_r$ and $\hat{S}_r$ and the consistency of $\hat{\alpha}$. It is well known that $A(\hat{\alpha}, \hat{S}_1, \hat{S}_2, \hat{G}_n)$ and the sample variance of $\{ \tilde{B}(U_i^{(1)}, V_i^{(1)}, U_i^{(2)}, V_i^{(2)}, \Delta_i), i = 1, ..., n\}$ are consistent estimators of $A(\alpha_0, S_1, S_2, G_{\alpha_0})$ and $Var\{\tilde{B}(U^{(1)}, V^{(1)}, U^{(2)}, V^{(2)}, \Delta)\}$, respectively. This suggests that the variance σ^2 can be consistently estimated by $\hat{\sigma}^2 = (A(\hat{\alpha}, \hat{S}_1, \hat{S}_2, \hat{G}_n))^{-2} \hat{\sigma}_1^2$, where

$$\hat{\sigma}_1^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left\{ \tilde{B}(U_i^{(1)}, V_i^{(1)}, U_i^{(2)}, V_i^{(2)}, \Delta_i) - \bar{B} \right\}^2,$$

and

$$\bar{B} = \frac{1}{n} \sum_{i=1}^{n} \tilde{B}(U_i^{(1)}, V_i^{(1)}, U_i^{(2)}, V_i^{(2)}, \Delta_i).$$

Note that in general, the estimator $\hat{\sigma}^2$ could be very technically involved due to the complexity of the estimator $\phi_{\hat{S}_r}$. Corresponding to this, we propose to use the bootstrap procedure for variance estimation. One simple and natural approach is to draw bootstrap samples of size n with replacement from the observed data $\{ (U_i^{(1)}, V_i^{(1)}, U_i^{(2)}, V_i^{(2)}, \Delta_i) ; i =$

$1, ..., n\,\}$ independently for M times, where M is a prespecified integer. This yields M estimators $\{\,\tilde{\alpha}_k\,;\, k = 1, ..., M\,\}$ of α based on each of M bootstrap samples. The variance of $\hat{\alpha}$ can then be naturally estimated by the sample variance of the $\tilde{\alpha}_k$'s. Alternatively, given the observed original data, a bootstrap sample $\{\,U_i^{(1)}, V_i^{(1)}, U_i^{(2)}, V_i^{(2)}, \Delta_i^*, i = 1, ..., n\,\}$ can be generated by only generating indicators

$$\Delta_i^* = (\tilde{\Delta}_{1i}^{(1)}, \tilde{\Delta}_{2i}^{(1)}, \tilde{\Delta}_{1i}^{(2)}, \tilde{\Delta}_{2i}^{(2)})\,,$$

where $(\tilde{\Delta}_{1i}^{(r)}, \tilde{\Delta}_{2i}^{(r)})$ are generated from a multinomial distribution with values $(1,0), (0,1)$ and $(0,0)$ and the corresponding probabilities $1 - \hat{S}_r(U_i^{(r)})$, $\hat{S}_r(U_i^{(r)}) - \hat{S}_r(V_i^{(r)})$ and $\hat{S}_r(V_i^{(r)})$, respectively, $r = 1, 2$.

3.5 Numerical Results

First we present some results from simulation studies conducted for assessing the performance of the inference procedure presented in the previous sections. In the study, we used the Clayton model

$$S(t_1, t_2) = [\,S_1^{1-\alpha}(t_1) + S_2^{1-\alpha}(t_2) - 1\,]^{1/(1-\alpha)}\,,\ \alpha > 0 \tag{3.4}$$

for the joint survival function with $S_r(t) = \exp(-0.1\,t)$ for the marginal survival function, $r = 1, 2$. For censoring intervals, we assumed that $U^{(1)}$ and $V^{(1)}$ followed uniform distributions such that $U^{(1)} \sim U(0, 3.98)$ and $V^{(1)} = U^{(1)} + W$ with $W \sim U(0, 6.02)$. The censoring intervals for T_2 were generated in the same way independently. The results below are based on 500 replications.

Table 3.1 presents the simulation results for estimation of α and the Kendall's τ by using the method proposed in the previous sections with $\hat{\tau} = (\hat{\alpha} - 1)/(\hat{\alpha} + 1)$. In the table, we considered the situations where $n = 200$ or 400 and $\alpha = 2$ or 3, giving $\tau = 1/3$ or $1/2$. In addition to $\hat{\alpha}$ and $\hat{\tau}$, for comparison, we also estimated α and τ by $\tilde{\alpha}$ and $\tilde{\tau}$, where

$\tilde{\alpha}$ is defined as the solution to $U(\alpha, S_1, S_2, \hat{G}_n) = 0$ with U given in equation (3.3) and $\tilde{\tau} = (\tilde{\alpha} - 1)/(\tilde{\alpha} + 1)$. Note that the difference between $\hat{\alpha}$ and $\tilde{\alpha}$ is that for $\tilde{\alpha}$, the marginal survival functions S_1 and S_2 were assumed to be known and used, while for $\hat{\alpha}$, their estimates were used.

The results in Table 3.1 include the estimated bias given by the averages of point estimators minus the true values (Bias), the sample standard errors of the point estimators (SSE), the square root of the average of the estimated variances (ESE), and the 95% empirical coverage probabilities (CP). For variance estimation of the two estimators of α, the simple bootstrap method with $M = 200$ described in Section 3.4 was used, while for that of the two estimators of τ, by the delta method, we used $\sigma_\tau = 2\,\sigma_\alpha/(\alpha+1)^2$ with α replaced by its estimator, where σ_τ and σ_α denote the estimated standard errors of the estimators of τ and α, respectively. It can be seen from the table that all estimators seem to be unbiased and the estimated standard errors of the estimators defined in the previous sections are close to the sample standard errors. Also the method seems to give reasonable empirical coverage probabilities and as expected, the results are better when the sample size increases.

In the study, we investigated some other values of M for the bootstrap variance estimation and it seems that $M = 200$ used above is large enough for the situations considered here. To assess the asymptotic normality given in Theorem 3.1, we investigated the quantile plots of the standardized $\hat{\alpha}$ against the standard normal variable and they indicate that the normality approximation seems reasonable for the situations considered here.

Now we apply the proposed method to the bivariate interval-censored AIDS data discussed in Chapter 1 and in Goggins and Finkelstein (2000). Define T_1 and T_2 to be the times to the occurrences of CMV shedding in blood and urine, respectively, and assume that they follow the Clayton model (3.4). The application of the proposed method with $M = 200$ gave $\hat{\alpha} = 2.8070$ with the estimated standard error of 0.5307. This yielded $\hat{\tau} = 0.4646$ with the estimated standard error being 0.0732 by the delta method. Note that under the model

(3.4), $\alpha \to 1$ or $\tau \to 0$ means the independence of the CMV shedding times in blood and urine. The testing of $\alpha = 1$ against $\alpha > 1$ based on the standard normal distribution gave a p-value of 0.0003, while the testing of $\tau = 0$ against $\tau > 0$ based on the standard normal distribution gave a p-value of less than 0.0001. These suggest that the CMV shedding times in blood and urine were significantly correlated.

3.6 Concluding Remarks

A two-step statistical procedure for estimation of the association parameter for bivariate interval-censored failure time data under the copula model was proposed. Both finite and asymptotic properties of the proposed estimator were established with the simulation results indicating that the method works well for practical situations. The approach is a generalization of the corresponding approaches for bivariate right-censored or current status failure time data (Shih and Louis, 1995; Wang and Ding, 2000). Note that the censoring mechanism behind interval-censored data is much more complicated than that behind right-censored or current status data. This is because for the former case, one has to deal with two related censoring variables, while for the latter case, only one censoring variable is involved.

To investigate the association between two correlated variables, an alternative to the proposed approach is to direct estimate the Kendall's τ as Betensky and Finkelstein (1999). Although the method is intuitively appealing and simple, it seems very difficult to study the asymptotic properties of the method. A limitation of the proposed approach is that it applies only to situations where the joint survival function follows the copula model although it is one of the most commonly used models for bivariate data. The idea discussed here can be applied to the case where there exist covariates and they affect the joint survival function through marginal survival functions. For some of commonly used regression models for univariate failure time data, see Sun (2005).

Chapter 4

A Conditional Approach for Regression Analysis of Interval-censored Failure Time Data with the Additive Hazards Model

4.1 Introduction

This chapter considers regression analysis of interval-censored failure time data. To keep the simplicity, unlike the previous two chapters, we focus on univariate interval-censored data and the proposed method can be generalized to multivariate interval-censored data. In Chapter 5, we will return to multivariate survival data.

For regression analysis of interval-censored data, a few methods have been proposed. For example, Finkelstein (1986) is the first to consider fitting the proportional hazards model to general interval-censored data and Hunag (1996) studied the same problem for current status data and established asymptotic properties of the approach. Rossini and Tsiatis (1996) discussed regression analysis of current status data using the proportional odds model, while Lin et al. (1998) and Martinussen and Scheike (2002) considered the same, but using the additive hazards model. Also Rabinowitz et al. (1995) and Betensky et al. (2001) investigated the use of the accelerated failure time model for case 2 interval-censored data. More references can be found in Sun (2005). In this chapter, we propose an easy procedure for regression analysis of case 2 interval-censored data using the additive hazards model.

The remainder of this chapter is organized as follows. Section 4.2 introduces notations

and models that are used throughout the chapter. In Section 4.3, a conditional estimation approach is presented for estimating regression parameters of interest. The approach makes use of counting process theory and a main advantage of it is that it does not involve estimation of the cumulative baseline hazard function. The consistency and asymptotic normality of the proposed estimate of regression parameters are established. In Section 4.4, we show that the proposed method can be also applied to the case of informative censoring, when the failure time(s) and censoring time share the common unobserved random process. Simulation studies are conducted in Section 4.5 to evaluate finite sample properties of the proposed estimate under both independent and dependent censoring, and the breast cancer data is analyzed to illustrate the proposed method. Section 4.6 contains some concluding remarks.

4.2 Notations and Models

Consider a survival study that consists of n independent subjects. For subject i, let T_i denote the survival time of interest and $Z_i(t)$ a p-dimensional vector of covariates that may depend on time. Assume that T_i is not observable except for knowing that it belongs to an interval. Specifically, suppose that we observe two random variables U_i and V_i with $U_i \leq V_i$ and the indicator variables $\delta_{1\,i} = I(T_i < U_i)$, $\delta_{2\,i} = I(U_i \leq T_i < V_i)$ and $\delta_{3\,i} = 1 - \delta_{1i} - \delta_{2i}$, where I is the indicator function. Here U_i and V_i can be regarded as two examination times that belong to a sequence of random examination times. The variables $\delta_{1\,i}$, $\delta_{2\,i}$ and $\delta_{3\,i}$ indicate whether the survival event of interest for subject i has occurred before U_i, during the examination interval $[U_i, V_i)$, or after V_i, respectively. Then the observed data consist of $\{\,(U_i, V_i, \delta_{1\,i}, \delta_{2\,i}, \delta_{3\,i}, Z_i(t))\,;\, i = 1, 2, ..., n\,\}$. In this paper, we assume that the examination times U and V are independent of the survival time T given Z and that covariate process $Z(t)$ is completely observed.

In the following, we consider the analysis of observed interval-censored data using the additive hazards model. Specifically, we assume that the hazard function of T_i at time t

given the covariate process up to t is given by

$$\lambda_i(t \,|\, Z_i(s), s \le t) \;=\; \lambda_0(t) \;+\; \beta_0' \, Z_i(t) \;, \tag{4.1}$$

where $\lambda_0(t)$ is an unknown baseline hazard function and β_0 denotes the p-dimensional vector of regression parameters.

For censoring times U_i and V_i, note that $P(V_i \ge U_i) = 1$. In the following, it will be assumed that given Z_i, the marginal and conditional hazard functions of U_i and V_i are given by

$$\lambda_i^U(t \,|\, Z_i(s), s \le t) \;=\; \lambda_1(t) \, e^{\gamma_0' \, Z_i(t)} \tag{4.2}$$

and

$$\lambda_i^V(t \,|U_i = u_i, \; Z_i(s), s \le t) \;=\; \begin{cases} \lambda_2(t) \, e^{\gamma_0' \, Z_i(t)} & \text{if} \quad t \ge u_i \\ 0 & \text{if} \quad t < u_i \end{cases} \tag{4.3}$$

respectively. In the above, $\lambda_1(t)$ and $\lambda_2(t)$ denote unspecified baseline hazard functions and γ_0 is a p-dimensional vector of unknown regression parameters. Among others, Kelly and Lim (2000) and Prentice et al. (1981) discussed similar models for regression analysis of multivariate failure time data and recurrent event data, respectively.

For each i, define a 0-1 counting process $N_i^{(1)}(t) = (1 - \delta_{1i}) \, I(U_i \le t)$ and conditional on $U_i = u_i$, define $N_i^{(2)}(t) \;= \delta_{3i} \, I(V_i \le t)$ if $t \ge u_i$ and 0 if $t < u_i$. Then following the same arguments as those in Lin et al. (1998) and under models (4.1) - (4.3), we can derive the intensity functions of $N_i^{(1)}(t)$ and $N_i^{(2)}(t)$ as

$$I(U_i \ge t)\lambda_i^{(1)}(t \,|\, Z_i(s), s \le t) \quad \text{and} \quad I(u_i \le t \le V_i)\lambda_i^{(2)}(t \,|\, Z_i(s), u_i < s \le t),$$

respectively, where

$$\lambda_i^{(1)}(t \,|\, Z_i(s), s \le t) \;=\; \lambda_1^*(t) \, e^{-\beta_0' Z_i^*(t) + \gamma_0' Z_i(t)} \;=\; \lambda_1(t) \, e^{-\Lambda_0(t)} \, e^{-\beta_0' Z_i^*(t) + \gamma_0' Z_i(t)}, \tag{4.4}$$

and

$$\lambda_i^{(2)}(t \mid Z_i(s), u_i < s \le t) \;=\; \lambda_2^*(t)\, e^{-\beta_0' Z_i^*(t)+\gamma_0' Z_i(t)} \;=\; \lambda_2(t)\, e^{-\Lambda_0(t)}\, e^{-\beta_0' Z_i^*(t)+\gamma_0' Z_i(t)}. \qquad (4.5)$$

In the above,

$$Z_i^*(t) \;=\; \int_0^t Z_i(s)\, ds \;\text{ and }\; \Lambda_0(t) \;=\; \int_0^t \lambda_0(s)\, ds\ .$$

For notational convenience, we will assume that $\lambda_i^{(2)}(t \mid Z_i(s), s \le t) = 0$ for any $t \le u_i$ given $U_i = u_i$. It is apparent that model (4.4) is a marginal proportional hazards model, but model (4.5) is not. For model (4.5), the starting time point is u_i, the realization of U_i, and it is a conditional model. In the next section, we will use (4.4) and (4.5) to construct estimating equations for regression coefficients β_0 and γ_0.

4.3 Estimation Procedure

To estimate β_0 and γ_0, for $j = 0, 1$, define

$$S_{1,\beta}^{(j)}(t, \beta, \gamma) \;=\; n^{-1} \sum_{i=1}^{n} I(t \le U_i)\, e^{-\beta' Z_i^*(t)+\gamma' Z_i(t)}\, Z_i^{*\,(j)}(t)$$

and

$$S_{2,\beta}^{(j)}(t, \beta, \gamma) \;=\; n^{-1} \sum_{i=1}^{n} I(u_i < t \le V_i)\, e^{-\beta' Z_i^*(t)+\gamma' Z_i(t)}\, Z_i^{*\,(j)}(t)\ ,$$

where $Z_i^{*\,(0)}(t) = 1$ and $Z_i^{*\,(1)}(t) = Z_i^*(t)$. Motivated by Lin et al. (1998), who considered regression analysis of current status data using models (4.1) and (4.2), we propose to use the estimating function

$$\begin{aligned} U_\beta(\beta, \gamma) \;&=\; \sum_{i=1}^{n} \sum_{k=1}^{2} \int_0^\infty \left\{ Z_i^*(t) - \frac{S_{k,\beta}^{(1)}(t, \beta, \gamma)}{S_{k,\beta}^{(0)}(t, \beta, \gamma)} \right\} dN_i^{(k)}(t) \\ &=\; \sum_{i=1}^{n} (1 - \delta_{1i}) \left\{ Z_i^*(u_i) - \frac{S_{1,\beta}^{(1)}(u_i, \beta, \gamma)}{S_{1,\beta}^{(0)}(u_i, \beta, \gamma)} \right\} + \sum_{i=1}^{n} \delta_{3i} \left\{ Z_i^*(v_i) \;- \frac{S_{2,\beta}^{(1)}(v_i, \beta, \gamma)}{S_{2,\beta}^{(0)}(v_i, \beta, \gamma)} \right\} \end{aligned}$$

for estimation of β_0 given γ_0.

In $U_\beta(\beta,\gamma)$, the first term is the partial likelihood score function under (4.4) if one had only current status data and is unbiased. The second term is the partial likelihood score function obtained under model (4.5) if one considers only current status data given by the V_i's and thus also has mean 0 at γ_0 and β_0 due to the fact that each integral is a martingale given $U_i = u_i$. Thus $U_\beta(\beta,\gamma)$ is unbiased. The key idea here is to reduce general interval-censored data to current status data and similar ideas have been used by Betensky et al. (2001) among others.

For estimation of γ_0, one can easily develop an estimating function that is similar to $U_\beta(\beta,\gamma)$. On the other hand, note that for the U_i's and V_i's or models (4.2) and (4.3), complete data are available and thus it is more efficient to directly estimate γ_0 from them. To this end, define $\tilde{N}_i^{(1)}(t) = I(U_i \le t)$ and $\tilde{N}_i^{(2)}(t) = I(V_i \le t)$ if $t \ge u_i$ and 0 if $t < u_i$, $i = 1, ..., n$. Also define

$$S_{1,\gamma}^{(j)}(t,\gamma) = n^{-1} \sum_{i=1}^{n} I(t \le U_i)\, e^{\gamma' Z_i(t)}\, Z_i^{(j)}(t)$$

and

$$S_{2,\gamma}^{(j)}(t,\gamma) = n^{-1} \sum_{i=1}^{n} I(U_i < t \le V_i)\, e^{\gamma' Z_i(t)}\, Z_i^{(j)}(t)\ ,$$

$j = 0, 1$, where $Z_i^{(0)}(t) = 1$ and $Z_i^{(1)}(t) = Z_i(t)$. Then under models (4.2) and (4.3), a partial likelihood score function of γ_0 can be derived as

$$\begin{aligned} U_\gamma(\gamma) &= \sum_{i=1}^{n} \left[\int_0^\infty \left(Z_i(t) - \frac{S_{1,\gamma}^{(1)}(t,\gamma)}{S_{1,\gamma}^{(0)}(t,\gamma)} \right) d\tilde{N}_i^{(1)}(t) + \int_0^\infty \left(Z_i(t) - \frac{S_{2,\gamma}^{(1)}(t,\gamma)}{S_{2,\gamma}^{(0)}(t,\gamma)} \right) d\tilde{N}_i^{(2)}(t) \right] \\ &= \sum_{i=1}^{n} \left(Z_i(u_i) - \frac{S_{1,\gamma}^{(1)}(u_i,\gamma)}{S_{1,\gamma}^{(0)}(u_i,\gamma)} \right) + \sum_{i=1}^{n} \left(Z_i(v_i) - \frac{S_{2,\gamma}^{(1)}(v_i,\gamma)}{S_{2,\gamma}^{(0)}(v_i,\gamma)} \right) \end{aligned}$$

(Lin, 1994).

Let $\hat{\gamma}$ be the solution to the equation $U_\gamma(\gamma) = 0$. Then we can estimate β_0 by $\hat{\beta}$ defined as the root to the equation $U_\beta(\beta,\hat{\gamma}) = 0$. Let $\hat{A}_\beta(\beta,\gamma) = -n^{-1}\, \partial U_\beta(\beta,\gamma)/\partial\beta$ and A_β denote the limit of $\hat{A}_\beta(\beta,\gamma)$ at $\beta = \beta_0$ and $\gamma = \gamma_0$. It can be easily shown that $\hat{\gamma}$ is consistent

and has an asymptotic normal distribution (Lin, 1994; Wei et al., 1989). The consistency of $\hat{\beta}$ can be similarly proved by noting the facts that $\hat{A}_\beta(\beta, \hat{\gamma})$ is positive semidefinite and its limit is assumed to be positive definite at β_0.

For the asymptotic distribution of $\hat{\beta}$, we show in the Appendix A that $n^{-1/2}\, U_\beta(\beta_0, \hat{\gamma})$ converges in distribution to a normal variable with mean zero and covariance matrix that can be consistently estimated by $\hat{\Gamma}(\hat{\beta}, \hat{\gamma})$ given in the Appendix A. This plus the equation

$$n^{1/2}\,(\,\hat{\beta}\,-\,\beta_0\,)\,=\,A_\beta^{-1}\,\{\,n^{-1/2}\,U_\beta(\beta_0,\hat{\gamma})\,\}\,+\,o_p(1)$$

given by the Taylor series expansion of $U_\beta(\hat{\beta}, \hat{\gamma})$ around β_0 shows that the distribution of $n^{1/2}\,(\,\hat{\beta}\,-\,\beta_0\,)$ can be asymptotically approximated by the normal distribution with mean zero and covariance matrix Σ that can be consistently estimated by

$$\hat{\Sigma}\,=\,\hat{A}_\beta(\hat{\beta},\hat{\gamma})^{-1}\,\hat{\Gamma}(\hat{\beta},\hat{\gamma})\,[\hat{A}_\beta(\hat{\beta},\hat{\gamma})^{-1}]'\,.$$

For determination of $\hat{\beta}$ and $\hat{\gamma}$, note that both estimating functions $U_\beta(\beta, \gamma)$ and $U_\gamma(\gamma)$ are similar to the partial likelihood score functions arisen from right-censored failure time data under stratified proportional hazards models or multivariate right-censored failure time data under marginal proportional hazards models. Thus $\hat{\beta}$ and $\hat{\gamma}$ can easily obtained using any statistical software.

4.4 An Extension of Model Setup

In Section 4.2, we assume that the examination times U and V are independent of the survival time T given Z. However, this assumption may not hold in practice. In this section, to address this, we assume that there exists an unobservable random process $b(t)$ that characterizes the dependency between censoring time(s) and failure time, and given the covariate process and process $b(t)$, the examination times U and V and the failure time T are independent. The same idea was used by Zhang, Sun, and Sun (2006) for current status

data. This is a typical type of informative censoring.

Specifically, we assume the following models for U, V, and T:

$$\lambda_i^T(t\,|Z_i(s), b_i(s), s \le t) = \lambda_0(t) + \beta_0' Z_i(t) + b_i(t), \tag{4.6}$$

$$\lambda_i^U(t\,|Z_i(s), b_i(s), s \le t) = \lambda_1(t)\, e^{\gamma_0' Z_i(t) + b_i(t)} \tag{4.7}$$

and

$$\lambda_i^V(t\,|U_i = u_i, Z_i(s), b_i(s), s \le t) \;= \begin{cases} \lambda_2(t) e^{\gamma_0' Z_i(t) + b_i(t)} & \text{if} \quad t \ge u_i \\ 0 & \text{if} \quad t < u_i \end{cases} \tag{4.8}$$

where $b_i(t)$'s are $i.i.d.$ realizations of an unobservable random process $b(t)$, which is assumed to have mean 0. Here, the distribution of the process $b(t)$ is totally unspecified.

It is easy to show that model (4.6) can reduce to an additive hazard model since the survival function can be derived as

$$Pr(T > t | Z(s), s \le t) = E_b(Pr(T > t | Z(s), b)) = E_b(B_i(t)) \exp(-\Lambda(t) - \beta' Z_i^*(t)),$$

where $B_i(t) = \int_0^t b_i(s)ds$ and $Z_i^*(t)$ is defined as above. It is worthy to note that the $\mathbf{E}_b$ term is not subject specific since b_i's are $i.i.d.$ realizations of b.

Let the counting processes $N_i^{(1)}$, $N_i^{(2)}$,$\tilde{N}_i^{(1)}$, and $\tilde{N}_i^{(2)}$ be defined as before. Their intensity functions are

$$I(U_i \ge t)\, \mathbf{E}_b\{e^{-\int_0^t b_i(s)ds}\, e^{b_i(t)}\}\, e^{-\Lambda_0(t)}\, \lambda_1(t)\, e^{-\beta_0' Z_i^*(t) + \gamma_0' Z_i(t)}, \tag{4.9}$$

$$I(u_i < t \le V_i)\, \mathbf{E}_b\{e^{-\int_0^t b_i(s)ds}\, e^{b_i(t)}\}\, e^{-\Lambda_0(t)}\, \lambda_2(t)\, e^{-\beta_0' Z_i^*(t) + \gamma_0' Z_i(t)}, \tag{4.10}$$

$$I(U_i \ge t)\, \mathbf{E}_b\{b_i(t)\}\, \lambda_1(t)\, e^{\gamma_0' Z_i(t)}, \tag{4.11}$$

and

$$I(u_i < t \le V_i)\, \mathbf{E}_b\{b_i(t)\}\, \lambda_2(t)\, e^{\gamma_0' Z_i(t)}, \tag{4.12}$$

respectively, where u_i is the realization of U_i. Notice that none of the $\mathbf{E}_b$ terms is subject spe-

cific. Since none of the baselines and the distribution of $b(t)$ is specified, those nonparametric parts can be put together as one function in each intensity function. Thus, the intensities (4.9)-(4.12) reduce to the corresponding intensity function of counting process $N_i^{(1)}$, $N_i^{(2)}$, $\tilde{N}_i^{(1)}$, and $\tilde{N}_i^{(2)}$ in Sections 4.2 and 4.3.

This suggests that the proposed estimation procedure given in Section 4.4 can be applied here, which indicates that our proposed method is quite robust to such informative censoring.

4.5 Numerical Results

Simulation studies are carried out to assess the finite sample performance of the estimation approach proposed in the previous sections. In the first part, the failure times T_i's are generated from model (4.1) and the censoring times U_i's and V_i's are generated from models (4.2) and (4.3), respectively. For covariates, we assume that there exists only a single covariate Z_i that represents the treatment indicator and follows the Bernoulli distribution with success probability 0.5. For the results presented below, we take the baseline hazard functions $\lambda_0(t)$, $\lambda_1(t)$ and $\lambda_2(t)$ to be constants 1, 2, and 1, respectively. The percentages of left-, interval- and right-censored observations are about 1/3 when $\beta_0 = \gamma_0 = 0$.

Table 4.1 presents the simulation results with $\beta_0 = -0.2$, 0, or 0.2 and $\gamma_0 = -1$ or 1, and $n = 100$ or 200, respectively. For each setup, the results include the bias (Bias) given by the mean of 500 point estimates based on simulated data minus the true value of the parameters, the sample standard error (SSE) of the 500 point estimates, and the average of 500 estimated standard errors based on simulated data (SEE). The 95% empirical coverage probabilities were also calculated and given in the table. It can be seen from Table 4.1 that the proposed estimates of regression parameters appear to be unbiased. The sample standard error and the estimated standard error are quite close, suggesting that the proposed variance estimate is good. As expected, both standard errors decrease as sample

size increases. Also the empirical coverage probabilities seem quite close to the true 95% for the situations considered here.

In the second part of the simulation study, we apply the same procedure to the case of information censoring. The random process is taken to be a random variable $b/4$, where b follows a standard normal distribution. The purpose of taking a small random effect is to make sure the hazard function of failure time to be positive in the additive hazard model. The baseline hazard functions $\lambda_0(t)$, $\lambda_1(t)$ and $\lambda_2(t)$ are set to be 2, 4, and 1, respectively, and β_0 and γ_0 are each taken to be 0, 0.5, and -0.5 making 9 setups in total. The same statistics are calculated as above for 500 replications with sample size $n = 100$. The results are summarized in Table 4.2. Again, the estimates seem unbiased, SSEs and SEEs are quite close, and the 95% coverage probabilities are very close to 0.95. These results suggest that the proposed method works well for the case of informative censoring.

To illustrate the proposed estimation approach, we apply it to the breast cancer data mentioned in Section 1.1.1. As discussed before, the study consists of 94 early breast cancer patients who were given either radiation therapy alone (46) or radiation therapy plus adjuvant chemotherapy (48). During the study, patients were supposed to be seen at clinic visits every 4 to 6 months. However, actual visit times differ from patient from patient and times between visits also vary. At the visits, physicians evaluated the cosmetic appearance of the patient such as breast retraction, a response that has a negative impact on overall cosmetic appearance. The goal of the study is to compare the two treatments with respect to the time to breast retraction, for which only interval-censored are available.

For the analysis, we define $Z_i = 1$ if the patient was given radiation therapy alone and 0 otherwise. For the determination of the U_i's and V_i's, we take U_i and V_i to be left and right end points of the censoring intervals for interval-censored observations. For left-censored observations, we let U_i to be the observation time and V_i the largest time point and for right-censored observations, U_i and V_i are set to be 0 and the observation time,

respectively. The application of the estimation procedure given in the previous section gives $\hat{\beta} = -0.0164$ and $\hat{\gamma} = -0.4261$ with the estimated standard errors being 0.0057 and 0.1637, respectively. This yields a p-value of 0.0041 for testing $\beta_0 = 0$ and suggests that the patients given radiation therapy alone have significantly lower risk to develop breast retraction than those given radiation therapy plus adjuvant chemotherapy. In other words, the adjuvant chemotherapy increases the risk of breast retraction and this result is similar to that given by Finkelstein (1986) using the proportional hazards model. The result also indicates that the observation times U_i's and V_i's seem to have different distributions for patients in the two treatment groups.

Although the result given here is similar to that obtained using the proportional hazards model, sometimes it is of interest to assess which of the two models is more appropriate to a given data set. For this, we obtained the separate maximum likelihood estimators of the survival functions corresponding to the two treatment groups in log and log-log scales, respectively, and presented them in Figure 4.2. Note that under the additive hazards model, the former should give a straight line passing the origin and the latter should give a line parallel to the x-axis. For reference, Figure 4.1 gives the two separate maximum likelihood estimators. Although there is no clear cut, Figures 4.1 and 4.2 suggest that the additive hazards model seems to be more reasonable for the data considered.

4.6 Concluding Remarks

In the preceding sections, regression analysis of general interval-censored failure time data is investigated using the additive hazards model. For estimation of regression parameters, an estimating equation-based approach is presented and asymptotic properties of the proposed estimates are established. Simulation studies suggest that the approach works well for practical situations. A major advantage of the presented method is that it does not involve estimation of the baseline cumulative hazard function $\Lambda_0(t) = \int_0^t \lambda_0(s)\, ds$ and also

it can be easily implemented.

In models (4.2) and (4.3), for simplicity, we assume that the effects of covariates on the observation times U_i's and V_i's are identical. The proposed inference approach can be easily generalized to the situation where the hazard functions of the U_i's and V_i's have the form

$$\lambda_i^U(t \mid Z_i(s), s \le t) \;=\; \lambda_1(t)\, e^{\gamma_1' Z_i(t)}$$

and

$$\lambda_i^V(t \mid U_i = u_i\, Z_i(s), u_i < s \le t) \;=\; \lambda_2(t)\, e^{\gamma_2' Z_i(t)}\,.$$

That is, covariates have different effects on the two observation times. In the models above, γ_1 and γ_2 are p-dimensional vectors of regression parameters representing the covariate effects on the U_i's and V_i's, respectively. Another generalization of the estimation procedure given here that should be useful is to apply some weight functions to $U_\gamma(\gamma)$ and $U_\beta(\beta, \gamma)$. Actually, one could use different weight functions for the two terms in each of the two estimating functions. One advantage for this is to increase the efficiency of resulting estimates of regression parameters.

It should be noted that as that given in Lin et al. (1998) for current status data, the method given in the previous sections may not be the most efficient. In addition to using some weight functions, an alternative is to derive efficient score functions of regression parameters (Martinussen and Scheike, 2002). In this case, one has to estimate $\Lambda_0(t)$ and a lot of more computational effort is needed compared to the method given here as it will be seen in the next chapter. More importantly, for asymptotic normality of the resulting regression parameter estimates, one may have to find an estimate of $\Lambda_0(t)$ that has $n^{1/2}$-convergence rate, which is not easy for given interval-censored data unless some assumptions about $\Lambda_0(t)$ are imposed.

Chapter 5

Efficient Estimation for Bivariate Current Status Data

5.1 Introduction

As discussed in Chapter 1, current status data arise in many fields including epidemiology and biomedical studies. Such data occur when the failure time of interest is never observed but can be determined only to be smaller or larger than a monitoring time. They are also referred to as case 1 interval-censored data (Groeneboom & Wellner 1992).

Regression analysis of univariate current status data has been studied by many researchers. Among others, Huang (1996) explored efficient estimation under the proportional hazards model. Huang and Rossini (1997) studied the same problem under the proportional odds model with the use of sieve method and Martinussen and Scheike (2002) used the additive hazards model.

When multiple failure times are of interest, estimating or testing the dependency of the failures is an important topic. For bivariate current status data, Wang and Ding (2000) proposed a two-stage estimation method under the assumption of a copula model for the joint survival function. They also proposed to test the independency by a 2×2 contingency table (Ding and Wang 2004).

In this chapter, we study the efficient estimation of regression parameter and association parameter simultaneously when bivariate current status data are available. By efficient estimation, we mean that the variance of the estimate reaches the information bound in

a specific model setting. In other words, there does not exist an estimate with a smaller variance under the same model setting.

Denote the two failure times of interest by T_1 and T_2 on each subject and let C denote the common monitoring time for T_1 and T_2. This is called univariate censoring, which is common for bivariate current status. Let $\delta_1 = I(T_1 \leq C)$ and $\delta_2 = I(T_2 \leq C)$ be the censoring indicators. Let X be a $p-$dimensional time independent covariate vector that T_1 and T_2 may depend on. Thus, the data structure is $(C, \delta_1, \delta_2, X)$. Assume there are n independent subjects in the study, then the observed data are $(C_i, \delta_{1i}, \delta_{2i}, X_i)$, for $i = 1, \ldots, n$.

Since we are interested in the dependency of T_1 and T_2, we need to specify their dependency structure. For this purpose, as in Chapter 3, we suppose that (T_1, T_2) follow a copula model with the joint survival function

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

where S_1 and S_2 are the marginal distribution of T_1 and T_2, respectively, $C_\alpha(\cdot\,,\cdot)$ is a mapping from $[0,1]^2$ to $[0,1]$, and α is a global association parameter. For more information about the copula model or Clayton model, please see page 33-34, Chapter 3.

The reminder of this chapter is organized as follows. Section 5.2 introduces the details of models in different cases. In Section 5.3, we derive the efficient score and information bound for the model setups and describe the estimation procedure. Simulation results are presented in Section 5.4 and followed by a real data application in Section 5.5. Section 5.6 gives some concluding remarks.

5.2 Model Setup

We assume that T_1 and T_2 both marginally follow a Cox model, i.e.,

$$S_k(t) = \exp(\, -\Lambda_{0\,k}\,(t)\, \exp(\beta_k' X)), k = 1, 2.$$

Here $\Lambda_{0k}(t)$ are the cumulative baseline hazard functions of T_1 and T_2, respectively. If T_1 and T_2 have the same covariate effects, write $\beta_1 = \beta_2 = \beta$. In this situation, the parameters of interest are $\theta = (\beta',\, \alpha)'$. If $\beta_1 \neq \beta_2$, define $\theta = (\beta_1',\, \beta_2',\, \alpha)'$. For simplicity, we will focus on the situation that the two marginal distributions are same, i.e., common baseline hazard function and common covariate effect.

In the following, we consider continuous censoring variable C and confine C to be in a finite interval $(0, M_0\,]$, where M_0 is a predetermined constant satisfying $P(T_1 \geq M_0) > 0$ and $P(T_2 \geq M_0) > 0$. For real data, τ can be taken to be the largest observation time. It is natural to define $C = \min(C', M_0)$, where C' is a positive random variable that may depend on covariate X. Assume that T_1 and T_2 are independent of C given covariate X. Let $\lambda_c\,(t|X)$ be the hazard function of C given X with the form $\lambda_c(t|X) = \lambda_{c0}\,(t) \exp(\omega' X)$ for $t < M_0$, where ω denotes the possible covariate effect on C. Let $g_c\,(t, X)$ be the joint density of C and X and assume that $g_c\,(t, X)$ is free of the parameter of interest, θ.

Define the following four counting processes

$$N_{11}(t) = \delta_1\, \delta_2\, I(C \leq t), \quad N_{10}(t) = \delta_1(1 - \delta_2)\, I(C \leq t),$$
$$N_{01}(t) = (1 - \delta_1)\, \delta_2\, I(C \leq t), \; N_{00}(t) = (1 - \delta_1)\, (1 - \delta_2)\, I(C \leq t).$$

Similar to the arguments of Lin et al. (1998), the intensity processes for the four processes N_{jm} , for $j = 0, 1$ and $m = 0, 1$, are

$$Y(t)\lambda_c(t|X)S_{jm}(t, \theta),$$

respectively, where $Y(t) = I(C \geq t)$ and

$$\begin{aligned}
S_{11}(\theta, t) &= P(T_1 \leq t, T_2 \leq t) = 1 - S_1(t) - S_2(t) + C_\alpha(S_1(t), S_2(t)), \\
S_{01}(\theta, t) &= P(T_1 > t, T_2 \leq t) = S_1(t) - C_\alpha(S_1(t), S_2(t)), \\
S_{10}(\theta, t) &= P(T_1 \leq t, T_2 > t) = S_2(t) - C_\alpha(S_1(t), S_2(t)), \\
S_{00}(\theta, t) &= P(T_1 > t, T_2 > t) = C_\alpha(S_1(t), S_2(t)).
\end{aligned}$$

Under the copula model assumption, the typical log-likelihood contribution is given by

$$\begin{aligned}
l(\theta) &= \delta_1 \, \delta_2 \, \log(S_{11}(\theta, C)) + (1 - \delta_1) \, \delta_2 \, \log(S_{01}(\theta, C)) \\
&\quad + \delta_1 \, (1 - \delta_2) \, \log(S_{10}(\theta, C)) + (1 - \delta_1) \, (1 - \delta_2) \, \log(S_{00}(\theta, C)) \\
&= \sum_{j=0}^{1} \sum_{m=0}^{1} \int_0^{M_0} \log S_{jm}(\theta, t) \, dN_{jm}(t) \qquad (5.1)
\end{aligned}$$

The original log-likelihood function contains the term $g(C, X)$ but is omitted here since that term does not include the parameters of interest. Note that S_{jm} and $l(\theta)$ are both functions of θ, Λ_{01}, Λ_{02}.

The log-likelihood given in (5.1) contains finite dimensional parameters θ and infinite-dimensional parameters $\Lambda_{01}(t)$ and $\Lambda_{02}(t)$. In this case, projection method can be used to derive the efficient score and information bound of θ for this semiparametric model (Bickel et al. 1993).

For this, consider general parametric submodels of Λ_{01} and Λ_{02} with parameter η, which has the same dimension with θ. Let $\dot{l}_\theta$ and $\dot{l}_\eta$ denote the derivatives of the log-likelihood $l(\theta)$ in (5.1) with respect to θ and η, respectively. The efficient score can be found as the component of $\dot{l}_\theta$ that are orthogonal to the linear span formed by all possible $\dot{l}_\eta$.

Let $\dot{l}^*_\eta$ be the projection of $\dot{l}_\theta$ onto the linear span formed by all possible $\dot{l}_\eta$. Then for all possible $\dot{l}_\eta$, we have

$$E(\dot{l}_\theta - \dot{l}^*_\eta) * \dot{l}_\eta = 0. \qquad (5.2)$$

and the efficient score is $\dot{l}_{\theta^*} = \dot{l}_\theta - \dot{l}^*_\eta$ (Huang, 1996; Martinussen and Scheike, 2002).

5.3 Derivation of Efficient Score and Information Bound

Here, we will focus on the situations of common or different marginal baseline hazard functions with the same covariate effects. If the covariate effects are not same, we can always obtain common regression parameters by redefining covariates.

5.3.1 Common marginal baseline hazard function

We first start with the situation where the two marginal distributions are same, i.e., common baseline hazard function $\Lambda_0(t)$ and common covariate effect β.

Let $\theta_0 = (\beta_0', \alpha_0)'$ be the true value of the parameter θ. In this case, $C_\alpha(S_1(t), S_2(t))$ can be rewritten as $C_\alpha(S_1(t))$. Denote $D_\alpha = \frac{\partial}{\partial \alpha} C_\alpha(S_1)$ and $D_u = \frac{\partial}{\partial u} C_\alpha(u)|_{u=S_1}$.

It is easy to show that

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

for $j = 0, 1$ and $m = 0, 1$.

Define

$$a_{jm} = -\exp(\beta' X)\, S_1\, \Lambda_0\, [(-1)^j\, m + (-1)^m\, j + 2\,(-1)^{m+j}\, D_u]$$

and

$$Z_{jm} = (X', (-1)^{m+j}\, D_u \,/\, a_{jm})'$$

for $j = 0, 1$ and $m = 0, 1$. Then we have $\frac{\partial}{\partial \theta} S_{jm} = Z_{jm} a_{jm}$ for $j, m = 0, 1$.

The score function of θ can be written as

$$\dot{l}_\theta = \sum_{j=0}^{1} \sum_{m=0}^{1} \int \frac{Z_{jm} a_{jm}}{S_{jm}} dN_{jm},$$

where the notation $\int$ stands for $\int_0^{M_0}$ in this chapter if the range is not specified.

Define

$$dM_{jm}(t) = dN_{jm}(t) - Y(t)\lambda_c(t)S_{jm}dt$$

for $j, m = 0, 1$ and it is easy to show that $M_{jm}(t)$ is a martingale for $j, m = 0, 1$.

Observe that $\sum_{j=0}^{1}\sum_{m=0}^{1} S_{jm} = 1$, which implies that $\sum_{j=0}^{1}\sum_{m=0}^{1} Z_{jm}a_{jm} = 0$. Based on this fact, we have

$$\dot{l}_\theta = \sum_{j=0}^{1}\sum_{m=0}^{1}\int \frac{Z_{jm}\, a_{jm}}{S_{jm}} dN_{jm} = \sum_{j=0}^{1}\sum_{m=0}^{1}\int \frac{Z_{jm}\, a_{jm}}{S_{jm}} dM_{jm}.$$

Suppose $\frac{\partial}{\partial\eta}\log\Lambda_0(t) = b(t)$, then the score function for η is of the following form:

$$\dot{l}_\eta(b) = \sum_{j=0}^{1}\sum_{m=0}^{1}\int \frac{b\, a_{jm}}{S_{jm}}\, dN_{jm} = \sum_{j=0}^{1}\sum_{m=0}^{1}\int \frac{b\, a_{jm}}{S_{jm}}\, dM_{jm}.$$

Using the projection method, to derive the efficient score function for θ, we need only to find a function b^* satisfying (5.2) for any b. This yields that for any $t\epsilon(0,\, M_0]$,

$$\begin{aligned} 0 &= E\,(\dot{l}_\theta - \dot{l}^*_\eta)\,\dot{l}_\eta \\ &= \sum_{j=0}^{1}\sum_{m=0}^{1} E\Big\{\int_0^t \frac{(Z_{jm} - b^*)\, a_{jm}}{S_{jm}}\, dM_{jm} \int_0^t \frac{b\, a_{jm}}{S_{jm}}\, dM_{jm}\Big\} \\ &= \sum_{j=0}^{1}\sum_{m=0}^{1} E\Big\{\int_0^t \frac{(Z_{jm} - b^*)\, a_{jm}^2}{S_{jm}}\, b\, Y\, \lambda_c\, ds\Big\}. \end{aligned}$$

for any b. The above equation reduces to

$$\sum_{j=0}^{1}\sum_{m=0}^{1} E\,[\,\frac{(Z_{jm} - b^*)\, a_{jm}^2}{S_{jm}}\, Y\, \lambda_c\,] = 0$$

for any $t\epsilon(0,\, M_0]$. Solving this equation, we can obtain that

$$b^* = \frac{\sum_{j,m} E\,[\,Z_{jm}\, a_{jm}^2\, Y\, \lambda_c\, /\, S_{jm}\,]}{\sum_{j,m} E\,[\,a_{jm}^2\, Y\, \lambda_c\, /\, S_{jm}\,]}\,.$$

Then the efficient score for θ is thus

$$\dot{l}_{\theta^*} = \sum_{j=0}^{1}\sum_{m=0}^{1}\int \frac{(Z_{jm} - b^*)\, a_{jm}}{S_{jm}}\, dN_{jm}\,, \tag{5.3}$$

and the information for θ is given by

$$I(\theta) \;=\; E\,\dot{l}_{\theta^*}^{\otimes 2} = \sum_{j=0}^{1}\sum_{m=0}^{1} E\int \Big((Z_{jm} - b^*)\, a_{jm}\Big)^{\otimes 2}\, S_{jm}^{-1}\, Y\, \lambda_c\, dt, \tag{5.4}$$

where $a^{\otimes 0} = 1$, $a^{\otimes 1} = a$, and $a^{\otimes 2} = a\, a'$. Although not explicitly specified, a_{jm}, Z_{jm}, S_{jm}, and b^* are all functions of θ, Λ_{01}, and Λ_{02} at each time t in the above expressions.

5.3.2 Different marginal baseline hazard functions

In this subsection, we consider the situation where the two marginal distributions are different, i.e., we have different baseline hazard functions but common regression coefficients. Suppose the marginal hazard functions follow Cox model:

$$\lambda_k(t|X) \;=\; \lambda_{0k}(t)\, \exp(\beta' X), \quad (k = 1, 2).$$

The definitions of all the counting processes remain the same, however, some other notations need to change. Denote D_u and D_u the derivatives of $C_\alpha(u, v)$ with respect to u and v, respectively, at $u = S_1$ and $v = S_2$.

For $j = 0, 1$ and $m = 0, 1$, define

$$a_{jm}^{(1)} = -(-1)^j\, \exp(\beta' X)\, \Lambda_{01}\, S_1\, [m + (-1)^j\, D_u],$$

$$a_{jm}^{(2)} = -(-1)^m\, \exp(\beta' X)\, \Lambda_{02}\, S_2\, [j + (-1)^j\, D_v],$$

and $B_{jm} = (a_{jm}^{(1)} + a_{jm}^{(2)})^{-1}\, (-1)^{m+j}\, D_\alpha$. Let $a_{jm} = (a_{jm}^{(1)}, a_{jm}^{(2)})'$ and Z_{jm} be a $(p+1) \times 2$ matrix in the form of

$$Z_{jm} = \begin{pmatrix} X & X \\ B_{jm} & B_{jm} \end{pmatrix}.$$

Then the derivative of S_{jm} with respect to θ is given by $\frac{\partial}{\partial\theta}S_{jm} = Z_{jm} \cdot a_{jm}$.

Suppose $\frac{\partial}{\partial\eta}\log\Lambda_{01}(t) = b_1(t)$ and $\frac{\partial}{\partial\eta}\log\Lambda_{02}(t) = b_2(t)$, where $b_1(t)$ and $b_2(t)$ are both $p+1$ dimensional vectors. It can be shown that the derivative of the S_{jm} with respect to η is $\frac{\partial}{\partial\eta}S_{jm} = b \cdot a_{jm}$, where $b = (b_1, b_2)$.

Using the exactly same procedure as that in the previous subsection, we can obtain the same form of efficient score and information bound for θ but with different definitions of Z_{jm}, a_{jm}, and b^*. In particular, $b^* = H * G^{-1}$, where

$$G = \sum_{j=0}^{1}\sum_{m=0}^{1} E\left[\, a_{jm}^{\otimes 2}\, Y\, \lambda_c\, / S_{jm}\,\right]$$

and

$$H = \sum_{j=0}^{1}\sum_{m=0}^{1} E\left[\, Z_{jm}\, a_{jm}^{\otimes 2}\, Y\, \lambda_c\, / S_{jm}\right].$$

At each time point t, G is a 2×2 matrix and H a $(p+1)\times 2$ matrix.

5.3.3 Estimation procedure

Let a_{jmi}, S_{jmi}, Z_{jmi}, and Y_i be defined as a_{jm}, S_{jm}, Z_{jm} and Y with respect to subject i. We use the empirical version of the above efficient score as the estimating function, i.e.,

$$\dot{l}_{\theta^*}(\Lambda_0,\theta) = \sum_{i=1}^{n}\sum_{j=0}^{1}\sum_{m=0}^{1}\int_0^{M_0}\frac{(X_{jmi} - b_n^*)\, a_{jmi}}{S_{jmi}}\, dN_{jmi}(t)\,,$$

where a_{jmi}, S_{jmi} , Z_{jmi} and b_n^* are all functions of Λ_0, θ, and time t. Then we have

$$b_n^* = \frac{\sum_{i=1}^{n}\sum_{j,m}[Z_{jmi}\, a_{jmi}^2 Y_i\, \exp(X_i'\omega)\,/\,S_{jmi}]}{\sum_{i=1}^{n}\sum_{j,mi}[a_{jmi}^2\, Y_i\, \exp(X_i'\omega)\,/\,S_{jmi}]} \tag{5.5}$$

at each time $t\epsilon(0,\, M_0]$ in the case of common baseline hazard function, and

$$b_n^*(\Lambda_{01},\Lambda_{02},t) = H_n(\Lambda_{01},\Lambda_{02},t) * G_n^{-1}(\Lambda_{01},\Lambda_{02},t)$$

in the case of different baseline hazard functions, where

$$G_n = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{1}\sum_{m=0}^{1}\left[a_{jmi}^{\otimes 2} Y_i \exp(X_i'\omega)\,/S_{jmi}\,\right]$$

and

$$H_n = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{1}\sum_{m=0}^{1}\left[Z_{jmi}\, a_{jmi}^{\otimes 2}\, Y_i \exp(X_i'\omega)\,/\,S_{jmi}\right].$$

Note that the nuisance function $\Lambda_0(t)$ contained implicitly in the estimating function (5.3) is unknown and the censoring variable C is continuous. Thus, the number of jump points in the estimate of Λ_0 is in the order of $O(n)$. If there exists an estimate $\hat{\Lambda}_0$ of Λ_0 or estimates of Λ_{01} and Λ_{02} with $n^{1/3}$ convergence rate, then $\hat{\theta}$ defined as the root of equation $\dot{l}_{\theta^*}(\hat{\Lambda}_0, \theta) = 0$ or $\dot{l}_{\theta^*}(\hat{\Lambda}_{01}\,\hat{\Lambda}_{02}, \theta) = 0$ is efficient and its variance can be consistently estimated by $\hat{I}(\hat{\theta})^{-1}$, where

$$\hat{I}(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{1}\sum_{m=0}^{1}\int\Big(\frac{(\hat{Z}_{jmi} - \hat{b}_n^*)\hat{a}_{jmi}}{\hat{S}_{jmi}}\Big)^{\otimes 2} dN_{jmi}. \tag{5.6}$$

Here, the notation $\hat{f}$ means $f(\hat{\Lambda}_0)$ if f is a function of Λ_0. The proof is attached in Appendix C.

It is proved by Huang (1996) that under the proportional hazards model the nonparametric maximum likelihood estimate (NPMLE) of the cumulative baseline hazard has a convergence rate $n^{1/3}$ for univariate current status data with the existence of covariate. However, there is no close form for the NPMLE. Profile likelihood approach can be used to obtain the NPMLE for computation purpose. However, the limiting distribution of the estimate obtained through profile likelihood approach is still not clear yet even though it is consistent (Huang and Wellner 1996).

For this, in the following, we use the sieve method to obtain a consistent estimator of Λ_0. The idea is to approximate $\Lambda_0(t)$ by a step function with k jump points, where $k = n^{\epsilon}$. As the number of steps increases along with the sample size, the bias from approximation

tends to disappear. Such method was also applied in, for example, Rossini and Tsiatis (1996) and Huang and Ross(1997) for current status data under the proportional odds model. In this way, we only need to estimate k parameters for Λ_0.

In the following, for simplicity, we focus on the estimation procedure for the case of common marginal distributions and some minor changes are needed for the case of different baseline hazard functions in the estimation procedure.

The first step is to obtain a consistent estimator of the cumulative baseline hazard function Λ_0 in terms of k parameters. Marginal approach is used for this purpose. That is, considering the product of marginal likelihoods as the full likelihood by assuming that the two events are independent (Goggins and Finkelstein 2000; Kim and Xue 2002). Denote the maximizer of the product likelihood by $(\hat{\Lambda}, \tilde{\beta})$. One can use $\tilde{\beta}$ as the initial value for β in the third step below.

Second, calculate $\hat{b}_n^*$ at each observed C using the formula (5.5). In this part,the censoring effect ω can be estimated by using the partial likelihood based on right-censored data on C. If the censoring variable C is known to be independent of covariate X, then $\omega = 0$. Then there is no need to estimate ω.

Third, solve the efficient estimating equation $\dot{l}_{\theta^*}(\hat{\Lambda}_0, \theta) = 0$ to obtain $\hat{\theta}$ and calculate the variance estimator $\hat{I}(\hat{\theta})^{-1}$ through (5.6).

When the sieve method is used to approximate the unknown function Λ_0, there are only finite ($k + p + 1$ for the case of common marginal distribution) parameters to be estimated. In this case, the full likelihood method can also be used and is efficient under the assumption that the step function with k jump points is the true baseline. In this situation, the $k+p+1$ dimensional parameter $(\Lambda_0', \beta', \alpha)'$ can be estimated simultaneously. In the next section, this approach is used as a benchmark for evaluating the proposed method, and we expect that our proposed method works as well as the full likelihood method since they are both efficient.

5.4 Simulation Study

In this simulation, 100 subjects are randomly selected to be assigned to treatment and control groups with equal probability. The two failure times for each subject are generalized from Clayton-Oakes model $C_\alpha(u,v) = (u^{1-\alpha}+v^{1-\alpha}-1)^{1/(1-\alpha)}$ with the association parameter α being 2 and 3. For the marginal distribution of the two failure times, the baseline hazard function is taken to be 1 and the true treatment effect is set to be 0 and 1 .The censoring effect ω is taken to be 0 and 1 in the simulation. Since Kendall's τ has a more clear interpretation of association than α, we summarize the estimate of τ in the following simulation instead of $\hat{\alpha}$. For Clayton model, the relationship $\tau = (\alpha-1)/(\alpha+1)$ can be used to obtain $\hat{\tau}$ through $\hat{\alpha}$ and the standard error of $\hat{\tau}$ can be obtained by using the delta method.

As discussed above, the sieve method is used by approximating the baseline hazard function with a step function. The number of jump points is taken to be $n^{1/3}$ or $k = n^{2/5}$. Jump points are taken to allow approximately equal number of observations within each interval. Both the proposed method and full likelihood method under this situation are carried out with 1000 replication. In each of the methods, we calculate the bias, sample standard error (SSE), average of standard errors (SEE) obtained by using variance formula, and 95% coverage probability (CP) for the estimators of β and τ.

Tables 5.1 and 5.3 show the performance of the proposed method and the full likelihood method when the censoring variable is independent of the covariate, while Tables 5.2 and 5.4 show the results when the censoring variable is dependent of covariate with the true covariate effect $w = 1$ on the censoring variable.

The simulation results show that the proposed method works well in terms of small bias, SSE being close to SEE, and the estimated 95% coverage probability being close to 0.95. Also, the proposed method gives similar results as the full likelihood method under the finite number of parameters, especially the two estimated SEE's are very close, which confirms the fact that the estimate of the proposed method is efficient.

Simulation results are shown in Tables 5.1 and 5.2 when the number of sieve (jump points) is 5, corresponding to $\epsilon = 1/3$, and in Tables 3 and 4 in which the number of sieve is 7, corresponding to $\epsilon = 2/5$. The estimates show smaller biases and smaller variances when $\epsilon = 1/3$ than when $\epsilon = 2/5$. This suggests that increasing the number of sieve may not necessarily lead to better results. One possible reason is that estimation of the baseline function is not easy if there are too many parameters when sample size n is fixed. It is observed that results become poor when 10 jump points are used, especially the estimation of nonzero β (Results are not shown here). Similar performance is also reported in Rossini and Tsiatis (1996).

When sample size is taken to be 200, we observe smaller biases and smaller variances as expected than in the same setting when $n = 100$. The results when $\epsilon = 1/3$ are shown in Tables 5.5 and 5.6.

5.5 A Real Data Application

This section discusses an illustrative example from an animal tumorigenicity experiment conducted by National Toxicology Program (NTP) discussed in Section 1.1.4. It is a 2-year rodent carcinogenicity study of chloroprene consisting of F344/N rats and B6C3F$_1$ mice with both sexes. The experiment was described and summarized in Dunson and Dinse (2002), and contained a control group with no chloroprene and three dose groups with 50 rodents in each group. Rodents in the dose groups were exposed to chloroprene at the concentration of 12.8, 32, and 80 ppm, respectively, 6 hours per day, 5 days per week for up to 2 years. The occurrence of tumor was determined through a pathologic examination when the rodents died. Some rodents died during the study. Those rodents who did not die at the end of the 2-year study were sacrificed regardless of health condition. As in Dunson and Dinse (2002), we will focus on male rats from the control group and the 80mmp dose group and only consider adrenal and lung tumor. Thus, we have bivariate current status data with

univariate censoring and a covariate being the group indicator. Let T_1 and T_2 denote the age of onset of adrenal tumor and lung tumor, respectively, and C denote the death time. Let X denote the group indicator with 1 for the high-dose group and 0 for the control group.

First, we consider if there is any censoring effect between groups, i.e., if there is significant difference in death time between the high-dose and control groups. The death times for the sacrificed rats at the end of study are considered right-censored. Using partial likelihood method, we obtain $\hat{\omega} = 0.4888$ with standard error 0.2227, which suggests that the rats in the high-dose group tend to die earlier than those in the control group.

To check if the two types of tumors have the common baseline hazard function or survival function, we compute the nonparametric maximum likelihood estimators (NPMLE's) of the survival functions of adrenal tumor and lung tumor for both the control and high-dose groups. As it is shown in the Figure 5.1, the NPMLE's are quite different for adrenal tumor and lung tumor and this suggests that the marginal distributions for adrenal tumor and lung tumor are different.

To analyze this data set, the jump points are taken to be 18, 20, 21, 22, 24, and 25 to ensure approximately equal number of death times in each interval. Under the assumption of different baselines and common dose effect, the proposed method gives that the $\hat{\beta} = 0.4861$ with standard error 0.3752 and and $\hat{\tau} = 0.1902$ with standard error 0.2135 . The estimated association is positive but not significant and the dose effect is not significant for adrenal tumor or lung tumor.

To further analyze the data, we apply the proposed method assuming different baselines and different dose effects and obtain $\hat{\beta}_1 = 0.3450$ with standard error 0.3919 , $\hat{\beta}_2 = 1.2397$ with standard error 0.8474 , and $\hat{\tau} = 0.2373$ with standard error 0.2217 . These results are quite similar to those under the assumption of common dose effect. Thus, we conclude that the dose effect is not significant for either adrenal tumor or lung tumor. Also, it seems that there does not exist significant association between the occurrences of the two types of

tumors.

5.6 Concluding Remarks

In this chapter, we consider the efficient estimation of regression parameter and association parameter for bivariate current status data under the assumption that the joint survival function of the two interested events follow a copula model. We derive the efficient score and information bound for the parameters of interest. The key question to carry out the estimation procedure is to find an uniformly consistent estimate(s) of the unknown baseline hazard function(s) with $n^{1/3}$ convergence rate. For this, we propose to use the sieve method to approximate the baseline hazard function. Simulation results show that the proposed method works well for finite sample sizes.

In the simulation, we observe that the number of sieve seems to affect the inference results. When the number is too small, these parameters can not fully represent the variability of the data. On the contrary, when the number is too large, there are no enough samples to estimate these parameters. Thus, determining the optimal number of sieve is a worthy topic. More research is needed on this part.

As an alternative, when the covariate is a binary variable, one can use nonparametric method to obtain the NPMLE of the baseline survival function and thus the cumulative hazard function in the control group. This estimate converges at $n^{1/3}$ rate. However, doing that may lose efficiency since only the subjects in the control group are used.

In this chapter, to specify the dependence structure, we assume that the joint survival functions of the two events follow a copula model. However, our proposed method is not confined within the copula model. Any form of bivariate survival function can be taken and our proposed method remains valid as long as the marginal distributions are in the form of the proportional hazards model.

Chapter 6

Future Research

In this chapter, we briefly discuss some issues that are related to the research presented in Chapters 2, 3, 4, and 5 and worth more research in the future. In the following, we only consider case 2 interval-censored data.

6.1 Regression Analysis of Multivariate Interval-censored Data

As mentioned in Chapter 1, both marginal approach and random effect approach can be used to deal with regression analysis of multivariate interval-censored data. Surprisingly, only the proportional hazards model has been considered for the marginal distributions in these approaches (Goggins and Finkelstein, 2000; Kim and Xue, 2002).

Notice that many other models have been applied to univariate interval-censored data during recent years, such as the accelerated failure time model (Rabinowitz et al. 1995; Betensky et al. 2001) and the proportional odds model (Huang and Rossini 1997; Rabinowitz et al. 2000; Zhang et al. 2005). We would like to explore the possibility to generalize their approaches to regression analysis of multivariate interval-censored data. Both marginal approach and random effect approach will be considered in the generalization.

No matter what models are used for multivariate situation, a nature problem is whether the data support the models assumed. A solution to this is to consider model checking, which is another difficult topic. The approach we have proposed in Chapter 2 may be applied to check other marginal models.

6.2 The Estimation of Association Parameter when Covariates Exist

Estimating the dependency or association of correlated events is an important topic in survival analysis. We studied the estimation of association parameter for bivariate case 2 interval-censored data without covariate in Chapter 2. In Chapter 5, we considered the efficient estimation of regression parameter and association parameter simultaneously for bivariate current status data. Further interesting research questions include: Does the two-stage estimation procedure described in Chapter 2 still work when there are covariates available for case 2 interval-censored data? How can we obtain efficient estimation of the association parameter when only case 2 interval censored data are available?

For estimation of association parameter, most papers assume the copula model for the dependence structure. These methods may be sensitive to this model assumption. It is helpful to study how to check this assumption when interval-censored data are available. Test statistics can be constructed based on the martingale residuals if proper counting processes can be defined.

BIBLIOGRAPHY

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.

Bebchuk, J. D. and Betensky, R. A. (2000). Multiple imputation for simple estimation of the hazard function based on interval-censored data. *Statistics in Medicine* 19, 405-419.

Betensky, R. A. and Finkelstein, D. M. (1999). An extension of Kendall's coefficient of concordance to bivariate interval censored data, *Statistics in Medicine* 18, 3101-3109.

Betensky, R. A. Lindsey, J. C., Ryan, L. M. and Wand, M. P. (2002). A local likelihood proportional hazards model for interval-censored data. *Statistics in Medicine* 21, 263-275.

Betensky, R. A., Rabinowitz, D. and Tsiatis, A. A. (2001). Computationally simple accelerated failure time regression for interval censored data. *Biometrika* 88, 703-711.

Bogaerts, K., Leroy, R. Lessaffre, E. and Declerck, D. (2002). Computationally simple accelerated failure time regression for interval censored data. *Statistics in Medicine* 21, 3775-3787.

Cai, J. (1999). Hypothesis testing of hazard ratio parameters in marginal models for multivariate failure time data. *Lifetime Data Analysis* 5, 39-53.

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

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. (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.

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

Finkelstein, D. M., Goggins, W. B. and Schoenfeld, D. A. (2002). Analysis of failure time data with dependent interval censoring. *Biometrics* 58, 298-304.

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

Foutz, R. V. (1977). On the unique consistent solution to the likelihood equations. *J. Amer. Statist. Assoc.* 72, 147-148.

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

Genest, C. and Rivest, L.-P. (1993). Statistical inference procedures for bivariate Archimedean copulas. *J. Amer. Statist. Assoc.* 88, 1034-1043.

Gentlemen, R. and Geyer, C. J. (1994). Maximum likelihood for interval censored data: consistency and computation. *Biometrika* 81, 618-623.

Geskus, R. and Groeneboom, P. (1999). Asymptotically optimal estimation of smooth functionals for interval censoring, case 2. *The Annals of Statistics* 27, 627-674.

Goedert, J., Kessler, C., Adedort, L. and et al. (1989). A prospective-study of human immunodeficiency virus type-1 infection and the development of AIDS in sujects with hemophilia. *New England Journal of Medicine* 321, 1141-1148.

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. (1996). Lectures on inverse problems. In *Lecture Notes in Math.,* 1648. Springer-Verlag, Berlin.

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

Hougaard, P. (1986). A class of multivariate failure time distributions. *Biometrika* 73, 671-678.

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

Hsu, L. and Prentice, R. L. (1996). On assessing the strength of dependency between failure time variates. *Biometrika* 83, 491-506.

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

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

Huang, J. and Wellner, J. A. (1996). Interval censored survival data: a review of recent progrss. *Technical Report 310* Department of Statistics, University of Washington, Seattle.

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

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

Kelly, P. J. and Lim, L. L-Y. (2000). Survival analysis for recurrent event data: an application to childhood infectious diseases. *Statist. Med.*, 19, 13-33.

Kim, M. Y. and Xue, X. (2002). The analysis of multivariate interval-censored survival data. *Statistics in Medicine* 21, 3715-3726.

Klein, J. P. and Moeschberger, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data.* Springer.

Kooperberg, C. and Stone, C. J. (1992) Cox Logspline density estimation for censored data. *J. Comput. and Graph. Stat.*, 1, 301-328.

Kroner, B., Rosenberg, P., Adedort, L., Alvord, W, and Goedert, J. (1994). HIV-1 infection incidence among people with hemophilia in the United States and Western Europe, 1978-1990. *Journal of Acquired Immune Deficiency Syndromes* 7,279-286.

Lee A. J. (1990). *U-statistics : Theory and Practice.* Marcel Dekker, Inc., New York.

Lin, D. Y. (1994). Cox regression analysis of multivariate failure time data: the marginal approach. *Statistcs in Medicine*, 13, 2233-2247.

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

Lin, D. Y., Oakes, D. and Ying, Z. (1998). Additive hazards regression with current status data. *Biometrika*, 85, 289-298.

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

Pan, W. (2000). A multiple imputation approach to Cox regression with interval-censored data. *Biometrics*, 56, 199-203.

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

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.

Rabinowitz, D., Tsiatis, A. A. and Aragon J. (1995). Regression with interval-censored data. *Biometrika*, 82, 501-513.

Rossini, A. and Tsiatis, A. A. (1996). A semiparametric proportional odds regression model for the analysis of current status data. *J. Am. Statist. Assoc.*, 91, 713-721.

Rosenberg, P. S. (1995). Hazard function estimation using B-splines. *Biometrics*, 51, 874-887.

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

Satten, G. A. (1996). Rank-based inference in the proportional hazards model for interval-censored data. *Biometrika*, 83, 355-370.

Satten, G. A., Datta, S. and Williamson, J. M. (1998). Inference based on imputated failure times for the proportional hazards model with interval-censored data. *J. Am. Statist. Assoc.*, 93, 318-327.

Shih, J. H. and Louis, T. A. (1995). Inference on the association parameter in copula models for bivariate survival data. *Biometrics* 51, 1384-1399.

Spiekerman, C. F. and Lin, D. Y. (1996). Checking the marginal Cox model for correlated failure time data. *Biometrika* 83, 143-156.

Sun, J. (2005). Interval Censoring. *Encyclopedia of Biostatistics*, John Wiley & Sons Ltd., Second Edition, 2603-2609.

Tian, L. and Cai, T. (2004). On the Accelerated Failure Time Model for Current Status and Interval Censored Data. *Technical report 14*, Department of Biostatistics, Harvard University, Boston.

Turnbull, B. W. (1976). The empirical Distribution with Arbitrarily Grouped Censored and Truncated Data. *Journal of the Royal Statistical Society* B 38, 290-295.

Van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes with Applications to Statistics*. Springer-Verlag, New York.

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

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

Wei, L. J. (1984). Testing goodness of fit for proportional hazards model with censored observations. *J. Amer. Statist. Assoc.* 79, 649-652.

Wei, L. J., Lin, D. Y. and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *J. Am. Statist. Assoc.*, 84, 1065-1073.

Yue. H. and Chan, K. S. (1997). A dynamic frailty model for multivariate survival data. *Biometrics*, 53, 785-793.

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

Zhang, Z.G., Sun, L., Zhao, X. Q. and Sun, J. (2005). Regression analysis of interval-censored failure time data with linear transformation models. *The Canadian Journal of Statistics* 33, 1, 61-70.

Appendix

Appendix A: Proof of asymptotic properties of $\hat{\alpha}$ in Chapter 3

We will use the same notation defined in the previous sections and assume that the following regularity conditions hold:

(C1) S_r and H_r $(r = 1, 2)$ satisfy model conditions (M1)–(M3) and distribution conditions (D1)–(D4) of Geskus and Groeneboom (1999), where H_r is the joint distribution of $(U^{(r)}, V^{(r)})$.

(C2) $\Psi_r(t)$ $(r = 1, 2)$ satisfies functional conditions (F1)–(F3) of Geskus and Groeneboom (1999).

(C3) $\partial^3 l(\alpha, S_1, S_2, \boldsymbol{x}, \delta)/\partial\alpha^3$, $\partial^3 l(\alpha, S_1, S_2, \boldsymbol{x}, \delta)/\partial\alpha^2\partial S_1$ and $\partial^3 l(\alpha, S_1, S_2, \boldsymbol{x}, \delta)/\partial\alpha^2\partial S_2$ are continuous and bounded on the support of G_α for $\alpha \in \mathcal{N}(\alpha_0)$, where $\mathcal{N}(\alpha_0)$ is a compact neighbourhood of α_0.

(C4) $A(\alpha_0, S_1, S_2, G_{\alpha_0}) = -\sum_\delta \int \left[\partial l(\alpha_0, S_1, S_2, \boldsymbol{x}, \delta)/\partial\alpha\right]^2 g_{\alpha_0}(\boldsymbol{x}, \delta) d\boldsymbol{x}$ is negative, where the summation is over all possible δ.

First, we show the consistency of $\hat{\alpha}$.

It follows from (C3) that for all $\alpha \in \mathcal{N}(\alpha_0)$,

$$\begin{aligned}
&\left|U(\alpha, \hat{S}_1, \hat{S}_2, G_n) - U(\alpha, S_1, S_2, G_n)\right| \\
&\quad \leq \int \left|\frac{\partial}{\partial\alpha} l(\alpha, \hat{S}_1, \hat{S}_2, \boldsymbol{x}, \delta) - \frac{\partial}{\partial\alpha} l(\alpha, S_1, S_2, \boldsymbol{x}, \delta)\right| dG_n(\boldsymbol{x}, \delta)
\end{aligned}$$

$$\leq \int \left| \frac{\partial}{\partial \alpha} l(\alpha, \hat{S}_1, \hat{S}_2, \boldsymbol{x}, \delta) - \frac{\partial}{\partial \alpha} l(\alpha, S_1, \hat{S}_2, \boldsymbol{x}, \delta) \right| dG_n(\boldsymbol{x}, \delta)$$

$$+ \int \left| \frac{\partial}{\partial \alpha} l(\alpha, S_1, \hat{S}_2, \boldsymbol{x}, \delta) - \frac{\partial}{\partial \alpha} l(\alpha, S_1, S_2, \boldsymbol{x}, \delta) \right| dG_n(\boldsymbol{x}, \delta)$$

$$\leq M_0 \left(\sup_{0 \leq t \leq \tau_1} |\hat{S}_1(t) - S_1(t)| + \sup_{0 \leq t \leq \tau_2} |\hat{S}_2(t) - S_2(t)| \right),$$

where M_0 is some constant and $[0, \tau_1] \times [0, \tau_2]$ is the bounded support of $S(t_1, t_2)$. Hence it follows from $\sup_{0 \leq t \leq \tau_r} |\hat{S}_r(t) - S_r(t)| \to_p 0$ (Groeneboom and Wellner, 1992) that

$$\left| U(\alpha, \hat{S}_1, \hat{S}_2, G_n) - U(\alpha, S_1, S_2, G_n) \right| \to_p 0 \, . \tag{A.1}$$

Furthermore, using the Glivenko-Cantelli theorem and the Dominated Convergence Theorem, we have that for all $\alpha \in \mathcal{N}(\alpha_0)$,

$$U(\alpha, S_1, S_2, G_n) \to_p \mathcal{U}(\alpha), \tag{A.2}$$

where $\mathcal{U}(\alpha) = E\{U(\alpha, S_1, S_2, G_n)\}$. Note that $U(\hat{\alpha}, \hat{S}_1, \hat{S}_2, G_n) = \mathcal{U}(\alpha_0) = 0$. Then it follows from (A.1), (A.2), (C.4) and the inverse function theorem (Foutz, 1977) that $\hat{\alpha}$ is a consistent estimate of α_0 (e.g. Hsu and Prentice, 1996).

Now we show asymptotical normality of $\hat{\alpha}$.

First note that (C3) and the application of Taylor series expansion to $U(\hat{\alpha}, \hat{S}_1, \hat{S}_2, G_n)$ at $\alpha = \alpha_0$ yield

$$U(\hat{\alpha}, \hat{S}_1, \hat{S}_2, G_n) - U(\alpha_0, \hat{S}_1, \hat{S}_2, G_n) = A(\alpha_0, \hat{S}_1, \hat{S}_2, G_n)(\hat{\alpha} - \alpha_0) + O_p(|\hat{\alpha} - \alpha_0|^2). \tag{A.3}$$

Similarly to (A.1) and (A2), we have

$$A(\alpha_0, \hat{S}_1, \hat{S}_2, G_n) \to_p A(\alpha_0, S_1, S_2, G_{\alpha_0}) < 0 \tag{A.4}$$

by (C4). Based on (A.3) and (A.4), to prove Theorem 3.1, it is sufficient to show that

$n^{1/2}U(\alpha_0, \hat{S}_1, \hat{S}_2, G_n)$ is asymptotically normally distributed with mean zero and variance $Var\{B(\alpha_0, S_1, S_2, U^{(1)}, V^{(1)}, U^{(2)}, V^{(2)}, \Delta)\}$.

To investigate $n^{1/2}U(\alpha_0, \hat{S}_1, \hat{S}_2, G_n)$, note that we can rewrite it as

$$n^{1/2}U(\alpha_0, \hat{S}_1, \hat{S}_2, G_n) = R_{n1} + R_{n2} + R_{n3}\,, \tag{A.5}$$

where

$$R_{n1} = n^{1/2} \int \frac{\partial}{\partial \alpha} l(\alpha_0, S_1, S_2, \boldsymbol{x}, \delta) dG_n(\boldsymbol{x}, \delta),$$

$$R_{n2} = n^{1/2} \int \left[\frac{\partial}{\partial \alpha} l(\alpha_0, \hat{S}_1, \hat{S}_2, \boldsymbol{x}, \delta) - \frac{\partial}{\partial \alpha} l(\alpha_0, S_1, S_2, \boldsymbol{x}, \delta) \right] dG_{\alpha_0}(\boldsymbol{x}, \delta)$$

and

$$R_{n3} = n^{1/2} \int \left[\frac{\partial}{\partial \alpha} l(\alpha_0, \hat{S}_1, \hat{S}_2, \boldsymbol{x}, \delta) - \frac{\partial}{\partial \alpha} l(\alpha_0, S_1, S_2, \boldsymbol{x}, \delta) \right] d[G_n(\boldsymbol{x}, \delta) - G_{\alpha_0}(\boldsymbol{x}, \delta)]\,.$$

It can be checked that the first term R_{n1} is a sum of i.i.d. variables with mean zero:

$$R_{n1} = n^{-1/2} \sum_{i=1}^{n} \frac{\partial}{\partial \alpha} l(\alpha_0, S_1, S_2, U_i^{(1)}, V_i^{(1)}, U_i^{(2)}, V_i^{(2)}, \Delta_i).$$

For R_{n3}, note that $\sup_{0 \le t \le \tau_r} |\hat{S}_r(t) - S_r(t)| \to_p 0$ (Groeneboom and Wellner, 1992), $\sup_x |G_n(x) - G_{\alpha_0}(x)| = O_p(n^{-1/2})$ and $\partial l(\alpha_0, \hat{S}_1, \hat{S}_2, \boldsymbol{x}, \delta)/\partial \alpha$ is continuous and bounded. It then follows from the Dominated Convergence Theorem that R_{n3} converges to zero in probability. Now look at R_{n2}. Under (C3) (e.g. Wang and Ding, 2000), an application of von Mises expansions to R_{n2} gives

$$R_{n2} = n^{1/2} \int_0^{\tau_1} \Psi_1(t) d[\hat{S}_1(t) - S_1(t)] + n^{1/2} \int_0^{\tau_2} \Psi_2(t) d[\hat{S}_2(t) - S_2(t)] + o_p(1)\,.$$

Furthermore, it follows from (C1), (C2) and Theorem 3.2 of Geskus and Groeneboom (1999)

that for $r = 1, 2$,

$$\int_0^{\tau_r} \Psi_r(t) d[\hat{S}_r(t) - S_r(t)] = -\int_0^{\tau_r} \Phi_r(u, v, \delta_1^{(r)}, \delta_1^{(r)}) d[G_{rn}(u, v, \delta_1^{(r)}, \delta_2^{(r)}) - G_r(u, v, \delta_1^{(r)}, \delta_2^{(r)})], \tag{A.6}$$

where G_r is the subdistribution function of the observed vector $(U^{(r)}, V^{(r)}, \Delta_1^{(r)}, \Delta_2^{(r)})$ and G_{rn} is the empirical estimate of G_r. Thus it follows from (A.5) and (A.6) and the fact that $E\Phi_r(U, V, \Delta_1^{(r)}, \Delta_2^{(r)}) = 0$ that

$$n^{1/2} U(\alpha_0, \hat{S}_1, \hat{S}_2, G_n) = n^{-/2} \sum_{i=1}^{n} B(\alpha_0, S_1, S_2, U_i^{(1)}, V_i^{(1)}, U_i^{(2)}, V_i^{(2)}, \Delta_i) + o_p(1)\,.$$

The asymptotic normality of $n^{1/2} U(\alpha_0, \hat{S}_1, \hat{S}_2, G_n)$ then follows from the central limit theorem.

Appendix B: Proof of asymptotic normality of $n^{-1/2}U_\beta(\beta_0,\hat{\gamma})$ in Chapter 4.

Define

$$M_i^{(1)}(t) \;=\; N_i^{(1)}(t) \;-\; \int_0^t I(s \le U_i)\,\lambda_1^*(s)\,e^{-\beta_0' Z_i^*(s)+\gamma_0' Z_i(s)}\,ds\;,$$

$$M_i^{(2)}(t) \;=\; N_i^{(2)}(t) \;-\; \int_0^t I(u_i < s \le V_i)\,\lambda_2^*(s)\,e^{-\beta_0' Z_i^*(s)+\gamma_0' Z_i(s)}\,ds\;,$$

$$\tilde{M}_i^{(1)}(t) \;=\; \tilde{N}_i^{(1)}(t) \;-\; \int_0^t I(s \le U_i)\,\lambda_1(s)\,e^{\gamma_0' Z_i(s)}\,ds$$

and

$$\tilde{M}_{2i}^{(2)}(t) \;=\; \tilde{N}_i^{(2)}(t) \;-\; \int_0^t I(u_i < s \le V_i)\,\lambda_2(s)\,e^{\gamma_0' Z_i(s)}\,ds\;,$$

$i = 1,...,n$. Then $M_i^{(1)}(t)$, and $\tilde{M}_i^{(1)}(t)$ are martingales starting at 0, and $M_i^{(2)}(t)$ and $\tilde{M}_i^{(2)}(t)$ are martingales starting at u_i.

Also define

$$A_1 \equiv E\left(\int_0^\infty \{Z_1^*(t) - \frac{s_{1,\beta}^{(1)}(t,\beta_0,\gamma_0)}{s_{1,\beta}^{(0)}(t,\beta_0,\gamma_0)}\}^{\otimes 2} I(U_1 \ge t)\lambda_1^*(t)\,e^{-\beta_0' Z_1^*(t)+\gamma_0' Z_1(t)}\,dt\right),$$

$$A_2 \equiv E\left(\int_0^\infty \{Z_1^*(t) - \frac{s_{2,\beta}^{(1)}(t,\beta_0,\gamma_0)}{s_{2,\beta}^{(0)}(t,\beta_0,\gamma_0)}\}^{\otimes 2} I(U_1 < t \le V_1)\lambda_2^*(t)\,e^{-\beta_0' Z_1^*(t)+\gamma_0' Z_1(t)}\,dt\right),$$

$$\tilde{A}_1 \equiv E\left(\int_0^\infty \{Z_1^(t) - \frac{s_{1,\gamma}^{(1)}(t,\gamma_0)}{s_{1,\gamma}^{(0)}(t,\gamma_0)}\}^{\otimes 2} I(U_1 \ge t)\lambda_1(t)\,e^{\gamma_0' Z_1(t)}\,dt\right),$$

and

$$\tilde{A}_2 \equiv E\left(\int_0^\infty \{Z_1(t) - \frac{s_{2,\gamma}^{(1)}(t,\gamma_0)}{s_{2,\beta}^{(0)}(t,\gamma_0)}\}^{\otimes 2} I(U_1 < t \le V_1)\lambda_2(t)\,e^{\gamma_0' Z_1(t)}\,dt\right),$$

where $s_{l,\gamma}^{(j)}(t,\gamma)$ and $s_{l,\beta}^{(j)}(t,\beta,\gamma)$ denote the limits of $S_{l,\gamma}^{(j)}(t,\gamma)$ and $S_{l,\beta}^{(j)}(t,\beta,\gamma)$, respectively, $l = 1,2$, $j = 0,1$. Let $A_\gamma = A_1 + A_2$ and $B = \tilde{A}_1 + \tilde{A}_2$ and assume that both A_γ and B are positive definite. Also let $\hat{A}_\gamma(\beta,\gamma) = n^{-1}\,\partial U_\beta(\beta,\gamma)/\partial\gamma$ and $\hat{B}(\gamma) = -n^{-1}\,\partial U_\gamma(\gamma)/\partial\gamma$. Then A_γ and B are the limits of $\hat{A}_\gamma(\beta,\gamma)$ and $\hat{B}(\gamma)$ at β_0 and γ_0, respectively.

To investigate the asymptotic normality of $n^{-1/2}\,U_\beta(\beta_0,\hat{\gamma})$, first note that using the

Taylor series expansions of $U_\beta(\beta_0, \hat{\gamma})$ and $U_\gamma(\hat{\gamma})$ around γ_0, we have

$$n^{-1/2}\, U_\beta(\beta_0, \hat{\gamma}) \;=\; n^{-1/2}\, U_\beta(\beta_0, \gamma_0) \;+\; A_\gamma\, B^{-1}\, \{\, n^{-1/2}\, U_\gamma(\gamma_0)\,\} \;+\; o_p(1)\;.$$

Furthermore, following Lin et al. (1998), it can be shown that

$$n^{-1/2}\, U_\beta(\beta_0, \gamma_0) \;=\; n^{-1/2} \sum_{i=1}^{n} \{\, a_{1i}(\beta_0, \gamma_0) \;+\; a_{2i}(\beta_0, \gamma_0)\,\} \;+\; o_p(1)$$

and

$$n^{-1/2}\, U_\gamma(\gamma_0) \;=\; n^{-1/2} \sum_{i=1}^{n} \{\, b_{1i}(\gamma_0) \;+\; b_{2i}(\gamma_0)\,\} \;+\; o_p(1)\;,$$

where

$$a_{1i}(\beta, \gamma) \;=\; \int_0^\infty \left\{ Z_i^*(t) - \frac{s_{1,\beta}^{(1)}(t, \beta, \gamma)}{s_{1,\beta}^{(0)}(t, \beta, \gamma)} \right\} dM_i^{(1)}(t)\;,$$

$$a_{2i}(\beta, \gamma) \;=\; \int_0^\infty \left\{ Z_i^*(t) - \frac{s_{2,\beta}^{(1)}(t, \beta, \gamma)}{s_{2,\beta}^{(0)}(t, \beta, \gamma)} \right\} dM_i^{(2)}(t)\,,$$

$$b_{1i}(\gamma) \;=\; \int_0^\infty \left\{ Z_i(t) \;-\; \frac{s_{1,\gamma}^{(1)}(t, \gamma)}{s_{1,\gamma}^{(0)}(t, \gamma)} \right\} d\tilde{M}_i^{(1)}(t) \;\;,$$

and

$$b_{2i}(\gamma) \;=\; \int_0^\infty \left\{ Z_i(t) \;-\; \frac{s_{2,\gamma}^{(1)}(t, \gamma)}{s_{2,\gamma}^{(0)}(t, \gamma)} \right\} d\tilde{M}_i^{(2)}(t)\;.$$

These give that

$$n^{-1/2}\, U_\beta(\beta_0, \hat{\gamma}) \;=\; n^{-1/2} \sum_{i=1}^{n} \alpha_i(\beta_0, \gamma_0) \;+\; o_p(1)\;,$$

where

$$\alpha_i(\beta, \gamma) \;=\; a_{1i}(\beta, \gamma) + a_{2i}(\beta, \gamma) + A_\gamma\, B^{-1}\, \{\, b_{1i}(\gamma) + b_{2i}(\gamma)\,\}\,.$$

It thus follows from the multivariate central limit theorem or the U-statistic theory (Lee, 1990) that $n^{-1/2}\, U_\beta(\beta_0, \hat{\gamma})$ converges in distribution to a zero-mean normal random vector.

For estimation of the covariance matrix of $n^{-1/2}\, U_\beta(\beta_0, \hat{\gamma})$, define

$$\hat{a}_{1i}(\beta,\gamma) = \int_0^\infty \left\{ Z_i^*(t) - \frac{S_{1,\beta}^{(1)}(t,\beta,\gamma)}{S_{1,\beta}^{(0)}(t,\beta,\gamma)} \right\} \left\{ dN_i^{(1)}(t) - \frac{I(t \le U_i)\, e^{-\beta' Z_i^*(t) + \gamma' Z_i(t)}}{n S_{1,\beta}^{(0)}(t,\beta,\gamma)} dN^{(1)}(t) \right\},$$

$$\hat{a}_{2i}(\beta,\gamma) = \int_{u_i}^\infty \left\{ Z_i^*(t) - \frac{S_{2,\beta}^{(1)}(t,\beta,\gamma)}{S_{2,\beta}^{(0)}(t,\beta,\gamma)} \right\} \left\{ dN_i^{(2)}(t) - \frac{I(u_i < t \le V_i) e^{-\beta' Z_i^*(t) + \gamma' Z_i(t)}}{n S_{2,\beta}^{(0)}(t,\beta,\gamma)} dN^{(2)}(t) \right\},$$

$$\hat{b}_{1i}(\gamma) = \int_0^\infty \left\{ Z_i(t) - \frac{S_{1,\gamma}^{(1)}(t,\gamma)}{S_{1,\gamma}^{(0)}(t,\gamma)} \right\} \left\{ d\tilde{N}_i^{(1)}(t) - \frac{I(t \le U_i) e^{\gamma' Z_i(t)}}{n S_{1,\gamma}^{(0)}(t,\gamma)} d\tilde{N}^{(1)}(t) \right\}$$

and

$$\hat{b}_{2i}(\gamma) = \int_{u_i}^\infty \left\{ Z_i(t) - \frac{S_{2,\gamma}^{(1)}(t,\gamma)}{S_{2,\gamma}^{(0)}(t,\gamma)} \right\} \left\{ d\tilde{N}_i^{(2)}(t) - \frac{I(u_i < t \le V_i)\, e^{\gamma' Z_i(t)}}{n S_{2,\gamma}^{(0)}(t,\gamma)} d\tilde{N}^{(2)}(t) \right\},$$

where $N^{(1)}(t) = \sum_{i=1}^n N_i^{(1)}(t)$, $N^{(2)}(t) = \sum_{i=1}^n N_i^{(2)}(t)$, $\tilde{N}^{(1)}(t) = \sum_{i=1}^n \tilde{N}_i^{(1)}(t)$ and $\tilde{N}^{(2)}(t) = \sum_{i=1}^n \tilde{N}_i^{(2)}(t)$. Then the asymptotic covariance matrix of $n^{-1/2}\, U_\beta(\beta_0, \hat{\gamma})$ can be consistently estimated by $\hat{\Gamma} = \frac{1}{n} \sum_{i=1}^n \hat{\alpha}_i(\hat{\beta},\hat{\gamma})\, \hat{\alpha}_i'(\hat{\beta},\hat{\gamma})$, where

$$\hat{\alpha}_i(\hat{\beta},\hat{\gamma}) = \hat{a}_{1i}(\hat{\beta},\hat{\gamma}) + \hat{a}_{2i}(\hat{\beta},\hat{\gamma}) + \hat{A}_\gamma(\hat{\beta},\hat{\gamma})\, \hat{B}(\hat{\gamma}) \{ \hat{b}_{1i}(\hat{\gamma}) + \hat{b}_{2i}(\hat{\gamma}) \} .$$

Appendix C: Proof of the asymptotical efficiency of $\hat{\theta}$ in Chapter 5.

In this part, we will show that the proposed estimate $\hat{\theta}$ is asymptotically efficient, i.e., the variance of $\hat{\theta}$ reaches the information bound $I^{-1}(\theta_0)$. We only consider the case of different baseline hazard functions since similar but simpler arguments can be applied to the case of common baseline hazard function.

Denote

$$U_n(\theta, \Lambda_{01}, \Lambda_{02}) = \sum_{i=1}^{n} \sum_{j=0}^{1} \sum_{m=0}^{1} \int \frac{(Z_{jmi} - H_n G_n^{-1}) a_{jmi}}{S_{jmi}} \, dN_{jmi},$$

where $Z_{jmi}, a_{jmi}, S_{jmi}, H_n$, and G_n are all functions of θ, Λ_{01}, and Λ_{02}.

Note that $\hat{\theta}$ is defined as the root of equation $U_n(\theta, \hat{\Lambda}_{01}, \hat{\Lambda}_{02}) = 0$, where $\hat{\Lambda}_{01}$ and $\hat{\Lambda}_{02}$ are the estimates of Λ_{01} and Λ_{02}, respectively, with $n^{-1/3}$ convergence rate. We assume that the $I(\theta_0)$ defined in (5.4) is positive definite.

Lemma 5.1.

$$n^{-1/2} U_n(\theta_0, \Lambda_{01}, \Lambda_{02}) \xrightarrow{L} N(0, I(\theta_0)). \tag{C1}$$

Proof.

$$\begin{aligned}
& U_n(\theta_0, \Lambda_{01}, \Lambda_{02}) \\
= & \sum_{i=1}^{n} \sum_{j=0}^{1} \sum_{m=0}^{1} \int \frac{(Z_{jmi} - H_n G_n^{-1}) a_{jmi}}{S_{jmi}} \, dN_{jmi} \\
= & \sum_{i=1}^{n} \sum_{j=0}^{1} \sum_{m=0}^{1} \int \frac{(Z_{jmi} - H_n G_n^{-1}) a_{jmi}}{S_{jmi}} \, dM_{jmi} \\
& + \sum_{i=1}^{n} \sum_{j=0}^{1} \sum_{m=0}^{1} \int (Z_{jmi} - H_n G_n^{-1}) a_{jmi} Y_i \lambda_c dt \\
= & \sum_{i=1}^{n} \sum_{j=0}^{1} \sum_{m=0}^{1} \int \frac{(Z_{jmi} - H_n G_n^{-1}) a_{jmi}}{S_{jmi}} \, dM_{jmi} \\
= & \sum_{i=1}^{n} \sum_{j=0}^{1} \sum_{m=0}^{1} \int \frac{(Z_{jmi} - H G^{-1}) a_{jmi}}{S_{jmi}} \, dM_{jmi} + o_p(n^{1/2}).
\end{aligned}$$

In the above, we use the facts that $\sum_{j=0}^{1}\sum_{m=0}^{1} Z_{jmi}a_{jmi} = 0$ and $\sum_{j=0}^{1}\sum_{m=0}^{1} a_{jmi} = 0$ for each i. The last step holds since H_n and G_n are the empirical version of H and G.

Thus, $U_n(\theta_0, \Lambda_{01}, \Lambda_{02})$ can be approximated by a sum of i.i.d. martingales. By martingale central limit theorem, $n^{-1/2}U_n(\theta_0, \Lambda_{01}, \Lambda_{02})$ converges weekly to a normal variable with mean 0 and variance $I(\theta_0)$.

Lemma 5.2.

$$n^{-1/2}U_n(\theta_0, \hat{\Lambda}_{01}, \hat{\Lambda}_{02}) = n^{-1/2}U_n(\theta_0, \Lambda_{01}, \Lambda_{02}) + o_p(1). \qquad (C2)$$

Proof.

$$\begin{aligned} & n^{-1/2}U(\theta_0, \hat{\Lambda}_{01}, \hat{\Lambda}_{02}) \\ = \; & n^{-1/2}\sum_{i=1}^{n}\sum_{j=0}^{1}\sum_{m=0}^{1}\int \frac{(\hat{Z}_{jmi} - \hat{H}_n\hat{G}_n^{-1})\hat{a}_{jmi}}{\hat{S}_{jmi}} dM_{jmi} \\ & + n^{-1/2}\sum_{i=1}^{n}\sum_{j=0}^{1}\sum_{m=0}^{1}\int \frac{(\hat{Z}_{jmi} - \hat{H}_n\hat{G}_n^{-1})\hat{a}_{jmi}}{\hat{S}_{jmi}} S_{jmi}Y_i\lambda_c dt \\ = \; & V_{1n} + V_{2n}. \end{aligned}$$

First consider V_{1n}. Let

$$Q_{jmi}(\Lambda_{01}, \Lambda_{02}) = (Z_{jmi} - H_nG_n^{-1})a_{jmi}/S_{jmi},$$

$$Q_{jmi}^{(1)} = \frac{\partial}{\partial u}Q_{jmi}(u, v)|_{u=\Lambda_{01}, v=\Lambda_{02}},$$

and

$$Q_{jmi}^{(2)} = \frac{\partial}{\partial v}Q_{jmi}(u, v)|_{u=\Lambda_{01}, v=\Lambda_{02}}.$$

Expanding $Q_{jmi}(\hat{\Lambda}_{01}, \hat{\Lambda}_{02})$ around the true value $(\Lambda_{01}, \Lambda_{02})$ and using the bounded property of the second derivative of Q_{jmi} for any $t \,\epsilon\, [0,\, M_0]$, we have

$$\begin{aligned} Q_{jmi}(\hat{\Lambda}_{01}, \hat{\Lambda}_{02}) \;=\; & Q_{jmi}(\Lambda_{01}, \Lambda_{02}) + Q_{jmi}^{(1)}(\hat{\Lambda}_{01} - \Lambda_{01}) + Q_{jmi}^{(2)}(\hat{\Lambda}_{02} - \Lambda_{02}) \\ & + O(|\hat{\Lambda}_{01} - \Lambda_{01}|^2 + |\hat{\Lambda}_{02} - \Lambda_{02}|^2), \end{aligned}$$

which yields that

$$V_{1n} = n^{-1/2}U_n + n^{-1/2}\sum_{i=1}^{n}\sum_{j=0}^{1}\sum_{m=0}^{1}\int [Q_{jmi}^{(1)}(\hat{\Lambda}_{01} - \Lambda_{01}) + Q_{jmi}^{(2)}(\hat{\Lambda}_{02} - \Lambda_{02})]dM_{jmi} + o_p(1),$$

since $|\hat{\Lambda}_{01} - \Lambda_{01}|^2 + |\hat{\Lambda}_{02} - \Lambda_{02}|^2 = O(n^{-2/3})$. By Lemma A.1 of Lin & Ying (2001), the second term in the above expression is also $o_p(1)$. Thus, we have $V_{1n} = n^{-1/2}U_n + o_p(1)$.

Now consider V_{2n}. Expanding $S_{jmi}(\Lambda_{01}, \Lambda_{02})$ around the estimated value $(\hat{\Lambda}_{01}, \hat{\Lambda}_{02})$ gives that

$$S_{jmi} = \hat{S}_{jmi} + \hat{a}_{jmi}^{(1)}\hat{\Lambda}_{01}^{-1}(\Lambda_{01} - \hat{\Lambda}_{01}) + \hat{a}_{jmi}^{(2)}\hat{\Lambda}_{02}^{-1}(\Lambda_{02} - \hat{\Lambda}_{02}) + O(|\hat{\Lambda}_{01} - \Lambda_{01}|^2 + |\hat{\Lambda}_{02} - \Lambda_{02}|^2).$$

Since $\hat{a}_{jmi} = (\hat{a}_{jmi}^{(1)}, \hat{a}_{jmi}^{(2)})'$ and $\sum_{j,m}\hat{a}_{jmi}^{(k)} = 0$ for $k = 1, 2$, we have

$$\begin{aligned} V_{2n} &= n^{-1/2}\sum_{i=1}^{n}\sum_{j=0}^{1}\sum_{m=0}^{1}\sum_{k=1}^{2}\int \hat{Q}_{jmi}\left[\hat{a}_{jmi}^{(k)}\hat{\Lambda}_{0k}^{-1}(\Lambda_{0k} - \hat{\Lambda}_{0k})\right]Y_i\lambda_c dt + o_p(1) \\ &= n^{-1/2}\int \sum_{k=1}^{2}\hat{\Lambda}_{0k}^{-1}(\Lambda_{0k} - \hat{\Lambda}_{0k})\sum_{i=1}^{n}\sum_{j=0}^{1}\sum_{m=0}^{1}\hat{Q}_{jmi}\hat{a}_{jmi}^{(k)}Y_i\lambda_c dt + o_p(1). \end{aligned}$$

Note that $\sum_{i=1}^{n}\sum_{j=0}^{1}\sum_{m=0}^{1}\hat{Q}_{jmi}\hat{a}_{jmki}^{(k)}Y_i\lambda_c = 0$ by checking the formulation of $\hat{H}_n$ and $\hat{G}_n$. Thus, $V_{2n} = o_p(1)$.

Lemma 5.3.

$$n^{-1}\frac{\partial}{\partial\theta}U_n(\theta, \hat{\Lambda}_{01}, \hat{\Lambda}_{02})|_{\theta=\theta_0} = n^{-1}\frac{\partial}{\partial\theta}U_n(\theta, \Lambda_{01}, \Lambda_{02})|_{\theta=\theta_0} + o_p(1), \qquad (C3)$$

and

$$n^{-1}\frac{\partial}{\partial\theta}U_n(\theta, \Lambda_{01}, \Lambda_{02})|_{\theta=\theta_0} \xrightarrow{p} -I(\theta_0). \qquad (C4)$$

Proof. (C3) can be proved similarly as the arguments for (C2). To show (C4), we only need to show that

$$n^{-1}\frac{\partial}{\partial\theta}U_n(\theta, \Lambda_{01}, \Lambda_{02})|_{\theta=\theta_0} \xrightarrow{p} -I(\theta_0)$$

based on (C3).

$$
\begin{aligned}
& n^{-1}\frac{\partial}{\partial\theta}U_n(\theta,\Lambda_{01},\Lambda_{02})|_{\theta=\theta_0} \\
= & -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{1}\sum_{m=0}^{1}\int\frac{(Z_{jmi}-H_nG_n^{-1})a_{jmi}}{S_{jmi}^2}(Z_{jmi}a_{jmi})'dN_{jmi} \\
& +\frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{1}\sum_{m=0}^{1}\int[\frac{\partial}{\partial\theta}(Z_{jmi}-H_nG_n^{-1})a_{jmi}]S_{jmi}^{-1}dN_{jmi} \\
= & -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{1}\sum_{m=0}^{1}\int\frac{(Z_{jmi}-H_nG_n^{-1})a_{jmi}}{S_{jmi}^2}(Z_{jmi}a_{jmi})'dM_{jmi} \\
& -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{1}\sum_{m=0}^{1}\int\frac{(Z_{jmi}-H_nG_n^{-1})a_{jmi}}{S_{jmi}}(Z_{jmi}a_{jmi})'Y_i\lambda_c dt \\
& +\frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{1}\sum_{m=0}^{1}\int[\frac{\partial}{\partial\theta}(Z_{jmi}-H_nG_n^{-1})a_{jmi}]S_{jmi}^{-1}dM_{jmi} \\
& +\frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{1}\sum_{m=0}^{1}\int[\frac{\partial}{\partial\theta}(Z_{jmi}-H_nG_n^{-1})a_{jmi}]Y_i\lambda_c dt \\
= & -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{1}\sum_{m=0}^{1}\int\frac{(Z_{jmi}-H_nG_n^{-1})a_{jmi}}{S_{jmi}}(Z_{jmi}a_{jmi})'Y_i\lambda_c dt+o_p(1) \\
= & -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{1}\sum_{m=0}^{1}\int\left((Z_{jmi}-H_nG_n^{-1})a_{jmi}\right)^{\otimes 2}S_{jmi}^{-1}Y_i\lambda_c dt \\
& -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{1}\sum_{m=0}^{1}\int\frac{(Z_{jmi}-H_nG_n^{-1})a_{jmi}}{S_{jmi}}(H_nG_n^{-1}a_{jmi})'Y_i\lambda_c dt+o_p(1) \\
= & -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{1}\sum_{m=0}^{1}\int\left((Z_{jmi}-H_nG_n^{-1})a_{jmi}\right)^{\otimes 2}S_{jmi}^{-1}Y_i\lambda_c dt \\
& -\int(H_nG_n^{-1}H_n'-H_nG_n^{-1}G_nG_n^{-1}H_n')dt+o_p(1) \\
= & -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{1}\sum_{m=0}^{1}\int\left((Z_{jmi}-H_nG_n^{-1})a_{jmi}\right)^{\otimes 2}S_{jmi}^{-1}Y_i\lambda_c dt+o_p(1) \\
\xrightarrow{p} & -I(\theta_0).
\end{aligned}
$$

In the third step, we use the fact

$$\sum_{j=0}^{1}\sum_{m=0}^{1}\frac{\partial}{\partial\theta}(Z_{jmi}-H_nG_n^{-1})a_{jmi}=0$$

based on the facts that $\sum_{j=0}^{1}\sum_{m=0}^{1} Z_{jmi}a_{jmi} = 0$, $\sum_{j=0}^{1}\sum_{m=0}^{1} a_{jmi} = 0$ for each i, and H_n and G_n are free of j, m and i.

Theorem 5.1. Given that (C1)-(C4), $\hat{\theta}$ is asymptotically efficient.

Proof. For any θ in a neighborhood of θ_0, we can show that $n^{-1}U_n(\theta, \hat{\Lambda}_{01}, \hat{\Lambda}_{02})$ is asymptotically equivalent to $n^{-1}U_n(\theta, \Lambda_{01}, \Lambda_{02})$ by the arguments similar as the proof of (C2). Also, it is easy to show that $n^{-1}U_n(\theta, \Lambda_{01}, \Lambda_{02}) \to V(\theta)$ $a.s.$, where

$$V(\theta) = \sum_{j=0}^{1}\sum_{m=0}^{1} E \int \frac{(Z_{jm} - HG^{-1})a_{jm}}{S_{jm}}(\theta, \Lambda_{01}, \Lambda_{02})\, S_{jm}(\theta_0, \Lambda_{01}, \Lambda_{02}) Y_i \lambda_c dt.$$

Note that $U_n(\hat{\theta}, \hat{\Lambda}_{01}, \hat{\Lambda}_{02}) = V(\theta_0) = 0$. Thus, based on the above facts and (C3)-(C4), it follows the inverse function theorem (Foutz, 1977) that $\hat{\theta}$ is a consistent estimate of θ_0.

The normality of $\hat{\theta}$ is obtained by using Taylor expansion given (C1)-(C4) and the consistency of $\hat{\theta}$. $\hat{\theta}$ is asymptotically efficient since the variance of $\hat{\theta}$ converges to the information bound $I^{-1}(\theta_0)$.

Table 1.1: 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 1.2: Data on the occurance of adrenal and lung tumors by the time of death for 100 male rats in the NTP study of chloroprene given in Dunson and Dinse (2002)

age at death[a] (months)	Control group 0 ppm				High dose group 80 ppm			
11	0	0	0	0[b]	2	0	0	0
16	1	0	0	0	2	0	0	0
17	1	0	0	0	1	0	0	0
18	4	0	0	0	5	0	0	1
19	3	0	0	0	3	0	0	0
20	4	2	0	0	4	0	0	0
21	2	2	1	0	7	3	0	1
22	5	3	1	0	5	0	0	1
23	0	0	0	0	2	2	1	0
24	3	4	0	0	0	5	0	0
25	1	0	0	0	0	1	0	0
25[c]	5	8	0	0	0	2	0	2
Total	29	19	2	0	31	13	1	5

[a] Day of death data are grouped into month for this summary table.

[b] Number of rats with no tumors, only adrenal, only lung, and both tumors, respectively.

[c] Animals sacrificed at the end of study.

Table 2.1: Estimated sizes of the goodness-of-fit test

	β=0	β=0.10	β=0.25	β=0.50
$c = 0$	0.054	0.057	0.056	0.049
$c = 1$	0.053	0.051	0.046	0.048

Table 2.2: Estimated powers of the goodness-of-fit test

	λ_1	λ_2	λ_3
$\beta = 0.10$	0.128	0.542	0.993
$\beta = 0.25$	0.289	0.923	0.998
$\beta = 0.50$	0.606	0.970	0.995

Table 3.1: Simulation results for estimation of α and τ

τ	Estimate	Bias	SSE	ESE	CP
		$n = 200$			
1/3	$\tilde{\tau}$	-0.0018	0.0523	0.0534	0.9540
	$\tilde{\alpha}$	0.0109	0.2421	0.2442	0.9600
	$\hat{\tau}$	0.0003	0.0563	0.0589	0.9520
	$\hat{\alpha}$	0.0230	0.2594	0.2743	0.9540
1/2	$\tilde{\tau}$	0.0034	0.0469	0.0468	0.9460
	$\tilde{\alpha}$	0.0639	0.3945	0.3917	0.9520
	$\hat{\tau}$	0.0019	0.0557	0.0567	0.9400
	$\hat{\alpha}$	0.0697	0.4942	0.4850	0.9580
		$n = 400$			
1/3	$\tilde{\tau}$	0.0001	0.0385	0.0379	0.9420
	$\tilde{\alpha}$	0.0103	0.1732	0.1722	0.9460
	$\hat{\tau}$	0.0002	0.0425	0.0410	0.9340
	$\hat{\alpha}$	0.0131	0.1929	0.1876	0.9360
1/2	$\tilde{\tau}$	-0.0015	0.0326	0.0331	0.9500
	$\tilde{\alpha}$	0.0050	0.2649	0.2673	0.9540
	$\hat{\tau}$	0.0026	0.0363	0.0380	0.9540
	$\hat{\alpha}$	0.0426	0.2979	0.3137	0.9640

Table 4.1: Simulation results for estimation of β_0 and γ_0 (part 1)

		$n = 100$				$n = 200$			
true	est	Bias	SSE	SEE	CP	Bias	SSE	SEE	CP
$\gamma = 1$	$\hat{\gamma}$	0.0056	0.1660	0.1620	0.950	0.0076	0.1067	0.1142	0.954
$\beta = -0.2$	$\hat{\beta}$	-0.0019	0.3726	0.3578	0.950	-0.0046	0.2627	0.2513	0.952
$\gamma = 1$	$\hat{\gamma}$	0.0162	0.1648	0.1629	0.946	0.0022	0.1147	0.1148	0.950
$\beta = 0$	$\hat{\beta}$	-0.0159	0.4082	0.3862	0.942	-0.0163	0.2717	0.2692	0.952
$\gamma = 1$	$\hat{\gamma}$	0.0123	0.1672	0.1629	0.950	0.0123	0.1158	0.1151	0.950
$\beta = 0.2$	$\hat{\beta}$	0.0192	0.4440	0.4169	0.950	-0.0043	0.3076	0.2912	0.944
$\gamma = -1$	$\hat{\gamma}$	-0.0079	0.1635	0.1628	0.940	-0.0058	0.1183	0.1149	0.954
$\beta = 0.2$	$\hat{\beta}$	-0.0040	0.3760	0.3502	0.944	-0.0003	0.2635	0.2409	0.944
$\gamma = -1$	$\hat{\gamma}$	0.0029	0.1685	0.1623	0.948	-0.016	0.1187	0.1152	0.946
$\beta = 0$	$\hat{\beta}$	-0.0164	0.3361	0.3143	0.940	0.0211	0.2376	0.2229	0.948
$\gamma = -1$	$\hat{\gamma}$	-0.0223	0.1629	0.1633	0.944	-0.0154	0.1182	0.1152	0.954
$\beta = -0.2$	$\hat{\beta}$	-0.0154	0.3049	0.2897	0.950	-0.0079	0.1970	0.1985	0.952

Table 4.2: Simulation results for estimation of β_0 and γ_0 (part 2)

γ	β	Estimte	Bias	SSE	SEE	CP
0	0	$\hat{\gamma}$	0.0008	0.1461	0.1437	0.954
		$\hat{\beta}$	0.0212	0.5987	0.5643	0.952
	0.5	$\hat{\gamma}$	-0.0003	0.1481	0.1435	0.944
		$\hat{\beta}$	0.0130	0.6558	0.6397	0.950
	-0.5	$\hat{\gamma}$	0.0025	0.1513	0.1435	0.946
		$\hat{\beta}$	-0.0215	0.5247	0.5044	0.948
0.5	0	$\hat{\gamma}$	-0.0267	0.1499	0.1486	0.948
		$\hat{\beta}$	-0.0021	0.6573	0.6068	0.940
	.5	$\hat{\gamma}$	-0.0273	0.1517	0.1492	0.954
		$\hat{\beta}$	0.0385	0.7209	0.6872	0.950
	-.5	$\hat{\gamma}$	-0.0271	0.1501	0.1482	0.944
		$\hat{\beta}$	0.0498	0.5151	0.5416	0.958
-0.5	0	$\hat{\gamma}$	0.0343	0.1481	0.1479	0.942
		$\hat{\beta}$	0.0443	0.5917	0.5733	0.954
	.5	$\hat{\gamma}$	0.0293	0.1512	0.1487	0.946
		$\hat{\beta}$	0.0648	0.7254	0.6456	0.950
	-.5	$\hat{\gamma}$	0.0244	0.151	0.1489	0.942
		$\hat{\beta}$	0.0621	0.4904	0.5007	0.950

Table 5.1: Simulation results when the censoring variable is independent of covariate, i.e., $w = 0$. The number of sieve is $k = 5$ and the sample size is $n = 100$.

		Likelihood method				Proposed method			
		Bias	SSE	SEE	CP	Bias	SSE	SEE	CP
0	$\hat{\beta}$	-0.0004	0.2632	0.2609	0.939	-0.0036	0.2709	0.2629	0.939
1/3	$\hat{\tau}$	-0.0152	0.1199	0.1129	0.930	-0.0180	0.1198	0.1148	0.937
1	$\hat{\beta}$	0.0693	0.3241	0.2988	0.948	0.0663	0.3256	0.3068	0.959
1/3	$\hat{\tau}$	-0.0111	0.1185	0.1189	0.939	-0.0147	0.1183	0.1211	0.949
0	$\hat{\beta}$	0.0034	0.2731	0.2761	0.955	-0.0033	0.2805	0.2775	0.953
1/2	$\hat{\tau}$	-0.0100	0.1058	0.1055	0.947	-0.0126	0.1062	0.1066	0.946
1	$\hat{\beta}$	0.0754	0.3390	0.3180	0.948	0.0709	0.3490	0.3254	0.950
1/2	$\hat{\tau}$	-0.0032	0.1104	0.1123	0.939	-0.0061	0.1110	0.1135	0.939

Table 5.2: Simulation results when the censoring variable is dependent of covariate and the censoring effect $w = 1$. The number of sieve is $k = 5$ and the sample size is $n = 100$.

		Likelihood method				Proposed method			
		Bias	SSE	SEE	CP	Bias	SSE	SEE	CP
0	$\hat{\beta}$	-0.0511	0.3102	0.3190	0.951	0.0188	0.3367	0.3265	0.949
1/3	$\hat{\tau}$	-0.0134	0.1227	0.1218	0.942	-0.0157	0.1226	0.1215	0.943
1	$\hat{\beta}$	0.0027	0.3482	0.32851	0.948	0.0565	0.3397	0.3394	0.957
1/3	$\hat{\tau}$	-0.0088	0.1154	0.1120	0.939	-0.0119	0.1153	0.1120	0.942
0	$\hat{\beta}$	-0.0624	0.3172	0.3350	0.957	0.0067	0.3474	0.3416	0.954
1/2	$\hat{\tau}$	-0.0163	0.1127	0.1116	0.932	-0.0182	0.1126	0.1114	0.934
1	$\hat{\beta}$	0.0152	0.3684	0.3476	0.943	0.0677	0.3600	0.3590	0.960
1/2	$\hat{\tau}$	-0.0103	0.1070	0.1049	0.930	-0.0132	0.1074	0.1049	0.934

Table 5.3: Simulation results when the censoring variable is independent of covariate, i.e., $w = 0$. The number of sieve is $k = 7$ and the sample size is $n = 100$.

		Likelihood method				Proposed method			
		Bias	SSE	SEE	CP	Bias	SSE	SEE	CP
0	$\hat{\beta}$	-0.0467	0.2522	0.2667	0.942	-0.0537	0.2938	0.2630	0.925
1/3	$\hat{\tau}$	-0.0206	0.1111	0.1144	0.949	-0.0206	0.1109	0.1151	0.951
1	$\hat{\beta}$	0.1153	0.3285	0.3098	0.953	0.0894	0.3360	0.3119	0.939
1/3	$\hat{\tau}$	-0.0114	0.1170	0.1217	0.950	-0.0112	0.1160	0.1225	0.955
0	$\hat{\beta}$	-0.0428	0.2477	0.2825	0.962	-0.0496	0.2954	0.2769	0.949
1/2	$\hat{\tau}$	-0.0131	0.1030	0.1063	0.955	-0.0151	0.1028	0.1072	0.957
1	$\hat{\beta}$	0.1293	0.3542	0.3305	0.948	0.0886	0.3699	0.3271	0.929
1/2	$\hat{\tau}$	-0.0140	0.1111	0.1153	0.956	-0.0130	0.1099	0.1158	0.954

Table 5.4: Simulation results when the censoring variable is dependent of covariate and the censoring effect $w = 1$. The number of sieve is $k = 7$ and the sample size is $n = 100$.

		Likelihood method				Proposed method			
		Bias	SSE	SEE	CP	Bias	SSE	SEE	CP
0	$\hat{\beta}$	-0.0834	0.2499	0.3244	0.964	-0.0849	0.3336	0.3244	0.941
1/3	$\hat{\tau}$	-0.0095	0.1142	0.1247	0.962	-0.0113	0.1156	0.1244	0.953
1	$\hat{\beta}$	0.0926	0.3658	0.3417	0.945	0.1026	0.3741	0.3493	0.941
1/3	$\hat{\tau}$	-0.0241	0.1116	0.1151	0.944	-0.0261	0.1116	0.1155	0.942
0	$\hat{\beta}$	-0.0566	0.2612	0.3409	0.964	-0.0560	0.3646	0.3409	0.943
1/2	$\hat{\tau}$	-0.0190	0.1147	0.1134	0.942	-0.0208	0.1160	0.1133	0.941
1	$\hat{\beta}$	0.1180	0.3796	0.3575	0.934	0.1258	0.3938	0.3668	0.934
1/2	$\hat{\tau}$	-0.0207	0.1052	0.1075	0.949	-0.0220	0.1048	0.1077	0.949

Table 5.5: Simulation results when the censoring variable is independent of covariate, i.e., $w = 0$. The number of sieve is $k = 6$ and the sample size is $n = 200$.

		Likelihood method				Proposed method			
		Bias	SSE	SEE	CP	Bias	SSE	SEE	CP
0	$\hat{\beta}$	-0.0062	0.1711	0.1817	0.967	-0.0058	0.1735	0.1815	0.966
1/3	$\hat{\tau}$	-0.0051	0.0809	0.0790	0.953	-0.0070	0.0809	0.0795	0.952
1	$\hat{\beta}$	0.0312	0.2075	0.2036	0.957	0.0278	0.2103	0.2063	0.950
1/3	$\hat{\tau}$	-0.0029	0.0853	0.0830	0.937	-0.0052	0.0854	0.0836	0.939
0	$\hat{\beta}$	-0.0092	0.1831	0.1919	0.958	-0.0098	0.1876	0.1914	0.955
1/2	$\hat{\tau}$	0.0008	0.0694	0.0736	0.957	-0.0009	0.0694	0.0739	0.961
1	$\hat{\beta}$	0.0377	0.2213	0.2162	0.958	0.0352	0.2217	0.2198	0.948
1/2	$\hat{\tau}$	0.0030	0.0826	0.0785	0.929	0.0012	0.0829	0.0787	0.932

Table 5.6: Simulation results when the censoring variable is dependent of covariate with censoring effect $w = 1$. The number of sieve is $k = 6$ and the sample size is $n = 200$.

		Likelihood method				Proposed method			
		Bias	SSE	SEE	CP	Bias	SSE	SEE	CP
0	$\hat{\beta}$	-0.0561	0.2107	0.2232	0.950	-0.0017	0.2207	0.2247	0.959
1/3	$\hat{\tau}$	-0.0020	0.0858	0.0858	0.952	-0.0034	0.0860	0.0854	0.943
1	$\hat{\beta}$	-0.0049	0.2357	0.2250	0.935	0.0287	0.2322	0.2275	0.941
1/3	$\hat{\tau}$	-0.0030	0.0828	0.0785	0.938	-0.0054	0.0828	0.0784	0.934
0	$\hat{\beta}$	-0.0663	0.2224	0.2338	0.954	-0.0174	0.2356	0.2347	0.956
1/2	$\hat{\tau}$	-0.0053	0.0768	0.0782	0.952	-0.0067	0.0766	0.0780	0.951
1	$\hat{\beta}$	-0.0023	0.2493	0.2375	0.946	0.0311	0.2462	0.2393	0.947
1/2	$\hat{\tau}$	-0.0013	0.0748	0.0734	0.939	-0.0034	0.0751	0.0733	0.937

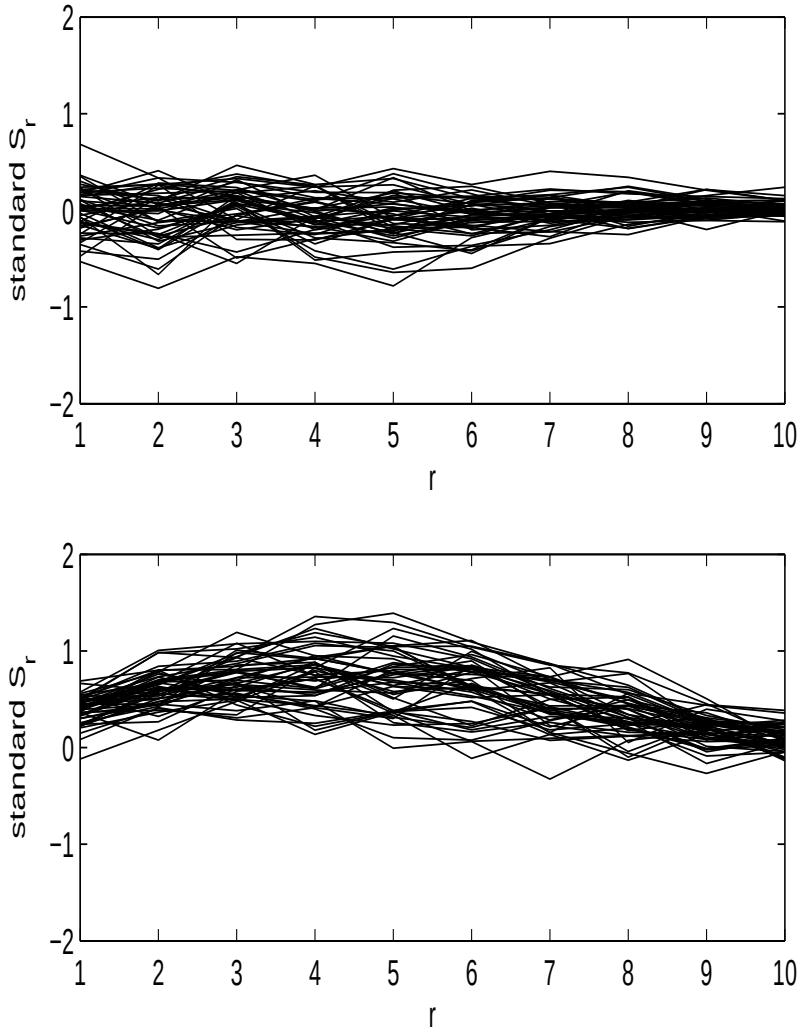

Figure 2.1: Standardized test processes based on simulated data with 50 replications; The top plot is under null hypothesis and the bottom plot is under alternative hypothesis.

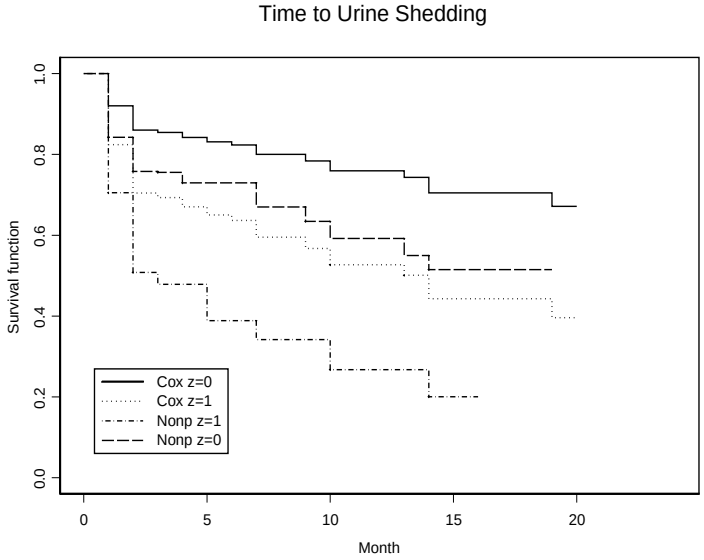

Figure 2.2: Estimates of survival functions under nonparametric setting and Cox model setting with the same baseline hazard function for urine shedding and blood shedding.

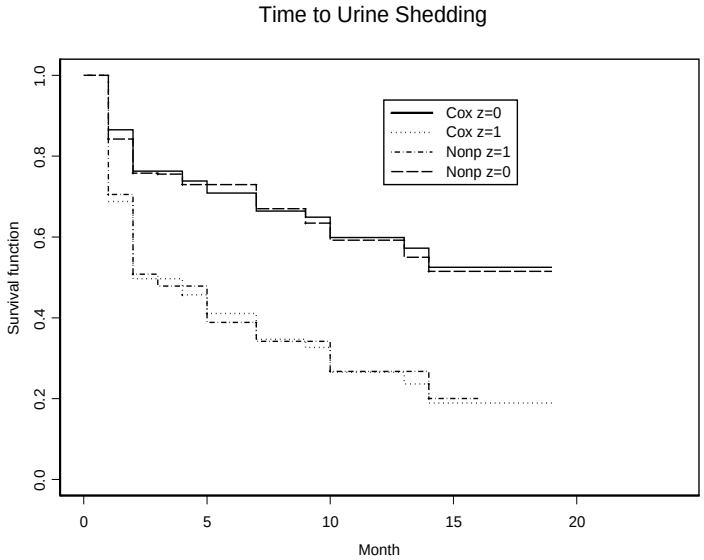

Figure 2.3: Estimates of survival functions under nonparametric setting and Cox model setting with different baseline hazard functions for urine shedding and blood shedding.

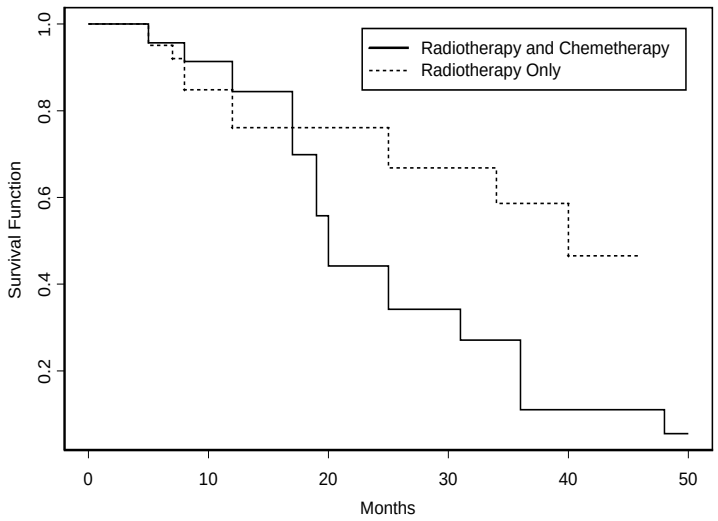

Figure 4.1: Maximum likelihood estimators of the two survival functions

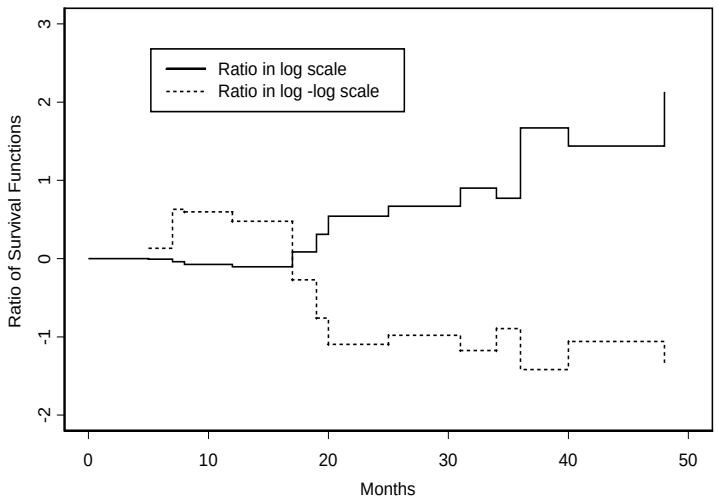

Figure 4.2: Ratios of estimated survival functions in log and log-log scales

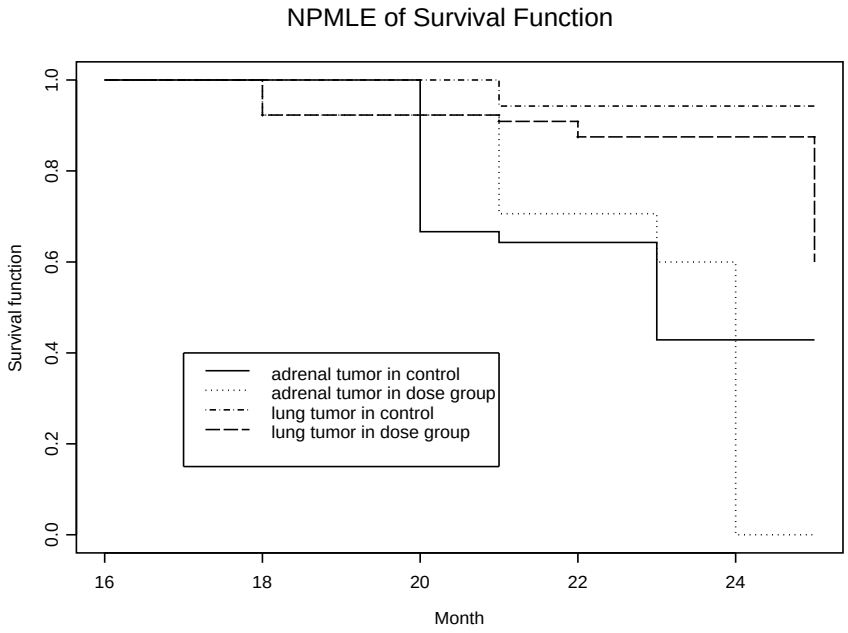

Figure 5.1: Nonparametric maximum likelihood estimators of the survival functions of adrenal tumor and lung tumor in the high dose group and the control group.

VITA

Lianming Wang was born on November 9, 1977 in Linyi City, Shandong Province, People's Republic of China. He received his B.A. in Mathematics from Shandong University in Ji'nan in 1999 and M.S. in Statistics from East China Normal University in Shanghai in 2002. In the same year, he joined the Department of Statistics at the University of Missouri-Columbia. He will receive his Ph.D. in Statistics in August 2006. As of August 2006, he will work as a Postdoctoral Fellow at the National Institute of Environmental Health Sciences in Research Triangle Park, North Carolina.