THE NONPARAMETRIC ANALYSIS OF INTERVAL-CENSORED FAILURE TIME DATA

A Dissertation

Presented to

the Faculty of the Graduate School

University of Missouri

In Partial Fulfillment

Of the Requirements for the Degree

Doctor of Philosophy

by

RAN DUAN

Dr. (Tony) Jianguo Sun, Dissertation Supervisor

May, 2013

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

THE NONPARAMETRIC METHODS FOR THE ANALYSIS OF INTERVAL-CENSORED DATA

presented by RAN DUAN

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. Subharup Guha ______________________________

Dr. Jing Qiu ______________________________

Dr. Davis Wade ______________________________

ACKNOWLEDGEMENTS

I owe my deepest gratitude to my esteemed advisor Dr. (Tony) Jianguo Sun for his generous support and help, wonderful guidance and endless patience, which enable me to complete this work and obtain the Ph.D degree. His continual encouragement and instruction lead me to a good start of my research life.

I extend my gratitude to my advisory committee members: Dr. Nancy Flournoy, Dr. Subharup Guha, Dr. Jing Qiu and Dr. Davis Wade for their insightful comments and suggestions on my work. Special thanks are due to Dr. Hui Zhao and Dr. Yanqing Feng for their helpful academic discussion and help.

I also want to express my deep gratitude to all faculty who taught me classes and assisted me in various ways during my course studies. I would also like to mention my special appreciation to Dr. Larry Ries for helping me to be a better instructor.

I want to thank our great staff Judy , Tracy and Kathleen for plenty of help. I would like to take this opportunity to extend many thanks to my fellow graduate students for making the department a strong one and I will always treasure the friendship and appreciate the help of friends here at Mizzou.

I am deeply indebted to my family, especially my parents, Tiansheng Zhu and Xiaochun Duan, for their unselfish love and support throughout my life. Its their company and care make my life so happy and wonderful.

Table of Contents

List of Tables

List of Figures

THE NONPARAMETRIC METHODS FOR THE ANALYSIS OF INTERVAL-CENSORED FAILURE TIME DATA

Ran Duan

Dr. (Tony) Jianguo Sun, Dissertation Supervisor

ABSTRACT

By interval-censored failure time data, we mean that the failure time of interest is observed to belong to some windows or intervals, instead of being known exactly. One would get an interval-censored observation for a survival event if a subject has not experienced the event at one follow-up time but had experienced the event at the next follow-up time. Interval-censored data include right-censored data (Kalbfleisch and Prentice, 2002) as a special case.

Nonparametric comparison of survival functions is one of the main tasks in failure time studies such as clinical trials. For interval-censored failure time data, a few nonparametric test procedures have been developed. However, due to the strict restrictions of existing nonparametric tests and practical demands, some new nonparametric tests need to be developed.

This dissertation consists of four parts. In the first part, we propose a new class of test procedures whose asymptotic distributions are established under both null and alternative hypotheses, since all of the existing test procedures cannot be used if one intends to perform some power or sample size calculation under the alternative hypothesis. Some numerical results have been obtained from a simulation study for assessing the finite sample performance of the proposed test procedure. Also we applied the proposed method to a real data set arising from an AIDS clinical trial concerning the opportunistic infection cytomegalovirus (CMV).

The second part of this dissertation will focus on the nonparametric test for interval-censored data with unequal censoring. As we know, one common drawback or restriction of the nonparametric test procedures given in the literature is that they can only apply to situations where the observation processes follow the same distribution among different treatment groups. To remove the restriction, a test procedure is proposed, which takes into account the difference between the distributions of the censoring variables. Also the asymptotic distribution of the test statistics is developed by counting process and martingale theory. For the assessment of the performance of the procedure,

a simulation study is conducted and suggested that it works well for practical situations. An illustrative example from a study aiming to investigate the HIV-1 infection risk among hemophilia patients is provided.

The third part of this dissertation deals with the regression analysis of multivariate interval-censored data with informative censoring. Multivariate interval-censored failure time data often occur in the clinical trial that involves several related event times of interest and all the event times suffer interval censoring. Different types of models have been proposed for the regression analysis (Zhang et al.(2008); Tong et al.(2008); Chen et al.(2009); Sun (2006)). However, most of these methods only deal with the situation where observation time is independent of the underlying survival time completely or given covariates. In this chapter, we discuss regression analysis of multivariate interval-censored data when the observation time may be related to the underlying survival time. An estimating equation based approach is proposed for regression coefficient estimate with the additive hazards frailty model and the asymptotic properties of the proposed estimates are established by using counting processes. A major advantage of the proposed method is that it does not involve estimation of any baseline hazard function. Simulation results suggest that the proposed method works well for practical situations.

Finally, we will talk about the directions for future research. One is about the nonparametric test for interval-censored data with informative censoring. The other is about multiple generalized log-rank test for interval censored data.

Chapter 1

Introduction

1.1 Data Structure

1.1.1 Interval-Censored Data and Censoring Mechanism

Researchers working with survival data often face with issues associated with incomplete data, especially censoring issues. One important type of censored data is called interval-censored data. By interval-censored data, we mean that study subjects are not under continuous observation. As a result, the survival times could not be observed exactly and one can only observe a time interval within which the event has occurred. Exact, right-censored and left-censored failure time data are special cases of interval-censored data. The exact failure time data occur when the censoring interval is reduced to a single point and interval-censored data become right-censored data or left-censored data when the right boundary of the interval is infinity or the left boundary is zero.

Interval-censored data occur in all kinds of areas, for instance, epidemiology, finance, social science, etc.. One real example comes from an oncology phrase III trial for breast cancer. In its statistical analysis plan, cancer progression time was predetermined as the secondary end point. During the trail, patients needed to visit the physician periodically in order to take the treatment therapy and the visiting times of each breast cancer patient were recorded, which was a periodic discontinuous observation process. Therefore, we only knew that the progression time of a breast cancer patient fell into a time interval, which was from the last visiting time with no cancer progression to the first visiting time with cancer progression.

There are three types of interval-censored data. Case I interval-censored data (Groeneboom and Wellner, 1992) , which also known as current status data, refer to the situation where each study subject only has one observation time and the failure time is either left-censored or right-censored. One commonly used notation of current status data is

$$\{C, \delta = I(T \leq C\}, \tag{1.1}$$

where C denotes the observation time and δ is the censoring indicator (Sun 2006). Current status data often occur in demographical studies. For example, the ages of patients with respect to the incidence of a disease, of which the exact incidence times are hard to measure.

Case II interval-censored data (Groeneboom and Wellner, 1992) refer to the situation that a time interval (L, R) instead of the exact failure time has been observed for a study subject. Here L is the lower boundary and R is the upper boundary of the interval. One way to present interval-censored data is

$$\{U,\ V,\ \delta_1 = I(T \leq U),\ \delta_2 = I(U < T \leq V),\ \delta_3 = 1 - \delta_1 - \delta_2\}, \tag{1.2}$$

where U and V represent observation processes. Here U and V are two random variables satisfying $U \leq V$ and δ is the censoring indicator (Sun, 2006). A more general situation is called Case K interval-censored data (Schick and Yu, 2000), which assume that each study subject has K observation points. The failure time falls in one of the $K + 1$ intervals and K is an random variable. Then the observation information can be presented in the form of

$$\{K,\ U_j,\ \delta_j = I(U_{j-1} < T \leq U_j),\ j = 1, \ldots, K,\ U_0 = 0\}. \tag{1.3}$$

The third type is doubly censored data (Sun, 1995), which occur under the situation that the objective of a clinical trial involves two related events. Then each individual has two event times and both of them are interval-censored. A common source of doubly censored failure time data is disease progression studies. For example, in a clinical trial, researchers are interested in tumor progression time as well as failure time for patients. Both of them are measured by the visiting time of patients. Therefore, doubly censored data have been collected at the end of the study.

1.1.2 Informative Censoring and Unequal Censoring

One common assumption of interval-censored data is that the observation process and event time are independent, which could be violated in practice. For example, a patient may withdraw from the study because the tumor grows too fast. Thus early withdraw may indicate a sooner death than expected. On the contrary, a patient may withdraw from the study because he is getting better and there is no need for him to take such intense therapy. Under this situation, earlier drop off may indicate a longer survival. This type of censoring mechanism is called informative censoring or dependent censoring, which means that the observation process and event time are dependent.

Besides informative censoring, one other censoring mechanism, often mentioned in literature, is unequal censoring mechanism. Unequal censoring means the distributions of observation process are related with treatment assignment. For example, therapy in treatment group requires patients to take blood test monthly while patients in control group only need to take blood test quarterly. Therefore, patients in treatment group have more chance to explore to the doctor which might affect their therapeutic outcome and event time. Unequal distributions of the observation process may add bias on the estimation of treatment effect.

1.1.3 Three Examples

1.1.3.1 Current Status Data

A breast feeding data (M. U. Ferreira, 1996) had been collected from a study performed in Santo Andre, Sao Paulo, Brazil in 1991. This study randomly selected a sample of children aged between 0 to 1 from 22 public health centers. Complete information was available for 2411 children. One objective of this study was to estimate the distribution of time to weaning.

Mothers were interviewed during the three months investigation and asked about their infant feeding practice. The data consisted of the current age of children and the indicator, weaned or not, at the time of survey. Total breastfeeding children include both exclusive breastfed children and partial breastfed children. Children's age were measured in days, which was from the data of birth recorded in their health card to the date of interview. Here the current age of the child was the observation time C and whether or not weaned at the time of survey was the censoring indicator δ. Since direct queries about time to wean yields severe measurement error in practice, current status data are favorable in this study (Grummer, 1993).

1.1.3.2 Interval-Censored Data

A main source of interval-censored data is medical study with periodic follow up. A breast cosmesis data had been produced from a retrospective study on early breast cancer patients at Joint Center for Radiation in Boston between 1976 and 1980.

It is known that adjuvant chemotherapy improves the relapse-free and overall survival for patients treated initially by mastectomy. However, both experimental and clinical evidence show that chemotherapy enhances the severe response of normal tissue to radiation therapy. Acute skin reactions are even worse when patients taking adjuvant chemotherapy along with radiation treatment for breast cancer. Moreover, the long-term impact of chemotherapy treatment on radiation therapy of the breast is still unknown. Therefore, researches purposed to compare the patients who were given adjuvant chemotherapy followed by radiation treatment to those who received only the radiation treatment, to determine the effect of chemotherapy on the cosmetic state.

Patients had been scheduled a periodic clinic visits every 4 to 6 months, but the actual visit time differs patients by patients. No exact time was observed. Data presented in (Finkelstein and Wolfe, 1985) contained the time interval of cosmetic deterioration for 94 early breast cancer patients. 46 of them were treated by Radiotherapy only and the rest of them were treated with adjuvant chemotherapy in conjunction with primary radiation treatment. More details of this data set can be found in (Finkelstein and Wolfe, 1985; Sun, 2006).

1.1.3.3 Bivariate Interval-Censored Data

A data set had been collected from an AIDS clinical trial, which had been conducted by AIDS Clinical Trial Group (ACTG)181 on HIV-infected individuals. During the trial, blood and urine samples were collected from the patients every time they visit the clinical center to test for the presence of CMV. The time to CMV shedding in blood and in urine are two event times of interest and both of them are interval censored.

Patients are classified into two groups based on their CD4 cell counts, which is used to indicate the status of a person's immune system or the stage of HIV infection. Patients with their CD4 cell count less than 75(cell/μl) is assigned to group 1 and group 2 otherwise. For the data set, one problem of interest is to determine the relationship between CMV shedding and CD4 cell counts. This data set had been first studied by Goggins and Finkelstein (2000).

1.1.3.4 Interval-Censored Data with Informative Censoring

A randomized study on the prophylaxis of pneumocystis carinii pneumonia(PCP) described in Lin (1996) is one example of interval-censored data with informative censoring. Researchers enrolled 310 AIDS patients who had recovered from PCP. 154 of them received trimethoprim sulfamethoxazole(TS) and the rest 156 patients received aerosolised pentamidine (AP). Finally, there were 43 patients died in TS group, among which 36 deaths happened prior to recurrences of PCP. For AP group, there were 47 deaths and 36 of them occurred before relapse of PCP. Some of the patients were withdrawn from the trial because of health issues. Therefore, statistician needs to concern about issues of dependent censoring due to early deaths and selected withdrawn when doing treatment comparison.

1.2 The Analysis of Interval-Censored Data

1.2.1 Nonparametric Comparison of Univariate Interval-Censored Data

Survival comparison is usually one of main goals in survival studies. Finklestein (1986), in her paper, assumed that the survival time follows Cox model and first developed a score test for interval-censored data. However, in most of practical problems, the proportional hazards assumption is too restrict. More nonparametric test procedures have been developed to deal with treatment comparison problems for interval-censored data.

In the following subsection, we discuss five different types of nonparametric tests for interval-censored data. The first one is a Wilcoxon type test (Sun, 1999), which can also be used to address the comparison problem of interval-censored data with unequal censoring after adjustment. The second one is a rank based procedure, a generalization of log rank test on interval-censored data. The third one is a survival based procedure, which considered the difference of survival function among treatment groups. The forth one is called generalized log-rank test (Zhao and Sun, 2005), which is a generalization of log rank test presented in Peto (1972)'s paper. This is one of the most commonly used nonparametric tests for interval-censored data nowadays. And the last one is the imputation test.

1.2.1.1 Wilcoxon Type Test

The Wilcoxon type test (Sun and Kalbfleisch, 1993) was first developed to deal with the treatment comparison problem of current status data. We notice that most of the existing procedures assume that the censoring mechanism is the same for different treatments. That is C_i follows the same distribution for subjects in different groups. Sun (1999) extended the restriction and proposed a test which allows the distributions of C_i's to depend on treatment assignment. To give a representative of such procedures, in the following, we describe the test proposed by Sun (1999).

Suppose there are n independent subjects enrolled in study. The observed data for subject i consist of $\{(C_i, \delta_i, Z_i)\}$ for $i = 1, \ldots, n$. For simplicity, we only consider two groups comparison here. C_i is the observed time and δ_i is the censoring indicator. $\delta = 1$ represents the event has occurred ; $\delta = 0$ represents the event has not occurred by the observed time. Z_i is the group indicator. $Z_i = 0$ when subject i belongs to control group and $Z_i = 1$ when i has been assigned to treatment group. If the observation time follows the same distribution, we can use the following Wilcoxon statistic

$$U_1 = \sum_i \sum_j (Z_i - Z_j)(\delta_i - \delta_j), \qquad (1.4)$$

to test the hypothesis $H_0 : S_1(t) = S_2(t)$. It can be proved that the above test statistic is equivalent to

$$U_1 = \sum_{i=1}^{n} (Z_i - \bar{Z})\delta_i,$$

where $\bar{Z} = \sum_{i=1}^{n} Z_i/n$. And under null hypothesis, U_1 has asymptotic normal distribution. However, when the distribution of C_i is dependent with Z_i, it may introduce bias to the test statistic. To correct the bias, they introduced a censoring indicator $N(t) = I(T \leq t)$. Then the observed data consist of $\{(C_i, N_i(C_i), Z_i)\ ,\ i = 1, \ldots, n\}$. To test $H_0 :\ S_1(t) = S_2(t)$ is equivalent to test the hypothesis $H_0 :\ E(N_i(t)|Z_i)$ is independent of Z_i (Sun, 1999). Motivated by (1.4), the test statistic can be written as

$$U_{12} = \sum_{i=1}^{n} (Z_i - \bar{Z}) N_i(C_i).$$

Suppose the hazard function of C_i follows a proportional hazards model

$$\lambda(t; Z_i) = \lambda_0(t) e^{Z_i \beta},$$

under the proportional hazards model assumption and null hypothesis, it can be shown that

$$E[N_i(C_i)|Z_i] = E[\int_0^\infty N_i(t) d\tilde{N}_i(t)|Z_i] = e^{Z_i'\beta} \int_0^\infty \lambda_0(t)\mu(t)[S_0(t)]^{exp(Z_i'\beta)} dt,$$

where $\mu(t)$ is the mean function of the $N_i(t)$, $\tilde{N}_i(t) = I(t \leq C_i)$ and $S_0(t) = exp[-\int_0^t \lambda_0(s)ds]$ which is the baseline survival function of the C_i. Then the test statistic can be rewritten as:

$$U_{13}(\beta) = \sum_{i=1}^{n} (Z_i - \bar{Z}) e^{-Z'\beta} \frac{N_i(C_i)}{\hat{S}_0(C_i; \beta)^{exp(Z_i'\beta)}},$$

where

$$\hat{S}_0(t;\beta) = exp\left[-\int_0^t \frac{d\tilde{N}(s)}{\sum_{i=1}^n I(s \le C_i)e^{Z_i'\beta}}\right].$$

1.2.1.2 Rank Based Test

The basic idea of rank based test is that, under null hypothesis, the summation of difference between observation and expectation of the number of failure events equals zero. Under null hypothesis, the survival functions of different treatment groups are the same. Therefore, the survival times of different groups should asymptotically share the same survival function. Then the difference between observation and expectation estimated by the pooled data should asymptotically equals zero. On the other hand, if null hypothesis is not satisfied, the test statistic based on the common survival function should no longer equal zero.

For case II interval-censored data, consider a survival study that involves n independent subjects. Let T_i denotes the survival time of interest for subject i, $i = 1, \dots, n$. Suppose that for subject i, we only observe $\{U_i, V_i, \Delta_i = I(T_i \le U_i), \Gamma_i = I(U_i < T_i \le V_i)\}$, where U_i and V_i are non-negative random variables independent of T_i such that $U_i < V_i$ with probability one, $i = 1, \dots, n$. This means that one only knows if T_i is smaller than U_i, between U_i and V_i, or larger than V_i. Assume that the study involves $p+1$ groups. Let $F_1(t), \dots, F_{p+1}(t)$ denote the cumulative distribution functions of the T_i's for the subjects in different treatment groups, respectively. To test the hypothesis $H_0 : F_1(t) =, \dots, = F_{p+1}(t)$, Sun (1996) and Sun et al. (2004) proposed the test statistic $U_{22} = (U_{22,1}, \ldots, U_{22,p+1})'$:

$$U_{22,l} = \sum_{j=1}^{m} \left(d_{jl} - \frac{n_{jl} d_j}{n_j} \right). \tag{1.5}$$

where d_j is the overall observed failure numbers of patients and n_j is the overall observed numbers of patients at risk at time s_j. The observed failure and risk numbers at time s_j for treatment group l are d_{jl} and n_{jl}. Under H_0, the test statistic U_{22} approximately follows normal distribution with mean zero and variance V_{22}. Zhao and Sun (2004) developed a multiple imputation approach to estimate variance matrix. There are other methods to implement the variance estimation, such as Fisher information matrix (Sun, 1996) or resampling approach (Sun, 2001).

The rank based test, as we can see, is a generalization of log rank test from right censored data to interval-censored data. And it can be reduced to log rank test if the data are all right-censored. Other similar approaches involve the weighted log rank test developed by Fleming and Harrington (1991) and , most recenlty, a generalized weighted log rank test proposed by Oller et al.(2012).

1.2.1.3 Survival Based Test

Another type of test statistic is called survival based procedure. Petroni and Wolfe (1994) developed a test statistic, which is a measure of distance between the survival functions of different treatment groups, for discrete survival time. Followed the same idea, Fang et al. (2002) and Zhang et al. (2001) moved their attention to continuous survival times. The basic idea of survival based test is very straightforward. The survival based test statistic is a measure of distance between two continuous functions. Therefore, under the null hypothesis, the expectation of the test statistic should asymptotically converge to zero.

Consider the two sample comparison problem. To test the null hypothesis, one can construct the following test statistic:

$$\int_0^\tau W(t)[\hat{S}_1(t) - \hat{S}_2(t)]dt. \tag{1.6}$$

Here τ is the largest observation time and $\hat{S}_1(t)$, $\hat{S}_2(t)$ are the NPMLE of $S_1(t)$, $S_2(t)$, which are the survival functions of treatment group 1 and 2, respectively. $W(t)$ is a weight process that can depend on observed data. We can see that this test statistic measures the weighted difference between the survival functions of the two groups.

Based on the basic idea of the test statistic (1.7), Fang (2002) introduced a test statistic

$$U_{31} = \sqrt{\frac{n_1 n_2}{n}} \int_0^\tau w(t)[\hat{S}_1(t) - \hat{S}_2(t)]dt.$$

Assume that $n_1/n \to p$ as $n \to \infty$, where $0 < p < 1$. Also assume $w(t)$ is a deterministic function with a bounded derivative on $[0, \tau]$. Under null hypothesis, as $n \to \infty$, the statistic has an asymptotic normal distribution with mean 0 and more detail about the consistent variance estimation could be found in Fang et al. (2002).

Under the same situation, let $\hat{H}$, $\hat{H}_1$ and $\hat{H}_2$ denote the empirical distributions of (U, V), U and V, respectively. Let $\hat{S}_0$ denote the NPMLE of the common survival function, Zhang (2001) constructed a test statistic

$$U_{32} = \sqrt{n}\int_0^\tau [w(u)\{\hat{S}_1(u) - \hat{S}_2(u)\}d\hat{H}_1(u) + w(v)\{\hat{S}_1(v) - \hat{S}_2(v)\}d\hat{H}_2(v)],$$

which can be approximated by the normal distribution under the null hypothesis. As we can see, the fundamental difference between rank based test and survival based test is that the former measures the differences between the estimated hazard functions while the latter relies on the differences between the estimated survival functions.

1.2.1.4 Generalized Log-Rank Test

Motivated by the test statistic present in Peto and Peto (1972), Zhao et al. (2005) proposed the following test statistic

$$U_\xi = \sum_{i=1}^{n} z_i \frac{\xi\{\hat{S}_n(L_i)\} - \xi\{\hat{S}_n(R_i)\}}{\hat{S}_n(L_i) - \hat{S}_n(R_i)}, \tag{1.7}$$

where ξ is a known function over $(0,1)$. Also we need to assume that $\lim_{x\to 0}\eta(x) = \lim_{x\to 1}\eta(x) = c_0$. In practice, different ξ can be used and will yield different test statistics. In that paper, they used $\xi(x) = xlog(x)$. Let $S_0(t)$ denote the common survival function under H_0 and $\hat{S}_n(t)$ be the NPMLE of $S_0(t)$. z_i is the treatment indicator.

In order to establish the asymptotic distribution of test statistic, they assume $F_0(t)$ has a support in $[0, M]$ with a continuous density function and $Pr(0 < U \leq V <$

$M) = 1$. Also assume that F_0 is a strict monotonic cumulative density function. It can be shown that, under the regularity conditions for the consistency of $\hat{F}_n$, as $n \to \infty$, $U_\eta/\sqrt{n}$ has an asymptotic normal distribution with mean zero and covariance matrix $\Sigma = (\sigma_{lr})_{k \times k}$ under H_0. k is the number of treatment groups. $\sigma_{lr} = p_l(1 - p_l)Q_0(K_0^2)$ if $l = r$, and $\sigma_{lr} = -p_l p_r Q_0(K_0^2)$ otherwise. More details can be found in (Zhao et al., 2005).

The key advantage of the generalized log rank test is that the test statistic has the asymptotic distribution and we can applied this method regardless of the distribution of the survival time. Also the variance estimation in this method is relatively easier to compute than other nonparametric tests.

1.2.1.5 Multiple Imputation Approach

A common method of dealing with missing data is to impute a value for each missing data. Basically, censored data differ from missing data since it provides incomplete information about the event time (Sun, 2006). However, for interval-censored data, the exact event time can still be treated as missing , since it is only known that the failure time falls into a time interval.

The proposal for multiple imputation approach is to replace the interval-censored data with right-censored data via multiple imputation technique and then applied the nonparametric test for right-censored data for the treatment compariation. Specifically, Pan (2000) replaced each interval-censored observation U_i, V_i with a failure time T_i, which satisfied $U_i < T_i \leq V_i < \infty$ and a censoring indicator δ_i. $\delta_i = 1$ if the subject was not censored. More details about the test statistic and variance-covariance matrix estimation can be found in their paper. Huang et al.(2008) draw the event time T_i^h from the conditional probability function

$$P(T_i^h = s_j | T_i^h \in [L_i, R_i]) = \frac{\alpha_{ij}\hat{w}_j}{\sum_{v=1}^{m} \alpha_{iv}\hat{w}_v}, \quad s_j \in [L_i, R_i],$$

here,

$$w_j = \frac{1}{n}\sum_{i=1}^{n} \frac{\alpha_{ij} w_j}{\sum_{v=1}^{m} \alpha_{iv} w_v}.$$

Then applied the log-rank test to the right-censored data and calculated the test statistic U^h. Repeat the test procedure for h from 1 to H. Let $\bar{U} = \sum_{h=1}^{H} U^h/H$. The proposed test statistic has the form:

$$\bar{U}^T(\hat{V}^{-1})\bar{U},$$

where V is the covariance matrix of $\bar{U}$.

Several authors had also considered the multiple imputation approach for interval-censored data. For example, Bebchuk and Betensky (2000) discussed the estimation of hazard function. Most recently, Fay et al. (2012) studied the log-rank test for interval-censored data when assessment times depend on treatment. They modified the multiple imputation log-rank tests of Huang et al. (2008) and showed through simulations that the modifications of the multiple imputation log-rank tests retain the type I error rate under the case of assessment-treatment dependence and the case of a small number of individuals in each treatment group.

1.2.2 Regression Analysis of Univariate Interval-Censored Data

For regression analysis, the primary objective is to estimate the covariate effects on the event time. In most cases, the hazard function and the baseline survival function are treated as a infinite dimension nuisance parameter.

In 1986, Finkelstain considered a proportional hazards model to fit the interval-censored data and proposed a score test for treatment comparison . Lin et al. (1998) constructed an additive hazards model to analyze the current status data and developed the asymptotic distribution of the parameter by counting process. Zeng et al. (2006) applied a full likelihood approach to study the efficient estimation of regression parameter in the same model. Wang et al. (2010) generalized the additive hazards model to Case II interval-censored data. other methods include the accelerated failure time model proposed by Betensky et al. (2001), proportional odds model (Huang and Rossini, 1997; Sun, 2006; Sun et al., 2007) and the linear transformation model which gives more flexibility for the relationship between the failure time T and covariate Z.

Suppose the study has n independent subjects and the observed time yields an case II interval-censored format.

$$\{U_i,\ V_i,\ Z_i,\ \delta_{1i} = I(T \leq U),\ \delta_{2i} = I(U < T \leq V),\ \delta_{3i} = 1 - \delta_{1i} - \delta_{2i}\},$$

here Z is the p dimensional vector of covariates. Let $S(t; Z_i)$ denote the survival function with covariates Z. The likelihood is proportional to

$$L = \prod_{i=1}^{n} [S(L_i, Z_i) - S(R_i, Z_i)].$$

For the analysis of interval-censored data, we first discuss the proportional hazards model, which uses the maximum likelihood approach to estimate the parameter.

1.2.2.1 Proportional Hazards Model

The proportional hazards model has been widely used in right-censored data. By partial likelihood function, one can estimate the regression parameter without specifying baseline hazard function. Also the asymptotic properties of regression parameter have been developed by counting process (Anderson and Gill, 1982). Finkelstein (1986) studied this approach for interval-censored data. Huang and Wellner (1996) proved the MLE of regression parameter is consistent and efficient and has asymptotic normal distribution with $n^{1/2}$ convergence rate.

Under the proportional hazards assumption, we have

$$\lambda(t) = \lambda_0(t) exp(Z_i'\beta). \tag{1.8}$$

The log likelihood function then has the form

$$l(\beta, S_0) = \sum_{i=1}^{n} log\{S_0(L_i)^{exp(Z_i'\beta)} - S_0(R_i)^{exp(Z_i'\beta)}\},$$

here $S_0(t)$ is the baseline survival function and β is the regression parameter.

To estimate the regression parameter, Finkelstein (1986) first proposed a maximum likelihood approach. As we can see, the likelihood function is only affected by the values of $S_0(t)$ and β. Let $0 = s_0 < s_1 < \ldots < s_{m+1} = \infty$ denote the ordered distinct observation time points of all time intervals $\{L_i, R_i; i = 1, \ldots, n\}$ and let $\alpha_{ij} = I(s_j \in [L_i, R_i]), j = 1, \ldots, m, i = 1, \ldots, n$. In particular, the contribution of the i-th observation to the likelihood (1) can be expressed as

$$\sum_{j=1}^{m} \alpha_{ij}[G(s_{j-1}|x_i) - G(s_j|x_i)].$$

Let $\gamma_j = log[-logS(s_j)]$. The log of the likelihood is expressed as

$$L = \sum_{i=1}^{N} log \sum_{j=1}^{m} \alpha_{ij}\{exp[-exp(x_i\beta + \gamma_{j-1})] - exp[-exp(x_i\beta + \gamma_j)]\}.$$

Then Newton-Raphson iteration can be used to get the maximum likelihood estimates (MLEs) $\hat{\gamma}, \hat{\beta}$ from the score statistic:

$$U = (\partial L/\partial\gamma', \partial L/\partial\beta').$$

Then under some regularity conditions, as $n \to \infty$, the asymptotic normality of $\hat{\beta}_n$ gives:

$$\sqrt{n}(\hat{\beta}_n - \beta) \to N(0, \Gamma^{-1}),$$

and Γ^{-1} can be estimated by fisher information matrix replacing β and α by their maximum likelihood estimators.

Among the recent work on proportional hazards model for interval-censored data, Zhang et al. (2010) developed a semiparametric MLE by using a spline based maximum likelihood approach and Heller (2011) proposed an weighted estimating equation method to estimate the regression parameter.

1.2.2.2 Additive Hazards Model

One other popular regression model for interval-censored data is called additive hazards model. For current status data, Lin et all 1998 constructed an additive hazards regression model and proposed an easy procedure to estimate the regression parameters, which do not need to estimate any nuisance parameters. Zhang et al. (2005) studied informative censoring under the same setting. Sun (2010) developed a multiple imputation procedure to estimate the parameter. To give a representative of the estimating procedures, in the following, we describe the one proposed by Wang et al. (2010).

Motivated by the idea in Lin et al. 1998, Wang et al. 2010 generalized the additive hazards model to Case II interval-censored data. They assumed that T_i has the hazard function

$$\lambda_i(t|Z_i) = \lambda_0(t) + \beta_0' Z_i(t). \tag{1.9}$$

Here λ_0 is an unknown baseline hazard function and β_0 is a p dimensional vector of regression parameters. They also modeled the two monitoring variables U and V by

Cox type hazards functions:

$$\lambda_i^U(t|Z_i) = \lambda_1(t)exp(\gamma_0' Z_i(t)),$$

$$\lambda_i^V(t|U_i, Z_i) = I(t > U_i)\lambda_2(t)exp(\gamma_0' Z_i(t)).$$

Here $\lambda_1(t)$ and $\lambda_2(t)$ denote the unspecified baseline hazards functions and γ_0 is the p dimensional regression parameter.

To estimate β_0 and γ_0, we define a counting process $N_i^{(1)} = (1 - \delta_{1i})$. Conditional on U_i, define $N_i^{(2)}(t) = \delta_{3i} I(V_i \leq t)$ when $t \geq U_i$ and $N_i^{(2)}(t) = 0$ elsewhere. The definition of $N_i^{(2)}$ indicates that V_i is only considered after U_i has been observed. By the properties of counting processes and the proportional assumption, the intensity functions of $N_i^{(1)}(t)$ and $N_i^{(2)}(t)$ have the form:

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

$$\lambda_i^{(2)}(t|U_i, Z_i) = I(t > U_i)\lambda_2(t)e^{\Lambda_0' Z_i^*(t) + \gamma_0' Z_i(t)},$$

here $Z_i^*(t) = \int_0^t Z_i(s)ds$ and $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$.

Followed the idea used in Lin 1998, they proposed the following estimating function $U_\beta(\beta, \gamma)$ to estimate β_0:

$$U_\beta(\beta, \gamma) = \sum_{i=1}^{n} [\int_0^\infty \left\{ Z_i^*(t) - \frac{S_{1,\beta}^{(1)}(t, \beta, \gamma)}{S_{1,\beta}^{(0)}(t, \beta, \gamma)} \right\} dN_i^{(1)}(t)]$$

$$+\sum_{i=1}^{n}[\int_0^\infty \left\{ Z_i^*(t) - \frac{S_{2,\beta}^{(1)}(t,\beta,\gamma)}{S_{2,\beta}^{(0)}(t,\beta,\gamma)} \right\} dN_i^{(2)}(t)].$$

Since complete data are available for γ_0, it is preferred to use the estimating function $U_\gamma(\gamma)$ for γ_0:

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

Then, we can first get the estimator of γ, $\hat{\gamma}$, by solving the equation $U_\gamma(\gamma) = 0$. later on, we can estimate β_0 through $\hat{\beta}$, which is the root of $U_\beta(\beta, \hat{\gamma}) = 0$. It can be shown that both $\hat{\beta}$ and $\hat{\gamma}$ are consistent estimator and have asymptotic normal distribution. More details of the asymptotic distribution can be found in their paper.

As we can see, they assumed that the failure time followed an additive hazards model and the observation process had a proportional hazards. By this setting, they can define a counting process based on failure time and censored time. Then the partial likelihood estimation procedure based on counting process can be used directly.

1.2.2.3 Accelerated Failure Time Model

The accelerated failure time (AFT) model is also widely used in survival analysis. There are few papers applying AFT model on interval-censored data. Huang and Wellner (1996) discussed the AFT model for both Case I and case II interval-censored data. Tian and Cai (2004) constructed a new parameter estimator and used MCMC resampling approach to obtain the point estimate of regression parameter and the estimator of variance covariance matrix.

The accelerated failure time (AFT) model specifies a linear relationship between $logT$ and Z.

$$log\ T = Z'\beta + W,$$

here β is a p dimensional regression parameter and W is an error variable with an unknown distribution function. Let $W^* = exp(W)$ and $\lambda_w(t)$ denote the hazard function of W^*. Therefore the hazard functions of T given Z have the forms

$$\lambda(t; Z) = \lambda_w(te^{-Z'\beta})e^{-Z'\beta}.$$

Then based on the set ups of interval-censored data, let F denote the distribution function of W, the likelihood function is proportional to: $L(\beta, F) = \prod_{i=1}^{n}[F(R_i(\beta)) - F(L_i(\beta))]$. Here $R_i(\beta) = log(R_i) - Z_i'\beta$ and $L_i(\beta) = log(R_i) - Z_i'\beta$. One common method to estimate the parameter β is an estimating equation approach based on linear rank statistics, which are defined as

$$S(b) = \sum_{i=1}^{n} Z_i c_i(b).$$

c_i is the weight for the sample with Z_i. More details about the estimation procedure can be found in Sun (2006).

1.2.2.4 Other regression models

Proportional Odds Model

An alternative to the proportional hazards model is the proportional odds model. It assumes that

$$log\frac{F(t|Z)}{1-F(t|Z)} = \lambda_0(t) + \beta' Z,$$

where $F(T|Z)$ is the cumulative distribution function of event time given covariate Z. λ is the baseline log odds, an unknown monotone increasing function. β represents the regression parameters.

For interval-censored data, Huang and Wellner (1997) proposed a maximum likelihood approach and established the asymptotic distribution of MLE of β by using the efficient score function. Huang and Rossini (1997) studied sieve estimation by using monotone spline functions to approximate the nuisance function.

Linear Transformation Model

All models mentioned above have a specific form of the effect of covariate. Zhang et al. (2005) presented a more general class of semi-parametric regression model, referred to the linear transformation model:

$$h(t) = \beta' Z + \epsilon,$$

where h is an unknown strictly increasing function and the distribution of ϵ is known or specified. β is a vector of regression parameter. The proportional hazards model and the proportional odds model are special cases of the linear transformation model. If ϵ follows the extreme value distribution, then the linear transformation model reduces to proportional hazards model and if ϵ follows the logistic distribution, then the linear transformation model reduces to proportional odds model. Another special case is that ϵ has standard normal distribution, the linear transformation model becomes a semi-parametric pro-bit model. Most recently, Chen and Sun (2010) considered the fitting of the model and proposed a multiple imputation approach for interval-censored data.

1.2.3 Analysis of Multivariate Interval-Censored Data

Multivariate time to event data often occur in the clinical study which involves several related events of interest. When the outcomes can not be directly observed but be measured by periodic clinical examination, the multivariate interval censored failure time data will be collected. For example, in the ACTG 181 study, researchers were interested in both blood shedding and urine shedding. But both failure times can not be exactly observed but be measured by the periodic blood test. One difficulty of inference procedure for multivariate interval-censored data is to deal with the association among failure time variables.

1.2.3.1 NPMLE of Survival Function

Sun(2006) discussed the procedure of getting NPMLE of survival function for multivariate interval censored data, which is actually an extension for univariate interval censored data. Suppose a survival study involves n independent subjects. For subject i, there exist k failure times denoted by $T_{1i}, \ldots, T_{ki}$, for $i = 1, \ldots, n$. Then the joint cumulative distribution function of failure times is $F(t_1, \ldots, t_n) = P(T_{1i} < t_{1i}, \ldots, T_{ki} < t_{ki})$. And suppose that the observed interval censored data for subject i has the form of

$$O_i = (L_{1i}, R_{1i}] \times, \ldots, \times (L_{ki}, R_{ki}]$$

.

For the determination of NPMLE of cumulative distribution function, let

$$S = \{S_l = (m_{1l}, n_{1l}] \times, \ldots, \times (m_{kl}, n_{kl}],\, l = 1, \ldots, p\}$$

denote the disjoint rectangles which contain all the possible support of the NPMLE of F. Then the likelihood function has the form

$$L(\mathbf{q}) = \prod_{i=1}^{n} = \prod_{i=1}^{n} (\sum_{j=1}^{p} \alpha_{il} q_l)$$

where $alpha_{il} = I(S_l \subseteq O_i)$ and $q_l = F(S_l)$. The NPMLE of F can be derived by maximizing the likelihood function with respect to q_l for $l = 1, \ldots, p$ under the restriction that $P_l > 0$ and $\sum_{i=1}^{p} = 1$.

1.2.3.2 Estimation of Association parameter

Here we want to discuss this problem in the context of bivariate interval censored data. To estimate the association between failure times, one common way is to assume that the joint survival function $S(t_1, t_2)$ can be written as a copula model as $S(t_1, t_2) = C_\alpha(S_1(t_1), S_2(t_2))$, here $S_1(t)$ and $S_2(t)$ are the marginal survival function for T_1 and T_2 respectively. One desirable feature of this copula model is that the marginal distributions do not depend on the choice of C_α, which is the association parameter. Therefore, one can model the marginal distribution and association structure separately. Wang et al.(2000) discuss this approach for current status data and Sun et al.(2006) considered this approach for bivariate interval censored data.

1.2.3.3 Regression Analysis of Multivariate Interval-Censored Data

Copula model could also be use in regression analysis for multivariate interval censored data. Wang (2008) proposed to use copula model to estimate regression coefficient and association parameter simultaneously. However, the two major approaches for regression analysis of multivariate interval censored data are marginal model approach and frailty model approach.

Marginal Model Approach

The marginal approach focus on the marginal distribution and leave the correlation between failure times arbitrary. Marginal model has been widely used for regression analysis of multivariate interval censored data. For example, Goggins et al. (2000) and Kim et al. (2012)considered a maximum likelihood approach based on marginal proportional hazard model. Chen et al. (2007) developed the marginal approach under the proportional odds model and Tong et al.(2008) developed such approach for fitting the additive hazard model.

Frailty Model Approach

To describe the dependence between failure times, anther commonly used approach is the frailty model, which introduces a common latent variable to characterize the correlation. One benefit of using frailty model compared to marginal approach is that it directly models the correlation between failure times.

Among others, Oakes (1989) considered the frailty model for bivariate failure time data. Hens et al. (2009) applied a frailty model to bivariate interval censored data. Chen et al.(2009) developed a frailty additive hazard model for multivariate current status data.

1.2.4 Analysis of Interval-Censored Data with informative censoring

By informative censoring, we mean that the fail time of interest T and the observation time C are dependent. As with informatively censored failure time data, the survival function of T is generally unidentifiable. Wang et al. (2012) proposed two estimates of the survival function by copula model and Frydman et al. (2009) proposed a nonparametric maximum likelihood approach. Kim et al.(2012) discussed the regression analysis with proportional hazard model under this type of data structure.

For case II informatively interval-censored data, the situation is quiet different from current status data and usually is more complicated. Since the observation time and event time are no longer independent, we can rewrite the likelihood of a single interval-censored observation as

$$Pr(L \leq T \leq R) = Pr(l \leq T \leq r | L = l, R = r) Pr(L = l, R = r).$$

Therefore, to perform the regression analysis, we need to specify a joint model for $Pr(l \leq T \leq r | L = l, R = r)$ and $Pr(L = l, R = r)$. Zhang et al.(2007) and Wang et al.(2010) discussed a joint modeling approach under additive hazards model framework, separately.

1.3 Outline of The Dissertation

The remainder of this dissertation is organized as follows. In Chapter 2, we discuss nonparametric comparison of survival functions when one observes only interval-censored failure time data (Peto and Peto, 1972; Sun, 2006; Zhao et al., 2008). For the problem, a few procedures have been proposed in the literature (Sun, 1999; Zhao and Sun, 2005; Fang et al.,2002). However, most of the existing test procedures determine the test results or p-values based on ad-hoc methods or the permutation approach. Furthermore for the test procedures whose asymptotic distributions have been derived, the results are only for the null hypothesis. In other words, no nonparametric test procedure exists with a known asymptotic distribution under the alternative hypothesis and thus can be employed to carry out the power and sample size calculation. In this chapter, a new class of generalized log-rank tests is proposed and their asymptotic distributions are derived under both null and alternative hypotheses. A simulation study is conducted to assess their performance for finite sample situations and an illustrative example is provided.

In Chapter 3, we will still focus on the nonparametric comparison of survival functions. However, we consider a situation that often occurs in practice but has not been discussed much: the comparison based on interval-censored data in the presence of unequal censoring. That is, one observes only interval-censored data and the distributions of or the mechanism behind censoring variables may depend on treatments and thus be different for the subjects in different treatment groups. For the problem, a test procedure is developed that takes into account the difference between the distributions of the censoring variables, and the asymptotic normality of the test statistics is given. For the assessment of the performance of the procedure, a simulation study is conducted and suggests that it works well for practical situations. The AIDS data mentioned above are analyzed with the proposed method.

In Chapter 4, we will discuss the regression problem for multivariate interval-censored fail time data. Multivariate interval-censored failure time data often occur in the clinical trial that involves several related event times of interest and all the event times suffer interval censoring. Different types of models have been proposed for the regression analysis. However, most of these methods only deal with the situation where observation time is independent of the underlying survival time completely or given covariates. In this chapter, we discuss regression analysis of multivariate interval-censored data when the observation time may be related to the underlying survival time. An estimating equation based approach is proposed for regression coefficient estimate with the additive hazards frailty model and the asymptotic properties of the proposed estimates are established by using counting processes. A major advantage of the proposed method is that it does not involve estimation of any baseline hazard function. Simulation results suggest that the proposed method works well for practical situations.

In Chapter 5, several future research directions are discussed.

Chapter 2

A New Class of Generalized Log Rank Tests for Interval-Censored Failure Time Data

2.1 Introduction

This chapter discusses nonparametric comparison of survival functions when one observes only interval-censored failure time data (Huang, 1999; Peto and Peto, 1972; Sun, 2006; Zhao et al., 2008). As we known, survival comparison is usually one of main goals in survival studies. For the case of right-censored failure time data, there exist a number of well-established procedures such as the weighted log-rank tests and the weighted Kaplain-Meier tests (Fleming and Harrington, 1991; Kalbfleisch and Prentice, 2002).

For the case of interval-censored failure time data, a few nonparametric test procedures have also been actually developed. For example, Finkelstein (1986) suggested a score test procedure, and Sun (1996) and Zhao and Sun (2004) generalized the log-rank test for right-censored data. However, most of the existing approaches for interval-censored data are ad-hoc generalizations of those for right-censored data and have unknown asymptotic properties (Sun, 2006). Some exceptions are the procedures proposed by Fang et al. (2002), Sun et al. (2005) and Zhao et al. (2008), in which the null asymptotic distribution of the test statistics were established. It is clear that all of these test procedures cannot be used if one intends to perform some power or sample size calculation as their asymptotic distributions under the alternative hypothesis are still unknown. In this paper, we propose a new class of test procedures whose asymptotic distributions are established under both null and alternative hypotheses.

The remainder of the chapter is organized as follows. We will begin in Section 2.2 with introducing some notation and assumptions that will be used throughout the paper and then present the new test statistics. The asymptotic distributions of the test statistics will be established in Section 2.3. In Section 2.4, we will present some numerical results obtained from a simulation study for assessing the finite sample performance of the proposed test procedures. An illustrative example is also given in Section 2.4. Section 2.5 contains some concluding remarks.

2.2 Generalized Log-rank Test Statistics

Consider a survival study that involves n independent subjects. Let T_i denote the survival time of interest for subject i, $i = 1, \ldots, n$. Suppose that for subject i, we only observe $\{U_i, V_i, \Delta_i = I(T_i \leq U_i), \Gamma_i = I(U_i < T_i \leq V_i)\}$, where U_i and V_i are non-negative random variables independent of T_i such that $U_i < V_i$ with probability one, $i = 1, \ldots, n$. This means that one only knows if T_i is smaller than U_i, between U_i and V_i, or larger than V_i. In other words, we only have interval-censored data on the T_i's. Assume that the study involves two groups, control (group 1) and treatment (group 2) groups. Let $F_1(t)$ and $F_2(t)$ denote the cumulative distribution functions of the T_i's for the subjects in the control and treatment groups, respectively. Suppose that the main goal is to compare the two groups or to test the hypothesis $H_0 : F_1(t) = F_2(t)$.

To construct the proposed test statistics, we first look at the test statistics given in Sun et al. (2005). For this, let $F(t)$ denote the common survival function under the null hypothesis H_0 and define

$$K_F(u, v, \delta, \gamma) = \delta \frac{\eta\{F(u)\} - c_0}{F(u)} + \gamma \frac{\eta\{F(v)\} - \eta\{F(u)\}}{F(v) - F(u)} + (1 - \delta - \gamma) \frac{c_0 - \eta\{F(v)\}}{1 - F(v)} \ .$$

Here η is a known function over $(0, 1)$ such that $\lim_{x \to 0} \eta(x) = \lim_{x \to 1} \eta(x) = c_0$, where c_0 is a constant. Also let $\hat{F}_n(t)$ denote the nonparametric maximum likelihood estimate of F based on all samples and S_l the set of indices for the subjects in group l, $l = 1, 2$.

To test H_0, Sun et al. (2005) proposed the following test statistic

$$U_{SZZ} = \left(\sum_{i \in S_1} K_{\hat{F}_n}(U_i, V_i, \Delta_i, \Gamma_i), \sum_{i \in S_2} K_{\hat{F}_n}(U_i, V_i, \Delta_i, \Gamma_i) \right)^T$$

and derived its null asymptotic distribution.

On the other hand, it easy to see that it would be difficult or impossible to derive the asymptotic distribution of U_{SZZ} under the alternative hypothesis partly because $\hat{F}_n$ is not well-defined if $F_1 \neq F_2$. To modify the test statistic U_{SZZ}, let n_1 and n_2 ($n_1 + n_2 = n$) denote the numbers of subjects in the control and treatment groups, respectively, and $\hat{F}_{n_1}$ and $\hat{F}_{n_2}$ the nonparametric maximum likelihood estimates of F_1 and F_2 based on the samples from the control and treatment groups, respectively. Naturally, by noting that

$$\sum_{i \in S_1} K_{\hat{F}_{n_1}}(U_i, V_i, \Delta_i, \Gamma_i) = 0$$

and

$$\sum_{i \in S_2} K_{\hat{F}_{n_2}}(U_i, V_i, \Delta_i, \Gamma_i) = 0\,,$$

one could define a new statistic as

$$\left(\sum_{i \in S_1} K_{\hat{F}_{n_2}}(U_i, V_i, \Delta_i, \Gamma_i), \sum_{i \in S_2} K_{\hat{F}_{n_1}}(U_i, V_i, \Delta_i, \Gamma_i) \right)^T .$$

by replacing $\hat{F}_n(t)$ with $\hat{F}_{n_1}$ or $\hat{F}_{n_2}$ in U_{SZZ}. However, it would still be difficult to derive the asymptotic distribution of the statistic given above.

To construct a workable test statistic, define

$$K_{F_1,F_2}(u,v,\delta,\gamma) = \delta\frac{\eta\{F_2(u)\} - c_0}{F_1(u)} + \gamma\frac{\eta\{F_2(v)\} - \eta\{F_2(u)\}}{F_1(v) - F_1(u)} + (1-\delta-\gamma)\frac{c_0 - \eta\{F_2(v)\}}{1 - F_1(v)} \ .$$

For testing the hypothesis H_0, we propose to use the statistic

$$\bar{U}_n = (\bar{U}_{n_1}, \bar{U}_{n_2})^T = \left(\frac{1}{\sqrt{n}} \sum_{i \in S_1} K_{\hat{F}_{n_1},\hat{F}_{n_2}}(U_i, V_i, \Delta_i, \Gamma_i), \frac{1}{\sqrt{n}} \sum_{i \in S_2} K_{\hat{F}_{n_2},\hat{F}_{n_1}}(U_i, V_i, \Delta_i, \Gamma_i) \right)^T .$$

In the next section, we will establish the asymptotic properties of $\bar{U}_{n_1}$ and $\bar{U}_{n_2}$ and hence present the resulting test procedure for H_0. Some comments will be given below on the determination of $\hat{F}_{n_1}(t)$ and $\hat{F}_{n_2}(t)$ as well as the selection of function η.

2.3 Asymptotic Distributions and Test Procedures

In this section, we will first establish the asymptotic distributions of $\bar{U}_{n_1}$ and $\bar{U}_{n_2}$ and then present the test procedure. For this, let H and h denote the distribution and density functions of (U_i, V_i), respectively, and λ_2 and ν_2 denote the Lebesgue measure on R^2 and counting measure on the set $\{(0,1),(1,0),(0,0)\}$, respectively. Define

$$q_F(u,v,\delta,\gamma) = h(u,v)\{F(u)\}^{\delta}\{F(v) - F(u)\}^{\gamma}\{1 - F(v)\}^{1-\delta-\gamma}$$

and similarly $q_{F_1}(u,v,\delta,\gamma)$ and $q_{F_2}(u,v,\delta,\gamma)$ with respect to $\lambda_2 \otimes \nu_2$. It is easy to see that $q_{F_l}(u,v,\delta,\gamma)$ is the density function of $(U_i, V_i, \Delta_i, \Gamma_i)$ for $i \in S_l$, $l = 1, 2$. Also for $l = 1, 2$, define $dQ_l \ = \ q_{F_l}\, d(\lambda_2 \otimes \nu_2)$ and

$$Q_{n_l}(u,v,\delta,\gamma) \ = \ \frac{1}{n_l} \sum_{i \in S_l} 1_{\{(U_i,V_i) \leq (u,v), (\Delta_i,\Gamma_i)=(\delta,\gamma)\}} \, .$$

Then we have

$$\bar{U}_{n_1} \ = \ \sqrt{n_1}\, Q_{n_1}(\, K_{\hat{F}_{n_1}, \hat{F}_{n_2}}\,) \ , \ \bar{U}_{n_2} \ = \ \sqrt{n_2}\, Q_{n_2}(\, K_{\hat{F}_{n_2}, \hat{F}_{n_1}}\,) \, .$$

For the result below, we will assume that the regularity conditions given in Groeneboom and Wellner (1992) for the strong consistency of $\hat{F}_{n_1}$ and $\hat{F}_{n_2}$ hold. Also following Sun et al. (2005), we will assume that $F_1(t)$ and $F_2(t)$ have their support in $[0, M]$ with continuous density functions, and that there exist $0 < \delta_0, \varepsilon_0 < M/2$ and $M_0 < M$ such that $Pr(U_i < \delta_0) = 0$, $Pr(U_i + \varepsilon_0 \leq V_i \leq M_0) = 1$, $0 < F_l(\delta_0) < F_l(M_0) < 1$ and $\min_{\delta_0 \leq t \leq M_0 - \varepsilon_0} [F_l(t + \varepsilon_0) - F_l(t)] \neq 0$, where M is a positive constant. These conditions usually hold for periodic follow-up studies such as clinical trials. The following theorem gives the asymptotic behavior of $\bar{U}_{n_1}$ and $\bar{U}_{n_2}$.

Theorem 1. Suppose that the assumptions described above hold and η is a bounded Lipschitz function on $[a, 1]$ for any finite positive number $a < 1$. Also suppose that as $n \to \infty$, $n_k/n \to p_k$, where $0 < p_k < 1$ and $p_1 + p_2 = 1$. Then we have, asymptotically,

$$\bar{U}_{n_1} = \sqrt{n_1}(Q_{n_1} - Q_1) \left\{ K_{F_1,F_2} - \tilde{\theta}_{g_1,F_1} \right\} + o_p(1)$$

and

$$\bar{U}_{n_2} = \sqrt{n_2}(Q_{n_2} - Q_2)\left\{K_{F_2,F_1} - \tilde{\theta}_{g_2,F_2}\right\} + o_p(1),$$

where g_l and $\tilde{\theta}_{g_l,F_l}$ are given in the Appendix.

The proof of the above theorem is sketched in the appendix. Let

$$\sigma_1^2 = Q_1\left[\left\{K_{F_1,F_2} - \tilde{\theta}_{g_1,F_1}\right\} - Q_1\left\{K_{F_1,F_2} - \tilde{\theta}_{g_1,F_1}\right\}\right]^2$$

and

$$\sigma_2^2 = Q_2\left[\left\{K_{F_2,F_1} - \tilde{\theta}_{g_2,F_2}\right\} - Q_2\left\{K_{F_2,F_1} - \tilde{\theta}_{g_2,F_2}\right\}\right]^2.$$

Define

$$S = \frac{\bar{U}_{n_1}^2/\sigma_1^2}{\bar{U}_{n_2}^2/\sigma_2^2}.$$

Then it follows from the theorem above that S has an asymptotic $F(1,1)$ distribution and furthermore, under the hypothesis H_0 and as $n \to \infty$, the distribution of $S_0 = \bar{U}_{n_1}^2/\bar{U}_{n_2}^2$ can be approximated by the $F(1,1)$ distribution. This suggests that one can carry out the test of the hypothesis H_0 by using the statistic S_0 based on the $F(1,1)$ distribution.

To implement the test procedure proposed above, one needs to determine $\hat{F}_{n_1}$ and $\hat{F}_{n_2}$ and select the function η. For the former, the simplest method is to apply the self-consistency algorithm given in Turnbull (1976). Some alternatives can be found in Sun (2006). For the latter, a common choice, which will be used below for the numerical study, is $\eta(x) = 1 - (1-x)\log(1-x)\,(1-x)^{\rho}\,x^{\gamma}$, where ρ and γ are some numbers between $[0, 1]$. More comments on this can be found in Sun et al. (2005).

As discussed above, in practice, one may be often interested in performing power calculation. For this based on the test procedure given above, for the given significance level α, let Z denote the random variable following the $F(1, 1)$ distribution and F_L and F_U be defined such that

$$P(Z < F_L) = \alpha/2 \quad \text{and} \quad P(Z > F_U) = \alpha/2\,.$$

Then the asymptotic power is given by

$$F_{1,1}\left(\frac{\sigma_2^2}{\sigma_1^2}F_L\right) + 1 - F_{1,1}\left(\frac{\sigma_2^2}{\sigma_1^2}F_U\right)$$

if F_1 and F_2 are known.

2.4 Numerical Studies

Now we report some results obtained from a simulation study conducted to assess the finite sample performance of the class of test procedures proposed in the previous sections and its application to a real set of interval-censored data. For the simulation study, we assumed that half of subjects are from the control group and the other half from the treatment group. To generate the survival times of interest, we considered two set-ups. One is to assume that T_i follows the exponential distribution with the mean

$\exp(\alpha + \beta z_i)$, where z_i is the treatment indicator, being equal to 0 for the subjects in the control group and 1 otherwise. The other is to generate T_i from the gamma distribution with the shape parameter equal to 2 and the scale parameter $1/(\alpha + \beta z_i)$.

To generate the censoring interval for subject i, we first generated U_{i1} and U_{i2} independently from the uniform distribution over $(1, \theta_1)$ and $(1, \theta_2)$, respectively. Here θ_1 and θ_2 are some positive constants chosen to give the desired percentages of left-censored, interval-censored and right-censored observations. Given U_{i1} and U_{i2}, we defined U_i to be the nearest integer to U_{i1} and V_i the nearest integer to the maximum of $U_{i1} + 1$ and $U_{i1} + U_{i2}$. Also we assumed that the study ended at $t = 10$ and thus defined V_i to be 10 if the V_i generated above is larger than 10. The results given below are based on 1000 replications.

Table 2.1 presents the empirical or estimated size and power of the proposed test procedure based on the simulated data generated from the exponential distribution with $\alpha = 2$, $\beta = -3, -2, -1.5, 0, 1.5, 2$ or 3. Here we used the η function given in Section 2.3 with different values of ρ and γ and the self-consistency algorithm for the determination of the maximum likelihood estimates $\hat{F}_{n_1}$ and $\hat{F}_{n_2}$. In the table, the first column gives the percentages of left-censored, interval-censored and right-censored observations in the generated data, which are roughly $(20\%, 20\%, 60\%)$ and $(17\%, 16\%, 67\%)$ for the two situations considered here. The results obtained under the gamma distribution are given in Table 2.2 and here we took $\alpha = 1$ and the same values for β as in Table 2.1. One can see from both Tables 2.1 and 2.2 that the proposed test procedure seems to give right size and have good power for the situations considered here.

To illustrate the proposed approach, we apply it to the set of interval-censored data discussed in Goggins and Finkelstein (2000) and Sun (2006) among others. The data arose from an AIDS clinical trial concerning the opportunistic infection cytomegalovirus (CMV). During the study, among other activities, blood and urine samples were collected from the patients at their clinical visits and tested for the presence of CMV, which is also commonly referred to as shedding of the virus. These samples and tests provide observed information on the two variables, the times to CMV shedding in blood and urine, respectively. The study consists of 204 patients who provided at least one urine and one blood samples during the study. For some patients, their shedding had already occurred at their first clinical visits or they had not yet started shedding by the end of the study, giving either left- or right-censored observations on their shedding times. For the other patients, their shedding times were observed to belong to some intervals given by the last negative and first positive blood or urine test, respectively.

In addition to the observed information about CMV shedding times in blood and in urine, the study also provided the range of each patient's baseline CD4 cell count. In particular, the patients were classified into two groups: these with their baseline CD4 cell counts less than 75 (cells/μl) and the others. Note that the CD4 cell count indicates the status of a person's immune system and is commonly used to measure the stage of HIV infection. For this data set, one problem of interest is to compare the two groups of patients with respect to their CMV shedding times. For this, we applied the test procedure developed in the previous sections to the data on the times to CMV shedding in blood and urine separately and the obtained results are presented in Table 2.3. They indicate that the CMV shedding times in both blood and urine were significantly different for the two groups of patients, especially in urine. In other words, the CMV shedding time seems to be significantly related to the baseline CD4 cell count and these results are similar to those obtained by others.

2.5 Concluding Remarks

This chapter discussed the nonparametric comparison of survival functions when only interval-censored failure time data are available. For the problem, a class of nonparametric tests was proposed and both finite sample and asymptotic properties of the presented approach were established. One major advantage of the proposed test procedure is that its asymptotic distribution is known under both null and alternative hypotheses, which makes both power and sample size calculation possible. In contrast, for all existing nonparametric test procedures, their asymptotic distribution is either unknown or known only under the null hypothesis. Note that another shortcoming for some existing test procedures is that the estimation or determination of the variance of the test statistics involve the dealing of high dimension matrices, which makes them unstable. It is easy to see that the proposed test procedure does not have the same problem.

It should be noted that there exist some limitations about the proposed nonparametric test procedures. One is that in the previous sections, it was assumed that no exact observation of survival time is observed. Although this may not be true in general, it holds in many situations such as studies with periodic follow-ups. Also one can apply the procedure if the distributions of interest have only finite support points. Of course, it would be useful to generalize the proposed approach to situations where observed data include both exact and interval-censored observations on the survival time of interest.

Another limitation of the proposed approach is that we only considered the situation where the distributions generating censoring intervals are same for the subjects in different treatment groups. Sometimes this may not be true as, for example, the subjects in different treatment groups may have different follow-up patterns in a periodic follow-up study. One specific example of this is given by a clinical trial in which patients receiving placebo treatment may feel worse compared to other patients and thus visit doctors more often. Among others, Sun (1999) discussed this problem for current status data, a special case of interval-censored data. However, there does not seem to exist a nonparametric test procedure similar to the one proposed here for this later situation.

Chapter 3

Nonparametric Comparison of Survival Functions Based on Interval-Censored Data With Unequal Censoring

3.1 Introduction

This chapter again discusses nonparametric comparison of survival functions when one observes only interval-censored failure time data. One common drawback or restriction of the nonparametric test procedures mentioned above and many given in the literature is that they can apply only to situations where the censoring variables follow the same distribution for the subjects in different treatment groups. In other words, they require that the censoring mechanism is the same for all subjects. In practice, however, this may not be true.

In the case of current status data, for example, the single observation time on each subject may the death time and depend on the treatment. It is easy to see that in these situations, the use of the test procedures that fail to take into account this fact could result in misleading or wrong results such as seriously overestimating or underestimating the treatment difference. Among others, Sun (1999) and Zhu et al.(2008) addressed this issue and gave some nonparametric test procedures that allow the dependence of the distributions of censoring variables on treatments. However, the former is only for current status data and the latter relies on some condition that may be restrictive in practice. In the following, we present a new test procedure that applies to more general situations.

The remainder of this chapter is organized as follows. We will first introduce some notation and assumptions that will be used throughout the chapter in Section 3.2. Section 3.3 will then present the new test procedure developed by using the same idea used in Sun (1999). It can be seen as a generalization of the procedure given in Sun (1999) and includes that proposed in Zhu et al. (2008) as a special case. Also in Section 3.3, the asymptotic distribution of the test statistic is given. Section 3.4 gives some results obtained from a simulation study conducted to evaluate the performance of the proposed approach and they indicate that it works well in practice. In Section 3.5, an illustrative example is presented and Section 3.6 contains some discussion and concluding remarks.

3.2 Notation and Assumptions

Consider a failure time study that involves n independent subjects and $p+1$ treatments. For subject i, let T_i denote the survival time of interest and assume that the observed information on T_i is given by

$$\{ U_i, V_i, \Delta_{1i} = I(T_i \leq U_i), \Delta_{2i} = I(U_i < T_i \leq V_i) \},$$

where U_i and V_i with $U_i \leq V_i$ are two random variables representing two observation times on the subject, $i = 1, \ldots, n$. That is, we have interval-censored data on the T_i's. It is easy to see that $\Delta_{1i} = 1$ means that the observation on T_i is left censored, while $\Delta_{3i} = 1 - \Delta_{1i} - \Delta_{2i} = 1$ corresponds to a right-censored observation on T_i. Define Z_i to be the p dimensional treatment indicator vector whose l element being equal to 1 if subject i is given treatment l and 0 otherwise, $l = 1, ..., p$, and let $S_j(t)$ denote the survival function of the subjects given treatment j, $j = 1, ..., p+1$. Then the observed data are $\{ U_i, V_i, \Delta_{1i}, \Delta_{2i}, Z_i; i = 1, ..., n \}$. Our goal is test the hypothesis

$$H_0 \, : \, S_1(t) \, = \, ... \, = \, S_{p+1}(t) \, .$$

In the following, we will assume that the distributions of U_i's and V_i's may depend on the treatment indicator Z_i, but they are independent of the survival time T_i given Z_i. To model the dependence, following Wang et al. (2010), we will assume that the hazard functions of U_i and V_i have the form

$$\lambda_i^U(t|Z_i) \,=\, \lambda_1(t)\, \exp(\gamma_1' Z_i) \tag{1}$$

and

$$\lambda_i^V(t|Z_i) \,=\, I(t > U_i)\, \lambda_2(t)\, \exp(\gamma_2' Z_i)\,, \tag{2}$$

respectively. In the above, $\lambda_1(t)$ and $\lambda_2(t)$ denote some unknown baseline hazard functions and γ_1 and γ_2 are vectors of regression parameters. Under models (1) and (2), the baseline survival functions of U_i and V_i have the forms

$$S_1(t) \,=\, \exp\left\{ -\int_0^t \lambda_1(s)\, ds \right\} =: \exp\{\, -\Lambda_1(t)\}\,,$$

and

$$S_2(t) \,=\, \exp\left\{ -\int_0^t I(s > U_i)\, \lambda_2(s)\, ds \right\} =: \exp\{\, -\Lambda_2(t)\}\ ,$$

respectively.

Before ending this section and presenting the proposed test statistic in the next section, we will briefly review the test statistic developed by Sun (1999) for current status data. For this, define $N_i(t) \,=\, I(t \geq T_i)$, $\tilde{N}_{1i}(t) \,=\, I(t \geq U_i)$, and $\tilde{N}_{2i}(t) \,=\, I(t \geq V_i)$ if $t > U_i$ and 0 otherwise, $i \,=\, 1, ..., n$. Then the hypothesis H_0 is equivalent to the hypothesis

$$H_0^* \,:\, E\{\, N_i(t) | Z_i \,\} \text{ is independent of } Z_i\,.$$

To test H_0^*, we pretend that one only observes current status data $\{\, U_i, \Delta_{1i} \,\}$ and in this case, Sun (1999) suggested to employ the test statistic

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} K_1(U_i, Z_i, \gamma_1) \;=\; \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Z_i - \bar{Z})\, \frac{N_i(U_i)\, \exp(-\gamma_1' Z_i)}{\hat{S}_1(U_i, \gamma_1)^{\exp(\gamma_1' Z_i)}}$$

assuming that γ_1 is known. In the above, $\bar{Z} \;=\; \sum_{i=1}^{n} Z_i \,/ n$ and

$$\hat{S}_1(t, \gamma_1) \;=\; \exp\left\{ -\int_0^t \frac{d\tilde{N}_1(s)}{\sum_{j=1}^{n} I(s \le U_j) \exp(\gamma_1' Z_j)} \right\} = \exp\Big(-\hat{\Lambda}_1(t, \gamma_1) \Big)\;,$$

where $\tilde{N}_1(t) \;=\; \sum_{i=1}^{n} \tilde{N}_{1i}(t)$. In the next section, we will generalize the procedure above to interval-censored data situations.

3.3 Nonparametric Test Procedure for Interval-censored Data

In this section, we will use the same notation as before. To test H_0, motivated by the procedure given in Sun (1999), we propose to use the statistic

$$U(\gamma_1, \gamma_2) \;=\; \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{ K_1(U_i, Z_i, \gamma_1) \,+\, K_2(U_i, V_i, Z_i, \gamma_2) \}$$

if γ_1 and γ_2 are known. In the above,

$$K_2(U_i, V_i, \gamma_2) \;=\; (Z_i - \bar{Z})\, \frac{N_i(V_i)(1-\Delta_{1i})\exp(-\gamma_2' Z_i)}{\hat{S}_2(V_i, \gamma_2)^{\exp(\gamma_2' Z_i)}}\,,$$

where

$$\hat{S}_2(t, \gamma_2) \;=\; \exp\left\{ -\int_0^t \frac{d\tilde{N}_2(s)}{\sum_{j=1}^n \, I(U_j < s \le V_j)\exp(\gamma_2' Z_j)} \right\} = \exp\Big(-\hat{\Lambda}_2(t, \gamma_2) \Big)$$

with $\tilde{N}_2(t) \;=\; \sum_{i=1}^n \, \tilde{N}_{2i}(t)$.

In practice, of course, γ_1 and γ_2 are generally unknown and need to be estimated. For this, again by following Wang et al. (2010), one can obtain consistent estimates by solving the following equations

$$U_1(\gamma_1) \;=\; n^{-\frac{1}{2}} \sum_{i=1}^n \int_0^\infty \left\{ Z_i - \frac{S_1^{(1)}(t, \gamma_1)}{S_1^{(0)}(t, \gamma_1)} \right\} d\tilde{N}_{1i}(t)$$

and

$$U_2(\gamma_2) \;=\; n^{-\frac{1}{2}} \sum_{i=1}^n \int_0^\infty \left\{ Z_i - \frac{S_2^{(1)}(t, \gamma_2)}{S_2^{(0)}(t, \gamma_2)} \right\} d\tilde{N}_{2i}(t)\,,$$

respectively. In the above,

$$S_1^{(j)}(t, \gamma_1) \;=\; n^{-1} \sum_{i=1}^n I(t \le U_i)\, \exp(\gamma_1' Z_i)\, Z_i^j$$

and

$$S_2^{(j)}(t, \gamma_2) \;=\; n^{-1} \sum_{j=1}^{n} I(U_i < t \leq V_i)\, \exp(\gamma_2' Z_i)\, Z_i^j \,,$$

$j \;=\; 0, 1$. Let $\hat{\gamma}_1$ and $\hat{\gamma}_2$ denote the estimates of γ_1 and γ_2 defined above. Then it is natural to perform the testing of the hypothesis H_0 based on the statistic $U(\hat{\gamma}_1, \hat{\gamma}_2)$.

To derive the asymptotic distribution of $U(\hat{\gamma}_1, \hat{\gamma}_2)$, define

$$U_3(\gamma_1, \gamma_2) \;=\; \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} K_1'(U_i, Z_i, \gamma_1), \frac{1}{\sqrt{n}} \sum_{i=1}^{n} K_2'(U_i, V_i, Z_i, \gamma_2) \right\}' .$$

Then we have $U(\gamma_1, \gamma_2) \;=\; C\, U_3(\gamma_1, \gamma_2)$, where $C \;=\; [I_p, I_p]$ with I_p denoting the $p \times p$ identity matrix. Let γ_{10} and γ_{20} denote the true values of γ_1 and γ_2, respectively, and define $U_4(\gamma_1, \gamma_2) \;=\; (U_1'(\gamma_1), U_2'(\gamma_2))'$. Note that asymptotically we have

$$U_3(\hat{\gamma}_1, \hat{\gamma}_2) \;\approx\; U_3(\gamma_{10}, \gamma_{20}) \;+\; \frac{\partial U_3(\gamma_{10}, \gamma_{20})}{\partial \gamma} \left\{ -\frac{\partial U_4(\gamma_{10}, \gamma_{20})}{\partial \gamma} \right\}^{-1} U_4(\gamma_{10}, \gamma_{20})$$

by the Taylor series expansion and the consistency of $\hat{\gamma}_1$ and $\hat{\gamma}_2$, where $\gamma \;=\; (\gamma_1', \gamma_2')'$. In the Appendix, we will show that one can approximate the joint distribution of $U_3(\gamma_{10}, \gamma_{20})$ and $U_4(\gamma_{10}, \gamma_{20})$ by the multivariate normal distribution with mean zero and the covariance matrix $\Gamma(\gamma_{10}, \gamma_{20})$ given in the Appendix. It thus follows that $U_3(\hat{\gamma}_1, \hat{\gamma}_2)$ asymptotically have the normal distribution with mean zero and the covariance matrix that can be consistently estimated by $A(\hat{\gamma}_1, \hat{\gamma}_2)\, \Gamma(\hat{\gamma}_1, \hat{\gamma}_2)\, A'(\hat{\gamma}_1, \hat{\gamma}_2)$, where

$$A(\gamma_1, \gamma_2) \;=\; \left(I_{2p}, \frac{\partial U_3(\gamma_1, \gamma_2)}{\partial \gamma} \left(-\frac{\partial U_4(\gamma_1, \gamma_2)}{\partial \gamma} \right)^{-1} \right) .$$

with I_{2p} denoting the $2p \times 2p$ identity matrix. This proves that one can approximate the distribution of $U(\hat{\gamma}_1, \hat{\gamma}_2)$ by the multivariate normal distribution with mean zero and the covariance matrix

$$V(\hat{\gamma}_1, \hat{\gamma}_2) \,=\, C\, A(\hat{\gamma}_1, \hat{\gamma}_2)\, \Gamma(\hat{\gamma}_1, \hat{\gamma}_2)\, A'(\hat{\gamma}_1, \hat{\gamma}_2)\, C' \,.$$

Hence one can test the hypothesis H_0 by using the statistic $X \,=\, U'(\hat{\gamma}_1, \hat{\gamma}_2)\, V^{-1}(\hat{\gamma}_1, \hat{\gamma}_2)\, U(\hat{\gamma}_1, \hat{\gamma}_2)$ based on the χ^2 distribution with the degrees of freedom p.

3.4 A Simulation Study

In this section, we report a simulation study conducted for the assessment of the finite sample performance of the test procedure proposed in the previous section. In the study, we considered the two sample comparison problem ($p \,=\, 1$) and assumed that the two treatment groups have the same number of subjects. To generate the survival times of interest, two distributions were used. One is to generate the T_i's from the exponential distribution with mean $\exp(\alpha + \beta Z_i)$ and the other is to assume that the T_i's follow the Gamma distribution with the shape parameter equal to 2 and the scale parameter $exp(\alpha + \beta Z_i)$. In both case, we took $\alpha \,=\, 2.5$. To generate censoring intervals, we first generated two random numbers U_{i1} and U_{i2} independently from the exponential distributions with means $1 \,/\, \exp(\lambda_1 + \gamma_1 Z_i)$ and $1 \,/\, \exp(\lambda_2 + \gamma_2 Z_i)$, respectively. Given U_{i1} and U_{i2}, we defined $U_i \,=\, U_{i1}$ and $V_i \,=\, U_{i1} + U_{i2}$. Here we chose $\lambda_1 = 0.1$ and $\lambda_2 = 0.05$ for the exponentially distributed T_i and $\lambda_1 = 0.06$ and $\lambda_2 = 0.05$ otherwise. The results given below are based on 1000 replications.

Table 3.1 presents the estimated size and power of the proposed test procedure based on the simulated data generated from the exponential distribution with $\alpha = 2.5$, $\beta = -4, -3.5, -2.5, 0,\ 2.5, 3.5$ or 4, $n = 150$, 300 or 400, and $\gamma_1, \gamma_2 = 0$, 0.1 or 0.2, respectively. The results show that the proposed test procedure seems to have the right size and reasonable power for the situations considered here. Also as expected, the estimated power increased when the sample size increased. The results obtained under the Gamma distribution are given in Table 3.2 with all other set-ups being the same as in Table 3.1. Again they seem to give similar conclusions.

To assess the normal approximation to the distribution of the test statistic, we investigated the quartile plots of the standardized test statistic against the standard normal variable. Figures 3.1 and 3.2 display these plots corresponding to the data given in Tables 3.1 and 3.2, respectively. They indicate that the normal approximation seems quite reasonable.

In practice, it would be interesting to know if one can still apply a test procedure that requires the equal censoring to the situation where there exists unequal censoring. To see this, we compared the proposed test procedure to the generalized log-rank test procedure developed by Sun et al. (2005). Table 3.3 gives the estimated sizes of the two methods based on the simulated data generated as in Table 3.1 with $n = 100$, $\alpha = 1.5$, $\lambda_1 = 0.2$ and $\lambda_2 = 0.2$. It is easy that the test given in Sun et al. (2005) could underestimated the size.

3.5 An Application

To illustrate the proposed test procedure, we apply it to the set of interval-censored data discussed in Goedert et al. (1989) and Sun (2006) among others. The data arose from a study aiming to investigate the HIV-1 infection risk among hemophilia patients. During the study, the patients are assigned to two different groups based on the average annual dose of the blood they received. The survival time of interest is the patients HIV-1 infection time for which only interval-censored data are available. The study consists of 368 patients and 236 of them received no factor VIII concentrate and 132 of them received low dose factor VIII concentrate ($< 20,000U$). The time units is quarters. Since the data are given in the format of $[L_i, R_i]$, we need to make some adjustment. Specifically, for the left censored subjects ($L_i = 0$), we took $U_i = R_i$ and V_i to be the largest observation time in the study. For the right censored subjects ($R_i = \infty$), we defined $V_i = L_i$ and U_i to be the smallest observation time in the study. For the subjects with interval-censored observations, we took $U_i = L_i$ and $V_i = R_i$. Then we could apply the proposed method to the adjusted data.

For the analysis, we first need to check the distributions of the observation times U_i's and V_i's. To do this graphically, we obtained the Kaplan-Meier estimates of the survival functions of the U_i's and V_i's separately and presented them in Figure 3.3. It seems that the first observation times have the same distribution between the two groups but the distributions of the second observation times are quite different between the two groups. We then fitted the models (1) and (2) separately and obtained $\hat{\gamma}_1 = 0.26$ and

$\hat{\gamma}_2 = , 0.578$ with the p-values testing the parameters equal to zero being 0.229 and almost zero, respectively. They suggest that there exists unequal censoring.

The application of the proposed test procedure yielded $X = 8.8351$, giving the p-value of 0.0030 based on the χ^2-distribution with the degree of freedom one. This indicates that the HIV-1 infection risks between the two groups were significantly different and the Factor VIII blood concentrate significantly increased the risk of infect $HIV-1$. This result is similar to that given in Sun (2006), which gave a p-value close to zero. The difference between the p-values may be because Sun (2006) did not consider the difference between the distributions of observation times or unequal censoring.

3.6 Concluding Remarks

This chapter discussed the nonparametric treatment comparison based on interval-censored failure time data with unequal censoring and presented a new test procedure for the problem. As discussed above, many procedures have been developed for the problem with equal censoring but the simulation study showed that they are not applicable to the situation considered here. The proposed test procedure can be regarded as a generalization of the method given proposed in Sun (1999) and its asymptotic distribution was established. The simulation study indicated that the method seems to perform reasonably well in practical situations.

We remark that the test procedure presented above applies to more general situations than the two sample test procedure proposed in Zhu et al. (2008). To see that, assume that $p = 1$, that is, there exist only two treatment groups. Let f_{U1} and f_{V1} denote the density functions of U and V for the subjects in treatment group 1 and f_{U2} and f_{V2} for the subjects in treatment group 2. Also let $Z = 0$ for the subjects in treatment group 1 and 1 otherwise. The procedure given in Zhu et al. (2008) requires that

$$\frac{f_{U2}(u)}{f_{U1}(u)} = \frac{f_{V2}(u)}{f_{V1}(u)}$$

for all u. One can easily show that models (1) and (2) give this condition if taking $\lambda_1(t) = \lambda_2(t)$ and $\gamma_1 = \gamma_2$.

Chapter 4

Regression Analysis of Multivariate Interval-Censored Data With Informative Censoring

4.1 Introduction

This chapter discusses regression analysis of multivariate interval-censored data when there exists the informative censoring issue. Multivariate interval-censored data often occurs in clinical trials that involves several related event times of interest and all the event times suffer interval censored. A number of authors have studied the regression analysis of multivariate interval-censored data (Zhang et al.(2008); Tong et al.(2008); Chen et al.(2009); Sun (2006)). However, most of methods need the assumption that the observation time is independent of the event time given covariate. In this chapter, we consider the situation where the observation time may depend on the event time of interest, which is often referred to informative censoring. An example of multivariate interval-censored data arises from an AIDS clinical trial conducted by AIDS Clinical Trial Group (ACTG) 181 (Goggins and Finkelstein (2000)).

One major difficulty for the analysis of multivariate failure time data compared to univariate failure time data is to deal with the association between related failure time variables. For multivariate interval-censored data, three major approaches have been developed for regression analysis. Wang et al. (2008) proposed to use copula model to estimate regression coefficient and association parameter jointly with the marginal proportional hazard model. Zhang et al.(2008) applied it to the marginal proportional odds model. Another commonly used approach is the marginal model approach, which focus on the marginal distribution and leave the correlation between failure times arbitrary. Chen et al. (2007) developed a marginal approach by using proportional odds model and Tong et al.(2008) developed such approach for additive hazard model. To describe the dependence between correlated failure time, people normally would employ the frailty model and introduce an latent variable to characterize the correlation. One benefit of using frailty model compared to marginal approach is that it directly models the correlation between failure times. Among others, Chen et al.(2009) developed a frailty additive hazard model for multivariate current status data. Nielsena et al.(2009) model multivariate failure time data by composite likelihood of pairwise frailty likelihoods and marginal hazards using natural cubic splines. Hens et al. (2009)

applied a frailty model to bivariate interval censored data.

Informative censoring is one of the severe issues that often occurs in failure time data analysis. For instance, some symptoms may happen before the failure time and patient with certain symptoms are more likely to visit the doctor than other patients. Under that situation, the observation processes are related to failure time process. For the regression analysis of informative current status data, Frydman et al. (2009) proposed a nonparametric maximum likelihood approach. Kim et al.(2012) discussed the regression analysis with proportional hazard model and Zhang et al.(2007) applied the additive hazard model and modeled the dependence through an unobservable random effect. For case II informatively interval-censored data, the situation is quiet different from current status data and usually is more complicated. Finkelstein et al.(2002) investigated univariate interval-censored data in the presence of dependent interval censoring and Wang et al.(2010) discussed a joint frailty modeling approach under additive hazards model framework.

In this chapter, we consider the regression analysis of multivariate interval censored data with informative censoring. Additive hazard model has been used and a latent variable was introduced in order to directly characterize the correlation between failure time and the dependence between failure time and observed time. The remainder of the chapter is organized as follows. Section 4.2 introduces notation, underlying model as well as the parameter estimate procedure for informatively multivariate current status data. Section 4.3.1 introduces notation and the fitted model for informatively multivariate interval-censored data. Section 4.3.2 describes the estimation procedure based on counting process framework. Simulation results are given in Section 4.4 that suggest that the proposed method works well for practical situations. In Section 4.5, we applied the proposed model to a real data set and Section 4.6 contains some concluding remarks.

4.2 Multivariate Current Status Data

In this section, we will discuss the regression analysis of multivariate current status data when the censoring time are dependent with failure time. Consider a clinical study that involves n independent subjects. For subject i, K failure time variables $T_{i1}, ..., T_{iK}$ have been observed, $i = 1, ..., n$. Furthermore, assume that only current status data are available for failure time $T_{ij}'s$ and let $C_{ij}'s$ denote the observation time for $T_{ij}'s$. Let $Z_i(t)$ denote a p-dimensional time-dependent covariate. To model the covariate effect, we assume that T_{ij} follows the additive hazard model

$$\lambda_{ij}(t|Z_i(s),\, b_i(s),\, s \leq t) \,=\, \lambda_{0j}(t) \,+\, b_i(t) \,+\, \beta' Z_i(t), \tag{4.1}$$

given $Z_i(t)$ and the latent variable $b_i(t)$, which could be time-dependent too. In the additive hazard model, $\lambda_{0j}(t)$ is the unknown baseline hazard function and β denotes the p-dimensional regression coefficient. Note that, for simplicity, the covariate effects β are assumed to be identical for all T_{ij}. For the observation time C_{ij}, we assume that the hazard function of C_{ij} has the form

$$\lambda_{ij}^{c}(t|Z_i(t),\, b_i(s),\, s \leq t) \;=\; \lambda_{1j}(t)\, exp\,\{\gamma' Z_i(t) \;+\; b_i(t)\}, \tag{4.2}$$

condition on $Z_i(t)$ and $b_i(t)$. Here $\lambda_{1j}(t)$ is an unknown baseline hazard function and γ denotes regression parameter. It can be seen that T_{ij} and C_{ij} are related through $b_i(t)$. For the rest part of this chapter, we assume that $b_i(t)'s$ are arbitrary processes with mean zero and given them, the $T_{ij}'s$ and $C_{ij}'s$ are independent.

Suppose the observed data are given by $\{(C_{ij}, \delta_{ij} = I(C_{ij} \leq T_{ij}), Z_i(t), t \leq C_{ij})\}$. To estimate the relative risk of failure time β and the relative risk of observed time γ, define a counting process $N_{ij}(t) \;=\; I\{C_{ij} \leq \min(T_{ij}, t)\}$, which equals 1 if the subject has not experience the event yet, and 0 otherwise. Also define a counting process for the observation time C_{ij}, $Y_{ij}(t) = I(C_{ij} \geq t)$. Motivated by Zhang et al. (2005) who considered univariate current status, one can estimate β and γ by using the estimating equations $U_\beta(\beta,\gamma) = 0$ and $U_\gamma(\beta,\gamma) = 0$, where

$$U_\beta(\beta,\gamma) \;=\; \sum_{j=1}^{K} \sum_{i=1}^{n} \int_0^\infty \left\{ Z_i^*(t) \;-\; \frac{S_\beta^{(1j)}(\beta,\gamma,t)}{S^{(0j)}(\beta,\gamma,t)} \right\} dN_{ij}(t),$$

and

$$U_\gamma(\beta,\gamma) \;=\; \sum_{j=1}^{K} \sum_{i=1}^{n} \int_0^\infty \left\{ Z_i(t) \;-\; \frac{S_\gamma^{(1j)}(\beta,\gamma,t)}{S^{(0j)}(\beta,\gamma,t)} \right\} dN_{ij}(t).$$

Here, $Z_i^*(t) = \int_0^t Z_i(s)ds$ for $i \;=\; 1,\ldots,n$, $j \;=\; 1,\ldots,K$. Also, $S^{0j}(\beta,\gamma,t)$, $S_\beta^{1j}(\beta,\gamma,t)$

and $S_{\gamma}^{1j}(\beta, \gamma, t)$ are

$$S^{0j}(\beta, \gamma, t) = \frac{1}{n} \sum_{i=1}^{n} Y_{ij}(t)\, e^{\{\gamma' Z_i(t) - \beta' Z_i(t)\}},$$

$$S_{\beta}^{1j}(\beta, \gamma, t) = \frac{1}{n} \sum_{i=1}^{n} Y_{ij}(t) Z_i^*(t)\, e^{\{\gamma' Z_i(t) - \beta' Z_i(t)\}},$$

and

$$S_{\gamma}^{1j}(\beta, \gamma, t) = \frac{1}{n} \sum_{i=1}^{n} Y_{ij}(t) Z_i(t)\, e^{\{\gamma' Z_i(t) - \beta' Z_i(t)\}}.$$

To get the asymptotic distribution of the parameter estimate, let $\hat{\theta} = (\hat{\beta}', \hat{\gamma}')$ denote the estimate of $\theta = (\beta', \gamma')$. It can be shown that $\hat{\theta}$ is a consistent estimate of θ_0, the true value of θ. Moreover, $\sqrt{n}(\hat{\theta} - \theta_0)$ has asymptotical normal distribution with mean zero and covariance matrix

$$V(\theta) = -\frac{\partial U(\theta)}{\partial \theta}$$

at $\theta = \hat{\theta}$,where $U(\theta) = (U_{\beta}(\beta, \gamma), U_{\gamma}(\beta, \gamma))$.

4.3 Multivariate Interval-Censored Data

4.3.1 Notation and Models

For multivariate interval-censored data, the observed information can be written as $\{Z_i, U_{ij}, V_{ij}, \delta_{1i}^{j}, \delta_{2i}^{j}\}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, K$. Let $Z_i(t)$ denote a p-dimensional time-dependent covariance for subject i,. U_{ij} and V_{ij} are the observation times and

δ^j_{1i} and δ^j_{2i} are the censoring indicators for event time T_{ij}. $\delta^j_{1i} = I(T_{ij} < U_{ij})$ and $\delta^j_{2i} = I(U_{ij} \leq T_{ij} < V_{ij})$. Under informative censoring condition, we need to assume that the distributions of $U'_{ij}s$ and $V'_{ij}s$ depend on the survival time $T'_{ij}s$ given Z_i. In order to model the dependence, following Wang et al. (2010), we assumed that the hazard functions of T_{ij}, U_{ij} and V_{ij} have the form of

$$\lambda^T_{ij}(t|\, Z_i(s), b_i(s), s \leq t) \;=\; \lambda_{0j}(t) \,+ \beta'_j Z_i(t) + b_i(t), \tag{4.3}$$

$$\lambda^U_{ij}(t|Z_i(s), b_i(s), s \leq t) \;=\; \lambda_{1j}(t)\, \exp(\gamma'_1\, Z_i(t) \,+\, b_i(t)), \tag{4.4}$$

and

$$\lambda^V_{ij}(t|U_{ij} = u_{ij}, Z_{ij}(s), b_{ij}(s), s \leq t) \;=\; I(t > U_i)\, \lambda_{2j}(t)\, \exp(\gamma'_2 Z_i(t) + b_i(t)), \tag{4.5}$$

respectively, for $i = 1\ldots, n$ and $j = 1\ldots, K$. In the above, $\lambda_{0j}(t)$ $\lambda_{1j}(t)$ and $\lambda_{2j}(t)$ denote some unknown baseline hazard functions for T_{ij}, U_{ij} and V_{ij}. The β'_js and γ'_js are the vectors of regression parameters.

The motivation of choosing additive hazard model to fit the failure time T_{ij} is that, if the hazard function of T_{ij} conditional on the latent variable b_i has additive format, then the marginal hazard function of T_{ij} with respect to b_i also has the similar additive format. It can be shown that the marginal hazard function of T_{ij} has additive hazard format as below,

$$\lambda_{ij}^T(t|Z_i(s), s \leq t) \;=\; \lambda_{0j}^*(t) + \beta' Z_i(t), \tag{4.6}$$

with

$$\lambda_{0j}^*(t) \;=\; \frac{\partial\, \Gamma_{0j}^*(t)}{\partial\, t},$$

and

$$\Gamma_{0j}^*(t) \;=\; \int_0^t \lambda_{0j}(s)ds \;-\; log[E(e^{(-\int_0^t b_i(s)ds)})].$$

As you can see from (4.6) that the latent variable are all included in the nuisance parameter, which do not need to be estimated. This nice feature of the additive hazard model could largely simplify the parameter estimating procedure as well as its asymptotic distribution derivation .

4.3.2 Estimation of Regression Parameter

In order to estimate the regression parameter $\beta' s$ and $\gamma' s$, let us first start with defining some notation. Let $\delta_{3i}^j = 1 - \delta_{1i}^j - \delta_{2i}^j$, which indicates whether subject i is right-censored or not. Then we want to define several counting processes. $N_{ij}^{(1)}(t) = (1 - \delta_{1i}^j)I(U_{ij} \leq t)$ and $N_{ij}^{(2)}(t) = \delta_{3i}^j I(U_{ij} \leq V_{ij} \leq t)$ are the counting process for the event time T_{ij}. $\tilde{N}_{ij}^{(1)} = I(U_{ij} \leq t)$ and $\tilde{N}_{ij}^{(2)} = I(V_{ij} \leq t)$ are for the observation time. According to the definition, the intensity functions of the four counting processes can be written as:

$$P(dN_{ij}^{(1)} = 1) \; = \; E_b[e^{-\int_0^t \{\lambda_{0j}(s)+b_i(s)\}ds+b_i(t)}]\,\lambda_{1j}(t)\,e^{\{-\beta_0' Z_i^*(t)+\gamma_1' Z_i(t)\}},$$

$$P(dN_{ij}^{(2)} = 1) \; = \; I(t \geq U_{ij})E_b[e^{-\int_0^t \lambda_{0j}(s)+b_i(s)ds+b_i(t)}]\,\lambda_{2j}(t)\,e^{\{-\beta_0' Z_i^*(t)+\gamma_2' Z_i(t)\}},$$

$$P(\tilde{N}_{ij}^{(1)} = 1) \; = \; E_b[b_i(t)]\;\lambda_{1j}(t)e^{\gamma_1' Z_i(t)},$$

$$P(\tilde{N}_{ij}^{(2)} = 1) \; = \; I(U_{ij} \;\leq t)E_b[b_i(t)]\;\lambda_{2j}(t)e^{\gamma_2' Z_i(t)}.$$

Since we have the complete data for the observation time $U_{ij}'s$ and $V_{ij}'s$, it is more efficient to directly estimate $\gamma_1's$ and$\gamma_2's$ from the observed data though the following estimating equation,

$$U_\gamma(\gamma_1, \gamma_2) = \sum_{j=1}^{k}\sum_{i=1}^{n}[\int_0^\infty Z_i(t) - \frac{S_{1,\gamma_1}^{1j}(t,\gamma_1)}{S_{1,\gamma_1}^{0j}(t,\gamma_1)}dN_{ij}^{(1)}(t) + \int_0^\infty Z_i(t) - \frac{S_{2,\gamma_2}^{(1j)}(t,\gamma_2)}{S_{2,\gamma_2}^{(0j)}(t,\gamma_2)}dN_{ij}^{(2)}(t),$$

$$= \sum_{j=1}^{k}\sum_{i=1}^{n}\left(Z_i(U_{ij}) - \frac{S_{1,\gamma_1}^{1j}(U_{ij},\gamma_1)}{S_{1,\gamma_1}^{(0j)}(U_{ij},\gamma_1)}\right) + \left(Z_i(V_{ij}) - \frac{S_{2,\gamma_2}^{(1j)}(V_{ij},\gamma_2)}{S_{2,\gamma_{2j}}^{(0)}(V_{ij},\gamma_2)}\right),$$

where

$$S_{1,\gamma_1}^{(mj)}(t,\gamma_1) = \frac{1}{n}\sum_{i=1}^{n} I(t \leq U_{ij})exp(\gamma_1' Z_i(t))Z_i(t)^m,$$

$$S_{2,\gamma_2}^{(mj)}(t,\gamma_2) = \frac{1}{n}\sum_{i=1}^{n} I(U_{ij} < t \leq V_{ij})exp(\gamma_2' Z_i(t))Z_i(t)^m,$$

$m = 0, 1$.

To estimate $\beta = (\beta_1, \ldots, \beta_k)$, again motivated by Wang et al. (2010), one can obtain consistent estimates by solving the estimating equation.

$$U_\beta(\beta, \gamma_1, \gamma_2) = \sum_{j=1}^{k} \sum_{i=1}^{n} \left[\int_0^\infty \left\{ Z_i^*(t) - \frac{S_{1,\beta_j}^{(1j)}(t, \beta_j, \gamma_1)}{S_{1,\beta_j}^{(0j)}(t, \beta_j, \gamma_1)} \right\} dN_{ij}^{(1)}(t) + \int_0^\infty \left\{ Z_i^*(t) - \frac{S_{2,\beta_j}^{1j}(t, \beta_j, \gamma_2)}{S_{2,\beta_j}^{0j}(t, \beta_j, \gamma_2)} \right\} dN$$

$$= \sum_{j=1}^{k} \sum_{i=1}^{n} (1 - \delta_{1i}^{(j)}) \left\{ Z_i^*(U_{ij}) - \frac{S_{1,\beta_j}^{(1j)}(U_{ij}, \beta_j, \gamma_1)}{S_{1,\beta_j}^{(0j)}(U_{ij}, \beta_j, \gamma_1)} \right\} + \delta_{3i}^{(j)} \left\{ Z_i^*(V_{ij}) - \frac{S_{2,\beta_j}^{1j}(V_{ij}, \beta_j, \gamma_2)}{S_{2,\beta_j}^{(0j)}(V_{ij}, \beta_j, \gamma_2)} \right\},$$

where

$$S_{1,\beta}^{(mj)}(t, \beta_j, \gamma_1) = \frac{1}{n} \sum_{i=1}^{n} I(t \leq U_{ij}) exp(-\beta_j' Z_{ij}^*(t) + \gamma_1' Z_i(t)) Z_i^*(t)^m,$$

$$S_{2,\beta}^{(mj)}(t, \beta_j, \gamma_2) = \frac{1}{n} \sum_{i=1}^{n} I(U_{ij} \leq t \leq V_{ij}) exp(-\beta_j' Z_i^*(t) + \gamma_2' Z_i(t)) Z_i^*(t)^m,$$

for $m = 0, 1$. And $Z_i^*(t) = \int_0^t Z_i(s) ds$.

Let $\hat{\gamma}$ be the solution to $U_\gamma(\gamma) = 0$. Then we can estimate β through the root of $U_\beta(\beta, \hat{\gamma}) = 0$. For the asymptotic distribution of $\hat{\beta}$, it can be proved that $\frac{1}{\sqrt{n}} U_\beta(\beta, \hat{\gamma})$ converges in distribution to a normal distribution with mean zero and a covariance matrix that can be estimated in the appendix. Then by Taylor expansion of $U_\beta(\hat{\beta}, \hat{\gamma})$ around β_0, the true value of β, we have

$$U_\beta(\hat{\beta}, \hat{\gamma}) = U_\beta(\beta_0, \hat{\gamma}) + \frac{\partial U_\beta(\beta_0, \hat{\gamma})}{\partial \beta} (\beta_0 - \hat{\beta}).$$

The distribution of $n^{-0.5}(\hat{\beta} - \beta_0)$ can be asymptotically approximated by the normal distribution with mean zero and covariate matrix given in appendix.

4.4 A Simulation Study

In this section, we conduct some simulation study to evaluate finite sample performance of the estimation approach in the previous sections. In the simulation study, we considered a multivariate case II interval censored data with informative censoring issue. For simplicity, we only consider the two dimensional case. To generate the survival times of interest $T_i = (T_{i1}, T_{i2})$, following the additive hazard model, exponential distribution are used. Here we generate the T_i from the exponential distribution with hazard $\lambda_{01} + \beta_1 Z_i + b_i$ and $\lambda_{02} + \beta_2 Z_i + b_i$, respectively and the treatment indicator Z_i is from the bernoulli distribution with success rate 0.5. Also, we assume that b_i is the latent variable follows the normal distribution with mean zero and variance 0.01.

For multivariate interval-censored data, we need to generate the censoring intervals for both T_{i1} and T_{i2}, that is U_{i1}, V_{i1}, U_{i2} and V_{i2}. For U_{i1} and V_{i1}, we first generated two random numbers U_{11} and U_{12} independently from the exponential distributions with means $1\,/\exp(\lambda_{11} + \gamma_1 Z_i + b_i)$ and $1/(\lambda_{21} exp(\gamma_2 Z_i + b_i)) - 1/(\lambda_{11} exp(\gamma_1 Z_i + b_i))$, respectively. Given U_{i1} and U_{i2}, we defined $U_{i1} = U_{11}$ and $V_{i1} = U_{i1} + U_{12}$. The generalization procedure of U_{i2} and V_{i2} are the same with U_{i1} and V_{i1}. The results given below are based on 1000 replications.

Table 4.1 presents the simulation results for each setup with sample size $n = 100$. The results include the bias (BIAS) given by the average of difference between point estimates and the true value, the sample standard errors (SSE), the true standard deviation of $1,000$ estimate, the sample errors estimate (SEE) representing the mean of the standard error estimates, and the 95 percent empirical coverage probability. It can be seen from Table 4.1 that the proposed approach worked very well in both non-informative and informative censoring situation. The biases of the proposed estimates were small. The means of the estimated standard deviations are close to the sample standard deviations, indicating the variance estimates seem reasonable. Moreover, the empirical coverage probabilities seemed quite close to 95 percent in all cases. Simulation results with different sample sizes were presented and indicate that when sample sizes increase, biases become smaller and the variance estimates become closer to the empirical variance estimates.

4.5 An Application

To illustrate the proposed test procedure, we apply it to the set of interval-censored data discussed in Goggins and Finkelstein (2000) and Sun (2006) among others. The data arose from AIDS observational study conducted by the AIDS Clinical Trials Group (ACTG). During the study, patients were scheduled to provide blood and urine samples at clinic visits every 12 weeks and every 4 weeks. The CD4 count was recorded at the entry time and patients were grouped by their CD4 counts. Urine samples and blood samples were tested in order to detect the presence of the cytomegalovirus (CMV) virus. There are 204 subjects in the study. 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 counts.

To investigate the relationship between CMV shedding and CD4 counts, we proposed additive hazard model to the real data, wihch gives the covariate effects $\beta_{blood} = 1.2056$ with estimated standard errors of 0.0435 and $\beta_{urine} = 0.8903$ with estimated standard errors of 0.4235. The results obtained here indicate that patients in late stage of HIV disease have higher risk of CMV shedding in either blood or urine. The similar conclusion was given by Goggins and Finkelstein (2000) and Sun (2006) using marginal grouped PH models.

4.6 Concluding Remarks

This chapter discussed regression analysis of multivariate interval-censored data using the additive hazards model when observation time may be related to survival time of interest. Additive hazard frailty model has been used and a latent variable was introduced in order to directly characterize the correlation between failure time and the dependence between failure time and observed time. This is the first paper, considering the correlation between multi-failure time variable and the dependence between observation time and event time together. The focus here is to estimate the regression effects and the counting processes have been used for developing the large sample properties of the proposed estimates. A major advantage of the proposed method is that it does not involve estimation of any baseline hazard function. The procedures can be easily implemented by using R and simulation studies and a real data application suggest that they perform well for practical situations.

However, one drawback of this model is that it assume the observation process follows Cox model, which may not be true in practice. Under that situation, our proposed method are theoretically invalid. To avoid this restriction, One common way is to use copula model to describe the association between multi-failure times as well as between observation time and event time. Anther direction for future research is to generalize the proposed method or to develop methods for situations where each subject is observed at a sequence of time points and survival time of interest is known only to belong to an interval.

Chapter 5

Future Research

In this chapter, we discuss several potential directions for future research that are related to the analysis of interval-censored data.

5.1 Nonparametric Test for Interval-Censored Data With Informative Censoring

In the previous chapters, we talked about the nonparametric test for interval-censored data. And the most important assumption for those approaches is that the censoring mechanism is independent with event time. However, the assumption may be violated in practice. Therefore, the hypothesis test for interval-censored data with informative censoring is also one of the important topics in survival analysis.

In Chapter 1, we briefly introduce several nonparametric tests for interval-censored data. To apply the generalized log-rank test, we need to have the nonparametric estimation of the common survival function among treatment groups. One solution to this problem is to use the estimation method in Wang et al.(2012). In their paper, they proposed two simple procedures to estimate the survival function under the copula model framework when one observes only current status data. Here we need to extend their methods to case II interval-censored data. For simplicity, we could make the assumption that the two observation processes are independent and split the interval-censored data into two current status data, then apply the estimation procedure.

5.2 Multiple Generalized Log-Rank Test for Interval-Censored Data

In Section 2, we discussed the nonparametric comparison problem of survival functions when only interval-censored failure time data are available. We proposed a class of nonparametric tests and established both finite sample and asymptotic properties of the presented approach. One major advantage of the proposed test procedure is that its asymptotic distribution is known under both null and alternative hypotheses, which makes both power and sample size calculation possible. One limitation of the proposed approach is that it can only compare two treatment groups. Since the two sample test statistics follows F distribution, we can extend the two sample test into a multiple test by using multiavriate F distribution. More work needs to be done about the asymptotic properties of the multiple test procedure.

APPENDIX

Appendix A:

Proof of Theorem 2.1 in Chapter 2

To prove the theorem, we will only need to prove the first part on $\bar{U}_{n_1}$ as the proof for the second part is similar. For this, note that we can rewrite $\bar{U}_{n_1}$ as

$$\bar{U}_{n_1} = \sqrt{n_1}(Q_{n_1} - Q_1)\left(K_{\hat{F}_{n_1},\hat{F}_{2,n_2}} - K_{F_1,F_2}\right)$$

$$+\sqrt{n_1}Q_1\left(K_{\hat{F}_{n_1},\hat{F}_{n_2}}\right) + \sqrt{n_1}(Q_{n_1} - Q_1)\left(K_{F_1,F_2}\right) . \tag{1}$$

For the second term at the right side of the above euqation, we have

$$\sqrt{n_1}Q_1\left(K_{\hat{F}_{n_1},\hat{F}_{n_2}}\right) = \sqrt{n_1}Q_1\left[\left\{K_{\hat{F}_{n_1},\hat{F}_{n_2}} - K_{\hat{F}_{n_1},F_2}\right\} - \left\{K_{F_1,\hat{F}_{n_2}} - K_{F_1,F_2}\right\}\right]$$

$$+\sqrt{n_1}Q_1\left\{K_{F_1,\hat{F}_{n_2}} - K_{F_1,F_2}\right\} + \sqrt{n_1}Q_1\left(K_{\hat{F}_{n_1},F_2}\right) \tag{2}$$

and

$$\sqrt{n_1} Q_1 \left(K_{\hat{F}_{n_1}, F_2} \right) = \sqrt{n_1} \int \left\{ K_{\hat{F}_{n_1}, F_2}(u, v, \delta, \gamma) q_{F_1}(u, v, \delta, \gamma) \right\} d(\lambda_2 \otimes \nu_2)$$

$$= \sqrt{n_1} \int \left\{ K_{\hat{F}_{n_1}, F_2}(u, v, \delta, \gamma) - K_{F_1, F_2}(u, v, \delta, \gamma) \right\} \left\{ q_{F_1}(u, v, \delta, \gamma) - q_{\hat{F}_{n_1}}(u, v, \delta, \gamma) \right\} d(\lambda_2 \otimes \nu_2)$$

$$+ \sqrt{n_1} \int K_{\hat{F}_{n_1}, F_2}(u, v, \delta, \gamma) q_{\hat{F}_{n_1}}(u, v, \delta, \gamma) \, d(\lambda_2 \otimes \nu_2)$$

$$+ \sqrt{n_1} \int K_{F_1, F_2}(u, v, \delta, \gamma) \left\{ q_{F_1}(u, v, \delta, \gamma) - q_{\hat{F}_{n_1}}(u, v, \delta, \gamma) \right\} d(\lambda_2 \otimes \nu_2) \, . \qquad (3)$$

It can be easily shown that

$$\sqrt{n_1} Q_1 \left[\left\{ K_{\hat{F}_{n_1}, \hat{F}_{n_2}} - K_{\hat{F}_{n_1}, F_2} \right\} - \left\{ K_{F_1, \hat{F}_{n_2}} - K_{F_1, F_2} \right\} \right] = o_p(1) \, ,$$

$$\sqrt{n_1} Q_1 \left\{ K_{F_1, \hat{F}_{n_2}} - K_{F_1, F_2} \right\} = 0 \, ,$$

$$\sqrt{n_1} \int \left\{ K_{\hat{F}_{n_1}, F_2}(u, v, \delta, \gamma) - K_{F_1, F_2}(u, v, \delta, \gamma) \right\} \left\{ q_{F_1}(u, v, \delta, \gamma) - q_{\hat{F}_{n_1}}(u, v, \delta, \gamma) \right\}$$

$$\times d(\lambda_2 \otimes \nu_2) = o_p(1) \, ,$$

and

$$\int K_{\hat{F}_{n_1}, F_2}(u, v, \delta, \gamma) q_{\hat{F}_{n_1}}(u, v, \delta, \gamma) d(\lambda_2 \otimes \nu_2) = 0 \, .$$

Thus it follows from (2) and (3) that

$$\sqrt{n_1} Q_1 \left(K_{\hat{F}_{n_1}, \hat{F}_{n_2}} \right) = -I_n + o_p(1) \, , \qquad (4)$$

where

$$I_n = \sqrt{n_1} \int K_{F_1,F_2}(u,v,\delta,\gamma) \left\{ q_{\hat{F}_{n_1}}(u,v,\delta,\gamma) - q_{F_1}(u,v,\delta,\gamma) \right\} d(\lambda_2 \otimes \nu_2) \, .$$

Now we will show that

$$\sqrt{n_1}\,(Q_{n_1} - Q_1)(K_{\hat{F}_{n_1},\hat{F}_{n_2}} - K_{F_1,F_2}) \to 0 \tag{5}$$

in probability as $n \to \infty$. For this, define

$$\mathcal{F} = \{F : F \text{ is a distribution function defined on } [0, M]\},$$

$$\mathcal{G} = \{F : F \in \mathcal{F},\, 0 < F(\delta_0) < F(M_0) < 1,\, \min_{\delta_0 \le t \le M_0 - \varepsilon_0} [F(t + \varepsilon_0) - F(t)] \neq 0\}$$

and

$$\mathcal{H} \,=\, \{K_{F_3,F_4}(u,v,\delta,\gamma) - K_{F_1,F_2}(u,v,\delta,\gamma) \,:\, (u,v) \in D,\, F_3, F_4 \in \mathcal{G}\}\ ,$$

where $D = \{(u,v) : \, u \geq \delta_0,\, u + \varepsilon_0 \leq v \leq M_0\}$. Because $\mathcal{F}$ is a $P - Donsker$ from the proof of Corollary 5.1 of Huang and Wellner (1995), $\mathcal{G}$ is a $P - Donsker$ by Theorem 2.10.1 of van der Vaart and Wellner (1996). Note that for any $F_3, F_4, F_5, F_6 \in \mathcal{G}$,

$(u, v) \in D$, we have

$$\begin{aligned}
&\left| \delta \frac{\eta(F_4(u)) - c_0}{F_3(u)} + \gamma \frac{\eta(F_4(v)) - \eta(F_4(u))}{F_3(v) - F_3(u)} + (1 - \delta - \gamma) \frac{c_0 - \eta(F_4(v))}{1 - F_3(v)} \right. \\
&\left. -\delta \frac{\eta(F_6(u)) - c_0}{F_5(u)} - \gamma \frac{\eta(F_6(v)) - \eta(F_6(u))}{F_5(v) - F_5(u)} - (1 - \delta - \gamma) \frac{c_0 - \eta(F_6(v))}{1 - F_5(v)} \right| \\
\leq \; & c \left[|F_3(u) - F_5(u)| + |F_3(v) - F_5(v)| + |F_4(u) - F_6(u)| + |F_4(v) - F_6(v)| \right]
\end{aligned}$$

for some constant c. Then it can be shown by using the bracket entropy theorem of van der Vaart and Wellner (1996, pp. 127-59) and the arguments similar to those used in Huang and Wellner (1995) that $\mathcal{H}$ is $P - Donsker$. Also note that $\hat{F}_{n_1}, \hat{F}_{n_2} \in \mathcal{G}$ for all n sufficiently large and as $n \to \infty$, we have

$$\int \{ |\hat{F}_{n_l}(u) - F_l(u)|^2 + |\hat{F}_{n_l}(v) - F_l(v)|^2 \} dP \longrightarrow 0$$

in probability from the strong consistency of $\hat{F}_{n_l}$ (Groeneboom and Wellner, 1992, pp. 85). Thus (5) is true based on this and the uniform asymptotic equcontinuity of the empirical process resulting from the Donsker property (van der Vaart and Wellner, 1996, pp. 168-71).

It follows from (1), (4) and (5) that we have

$$\bar{U}_{n_1} = \sqrt{n_1}(Q_{n_1} - Q_1) K_{F_1, F_2} - I_n + o_p(1) \,. \tag{6}$$

To finish the proof, next we will show that

$$I_n = \sqrt{n_1}(Q_{n_1} - Q_1)(\tilde{\theta}_{g_1,F_1}) + o_p(1)\,, \tag{7}$$

where $\tilde{\theta}_{g,F}$ is defined below. For this, note that

$$I_n = \sqrt{n_1}\int h_1(u)\frac{\eta(F_2(u)) - c_0}{F_1(u)}\{\hat{F}_{n_1}(u) - F_1(u)\}du$$

$$+\sqrt{n_1}\int h(u,v)\frac{\eta(F_2(v)) - \eta(F_2(u))}{F_1(v) - F_1(u)}[\{\hat{F}_{n_1}(v) - F_1(v)\} - \{\hat{F}_{n_1}(u) - F_1(u)\}]dudv$$

$$-\sqrt{n_1}\int h_2(v)\frac{c_0 - \eta(F_2(v))}{1 - F_1(v)}\{\hat{F}_{n_1}(v) - F_1(v)\}dv = \sqrt{n_1}\int g_1(t)\{\hat{F}_{n_1}(t) - F_1(t)\}dt\,,$$

where

$$g_1(t) = h_1(t)\frac{\eta(F_2(t)) - c_0}{F_1(t)} + \int_0^t h(u,t)\frac{\eta(F_2(t)) - \eta(F_2(u))}{F_1(t) - F_1(u)}du$$
$$- \int_t^M h(t,v)\frac{\eta(F_2(v)) - \eta(F_2(t))}{F_1(v) - F_1(t)}dv - h_2(t)\frac{c_0 - \eta(F_2(t))}{1 - F_1(t)}$$

with h_1 and h_2 being the marginal density functions of U_i and V_i, respectively.

Define

$$h^*(u,v) = \begin{cases} h(u,v), & \text{if } u \le v, \\ h(v,u), & \text{if } u > v, \end{cases}$$

and

$$d_F(x) = \frac{F(x)\{1 - F(x)\}}{h_1(x)\{1 - F(x)\} + h_2(x)F(x)}\,.$$

Let $\phi = \phi_{g,F}$ be the right-continuous solution to the following equation

$$\phi(x) = d_F(x) \left\{ g(x) - \int_0^x \frac{\phi(x) - \phi(x')}{|F(x) - F(x')|} h^*(x', x) dx' \right\} .$$

Also define

$$\tilde{\theta}_{g,F}(u, v, \delta, \gamma) = -\delta \frac{\phi_{g,F}(u)}{F(u)} - \gamma \frac{\phi_{g,F}(v) - \phi_{g,F}(u)}{F(v) - F(u)} + (1 - \delta - \gamma) \frac{\phi_{g,F}(v)}{1 - F(v)} .$$

Then it follows from Groeneboom (1996, pp. 149) that we have

$$I_n = \sqrt{n_1} \int g_1(t) \{\hat{F}_{n_1}(t) - F_1(t)\} dt = \sqrt{n_1}(Q_{n_1} - Q_1)(\tilde{\theta}_{g_1, \hat{F}_{n_1}})$$

$$= \sqrt{n_1}(Q_{n_1} - Q_1)(\tilde{\theta}_{g_1, \hat{F}_{n_1}} - \tilde{\theta}_{g_1, F_1}) + \sqrt{n_1}(Q_{n_1} - Q_1)\tilde{\theta}_{g_1, F_1}$$

$$= \sqrt{n_1}(Q_{n_1} - Q_1)(\tilde{\theta}_{g_1, F_1}) + o_p(1) ,$$

which is (7). Thus based on (6) and (7), we have

$$\bar{U}_{n_1} = \sqrt{n_1}(Q_{n_1} - Q_1) \left\{ K_{F_1, F_2} - \tilde{\theta}_{g_1, F_1} \right\} + o_p(1) ,$$

which is the first part of Theorem 1.

As pointed above, the second part of Theorem 1 can be proved similarly and in this case, we have

$$g_2(t) = h_1(t)\frac{\eta(F_1(t)) - c_0}{F_2(t)} + \int_0^t h(u,t)\frac{\eta(F_1(t)) - \eta(F_1(u))}{F_2(t) - F_2(u)}du$$

$$- \int_t^M h(t,v)\frac{\eta(F_1(v)) - \eta(F_1(t))}{F_2(v) - F_2(t)}dv - h_2(t)\frac{c_0 - \eta(F_1(t))}{1 - F_2(t)} \, .$$

Appendix B:

Asymptotic joint normal distribution of $U_3(\gamma_{10}, \gamma_{20})$ and $U_4(\gamma_{10}, \gamma_{20})$ in Chapter 3

Define

$$\tilde{M}_{1i}(t) = \tilde{N}_{1i}(t) - \int_0^t I(s \le U_i)\lambda_1(s)\exp(\gamma_{10}' Z_i(s))ds,$$

and

$$\tilde{M}_{2i}(t) = \tilde{N}_{2i}(t) - \int_0^t I(U_i < s \le V_i)\lambda_2(s)\exp(\gamma_{20}' Z_i(s))ds,$$

which are martingales. We want to derive the asymptotic joint distribution of $U_3(\gamma_{10}, \gamma_{20})$ and $U_4(\gamma_{10}, \gamma_{20})$, note that the first part of $U_3(\gamma_{10}, \gamma_{20})$ can be rewritten as

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^n K_1(U_i, Z_i, \gamma_{10}) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^n (Z_i - \bar{Z})\exp(-\gamma_{10}' Z_i)\int_0^\infty \frac{N_i(t)}{S_1(t)^{\exp(\gamma_{10}' Z_i)}} d\tilde{N}_{1i}(t) \\
&+ \frac{1}{\sqrt{n}}\sum_{i=1}^n (Z_i - \bar{Z})\exp(-\gamma_{10}' Z_i)\int_0^\infty N_i(t)\left\{\frac{1}{\hat{S}_1(t,\gamma_{10})^{\exp(\gamma_{10}' Z_i)}} - \frac{1}{S_1(t)^{\exp(\gamma_{10}' Z_i)}}\right\} d\tilde{N}_{1i}(t) \\
&= T_1^{(1)}(\gamma_0) + T_2^{(1)}(\gamma_0), \qquad (5.1)
\end{aligned}$$

where $\gamma_0 = (\gamma_{10}', \gamma_{20}')'$. Also note that $\hat{S}_1(t, \gamma_{10}) = \exp(-\hat{\Lambda}_1(t, \gamma_{10}))$, and

$$\sqrt{n}\{\hat{\Lambda}_1(t, \gamma_{10}) - \Lambda_1(t)\} \xrightarrow{P} \frac{1}{\sqrt{n}}\sum_{i=1}^n \int_0^t \frac{d\tilde{M}_{1i}(s)}{S_1^{(0)}(s, \gamma_{10})}.$$

It asymptotically yields that

$$T_2^{(1)}(\gamma_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^\infty \frac{R_1(t)}{S_1^{(0)}(t, \gamma_{10})} d\tilde{M}_{1i}(t), \tag{5.2}$$

where

$$R_1(t) = \frac{1}{n} \sum_{i=1}^{n} (Z_i - \bar{Z}) \int_t^\infty \frac{N_i(s) d\tilde{N}_{1i}(s)}{\hat{S}_1(s, \hat{\gamma}_1)^{\exp(\hat{\gamma}_1' Z_i)}}.$$

Similarly, we can rewrite the second part of $U_3(\gamma_{10}, \gamma_{20})$ as,

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \sum_{i=1}^{n} K_2(U_i, V_i, Z_i, \gamma_{20}) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Z_i - \bar{Z}) \exp(-\gamma_{20}' Z_i) \int_0^\infty \frac{N_i(t)(1 - \Delta_{1i})}{S_2(t)^{\exp(\gamma_{20}' Z_i)}} d\tilde{N}_{2i}(t) \\
&\quad + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Z_i - \bar{Z}) \exp(-\gamma_{20}' Z_i) \\
&\quad \times \int_0^\infty N_i(t)(1 - \Delta_{1i}) \left\{ \frac{1}{\hat{S}_2(t, \gamma_{20})^{\exp(\gamma_{20}' Z_i)}} - \frac{1}{S_2(t)^{\exp(\gamma_{20}' Z_i)}} \right\} d\tilde{N}_{2i}(t) \\
&= T_1^{(2)}(\gamma_0) + T_2^{(2)}(\gamma_0).
\end{aligned} \tag{5.3}$$

Since $\hat{S}_2(t, \gamma_{20}) = \exp(-\hat{\Lambda}_2(t, \gamma_{20}))$, and

$$\sqrt{n}\{\hat{\Lambda}_2(t, \gamma_{20}) - \Lambda_2(t)\} \xrightarrow{P} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^t \frac{d\tilde{M}_{2i}(s)}{S_2^{(0)}(s, \gamma_{20})}.$$

Hence, asymptotically, we have

$$T_2^{(2)}(\gamma_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^\infty \frac{R_2(t)}{S_2^{(0)}(t, \gamma_{20})} d\tilde{M}_{2i}(t), \tag{5.4}$$

where

$$R_2(t) = \frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z})\int_t^\infty \frac{N_i(s)(1-\Delta_{1i})d\tilde{N}_{2i}(s)}{\hat{S}_2(s,\hat{\gamma}_2)^{\exp(\hat{\gamma}_2' Z_i)}}.$$

For $U_4(\gamma_{10},\gamma_{20})$, since

$$\begin{aligned} U_1(\gamma_{10}) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^\infty \left\{ Z_i - \frac{S_1^{(1)}(t,\gamma_{10})}{S_1^{(0)}(t,\gamma_{10})} \right\} d\tilde{N}_{1i}(t) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^\infty \left\{ Z_i - \frac{S_1^{(1)}(t,\gamma_{10})}{S_1^{(0)}(t,\gamma_{10})} \right\} d\tilde{M}_{1i}(t), \end{aligned} \tag{5.5}$$

and

$$\begin{aligned} U_2(\gamma_{20}) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^\infty \left\{ Z_i - \frac{S_2^{(1)}(t,\gamma_{20})}{S_2^{(0)}(t,\gamma_{20})} \right\} d\tilde{N}_{2i}(t) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^\infty \left\{ Z_i - \frac{S_2^{(1)}(t,\gamma_{20})}{S_2^{(0)}(t,\gamma_{20})} \right\} d\tilde{M}_{2i}(t). \end{aligned} \tag{5.6}$$

Thus, it follows from equations (1)-(6) that the joint distribution of $U_3(\gamma_{10},\gamma_{20})$ and $U_4(\gamma_{10},\gamma_{20})$ is asymptotically normal with mean 0, and the covariance matrix Γ of $U_3(\gamma_{10},\gamma_{20})$ and $U_4(\gamma_{10},\gamma_{20})$ can be estimated by $\sum_{i=1}^{n} B_i B_i'/n$, where $B_i = (a_{1i} +$

$b_{1i}, a_{2i} + b_{2i}, \alpha_{1i}, \alpha_{2i})'$ for $i = 1, \ldots, n$, and

$$\begin{aligned}
a_{1i} &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Z_i - \bar{Z}) \exp(-\hat{\gamma}_1' Z_i) \frac{N_i(U_i)}{\hat{S}_1(U_i, \hat{\gamma}_1)^{\exp(\hat{\gamma}_1' Z_i)}}, \\
b_{1i} &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^\infty \frac{R_1(t)}{S_1^{(0)}(t, \hat{\gamma})} \left(d\tilde{N}_{1i}(t) - \frac{I(U_i < t \leq V_i) \exp(\hat{\gamma}_1' Z_i)}{n S_1^{(0)}(t, \hat{\gamma}_1)} d\tilde{N}_2(t) \right), \\
a_{2i} &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Z_i - \bar{Z}) \exp(-\hat{\gamma}_2' Z_i) \frac{N_i(V_i)(1 - \Delta_{1i})}{\hat{S}_2(V_i, \hat{\gamma}_2)^{\exp(\hat{\gamma}_2' Z_i)}}, \\
b_{2i} &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^\infty \frac{R_2(t)}{S_2^{(0)}(t, \hat{\gamma})} \left(d\tilde{N}_{2i}(t) - \frac{I(U_i < t \leq V_i) \exp(\hat{\gamma}_2' Z_i)}{n S_2^{(0)}(t, \hat{\gamma}_2)} d\tilde{N}_2(t) \right), \\
\alpha_{1i} &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^\infty \left(Z_i - \frac{S_1^{(1)}(t, \hat{\gamma}_1)}{S_1^{(0)}(t, \hat{\gamma}_1)} \right) \left(d\tilde{N}_{2i}(t) - \frac{I(t \leq U_i) \exp(\hat{\gamma}_1' Z_i)}{n S_1^{(0)}(t, \hat{\gamma}_1)} d\tilde{N}_1(t) \right), \\
\alpha_{2i} &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \int_0^\infty \left(Z_i - \frac{S_2^{(1)}(t, \hat{\gamma}_2)}{S_2^{(0)}(t, \hat{\gamma}_2)} \right) \left(d\tilde{N}_{2i}(t) - \frac{I(U_i < t \leq V_i) \exp(\hat{\gamma}_2' Z_i)}{n S_2^{(0)}(t, \hat{\gamma}_2)} d\tilde{N}_2(t) \right).
\end{aligned}$$

Appendix C:

Asymptotic normality of $n^{-0.5}U_\beta(\beta_0, \hat{\gamma})$

Define

$$M_{ij}^{(1)}(t) = N_{ij}^{(1)}(t) - \int_0^t I(s \le U_{ij})\lambda_{1j}^*(s)\exp(-\beta_j' Z_i^*(s) + \gamma_1' Z_i(s))ds,$$

$$M_{ij}^{(2)}(t) = N_{ij}^{(2)}(t) - \int_0^t I(U_{ij} < s \le U_{ij})\lambda_{2j}^*(s)\exp(-\beta_j' Z_i^*(s) + \gamma_2' Z_i(s))ds,$$

$$\tilde{M}_{ij}^{(1)}(t) = \tilde{N}_{ij}^{(1)}(t) - \int_0^t I(s \le U_{ij})\lambda_{1j}(s)\exp(\gamma_1' Z_i(s))ds,$$

and

$$\tilde{M}_{ij}^{(2)}(t) = \tilde{N}_{ij}^{(2)}(t) - \int_0^t I(U_{ij} < s \le V_{ij})\lambda_{2j}(s)\exp(\gamma_2' Z_i(s))ds,$$

where $\lambda_{1j}^*(t) = \lambda_{1j}(t)\, E(e^{-\int_0^t [\lambda_{0j}(s)+b_i(s)]ds + b_i(t)})$ and $\lambda_{2j}^*(t) = \lambda_{2j}(t)\, E(e^{-\int_0^t [\lambda_{0j}(s)+b_i(s)]ds + b_i(t)})$.

Also define

$$A_1 = \sum_{j=1}^n E\left(\int_0^\infty Z_1^*(t) - \frac{s_{1,\beta_j}^{(1)}(t,\beta_j,\gamma_1)}{s_{1,\beta_j}^{(0)}(t,\beta_j,\gamma_1)}^{\bigotimes 2} I(U_{1j} \ge t)\lambda_{1j}^*(t) exp(-\beta_j' Z_1^*(t) + \gamma_1' Z_1(t))dt\right),$$

$$A_2 = \sum_{j=1}^{n} E\left(\int_0^\infty Z_1^*(t) - \frac{s^{(1)}_{2,\beta_j}(t,\beta_j,\gamma_2)}{s^{(0)}_{2,\beta_j}(t,\beta_j,\gamma_2)}^{\bigotimes 2} I(U_{1j} < t \le V_{1j})\lambda^*_{2j}(t) exp(-\beta_j' Z_1^*(t) + \gamma_2' Z_1(t))dt\right),$$

$$\tilde{A}_1 = \sum_{j=1}^{n} E\left(\int_0^\infty Z_1(t) - \frac{s^{(1)}_{1,\gamma_{1j}}(t,\gamma_1)}{s^{(0)}_{1,\gamma_1}(t,\gamma_1)}^{\bigotimes 2} I(U_{1j} \ge t)\lambda_{1j}(t) exp(-\gamma_1' Z_1^*(t))dt\right),$$

and

$$\tilde{A}_2 = \sum_{j=1}^{n} E\left(\int_0^\infty Z_1(t) - \frac{s^{(1)}_{2,\gamma_2}(t,\gamma_2)}{s^{(0)}_{2,\gamma_2}(t,\gamma_2)}^{\bigotimes 2} I(U_{1j} < t \le V_{1j})\lambda_{2j}(t) exp(-\gamma_2' Z_1^*(t))dt\right),$$

where $s^{(m)}_{l,\gamma_1}(t,\gamma_1)$, $s^{(m)}_{l,\gamma_2}(t,\gamma_2)$, $s^{(m)}_{l,\beta_j}(t,\beta_j,\gamma_1)$ and $s^{(m)}_{l,\beta_j}(t,\beta_j,\gamma_2)$ denote the limits of $S^{(m)}_{l,\gamma_1}(t,\gamma_1)$, $S^{(m)}_{l,\gamma_2}(t,\gamma_2)$, $S^{(m)}_{l,\beta_j}(t,\beta_j,\gamma_1)$ and $S^{(m)}_{l,\beta_j}(t,\beta_j,\gamma_2)$, respectively, for $m = 0, 1$ and $l = 1, 2$. Let $A_\gamma = A_1 + A_2$ and $B = \tilde{A}_1 + \tilde{A}_2$, and assume A_γ and B are positive definite. Then we could use $\hat{A}_\gamma(\beta,\gamma) = -\frac{1}{n}\partial U_\beta(\beta,\gamma)/\partial\gamma$ and $\hat{B}(\gamma) = -\frac{1}{n}\partial U_\gamma(\gamma)/\partial\gamma$ to estimate A_γ and B, since A_γ and B are the limits of $\hat{A}_\gamma(\beta,\gamma)$ and $\hat{B}(\gamma)$ at β_0 and γ_0, respectively.

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

$$\frac{1}{\sqrt{n}}U_\beta(\beta,\hat{\gamma}) \;=\; \frac{1}{\sqrt{n}}U_\beta(\beta,\gamma) + A_\gamma B^{-1}\frac{1}{\sqrt{n}}U_\gamma(\gamma) + o_p(1).$$

Following Lin(1998), it can be shown that

$$\frac{1}{\sqrt{n}}U_\beta(\beta,\gamma) \ = \ \frac{1}{\sqrt{n}}\sum_{j=1}^{k}\sum_{i=1}^{n} a_{1ij}(\beta,\gamma) + a_{2ij}(\beta,\gamma) + o_p(1),$$

$$\frac{1}{\sqrt{n}}U_\gamma(\gamma) \ = \ \frac{1}{\sqrt{n}}\sum_{j=1}^{k}\sum_{i=1}^{n} c_{1ij}(\gamma) + c_{2ij}(\gamma) + o_p(1),$$

where

$$a_{1ij}(\beta,\gamma) = \int_0^\infty \left\{ Z_i^*(t) - \frac{s_{1,\beta_j}^{(1)}(t,\beta_j,\gamma_1)}{s_{1,\beta_j}^{(0)}(t,\beta_j,\gamma_1)} \right\} dM_{ij}^{(1)}(t),$$

$$a_{2ij}(\beta,\gamma) = \int_0^\infty \left\{ Z_i^*(t) - \frac{s_{2,\beta_j}^{(1)}(t,\beta_j,\gamma_2)}{s_{2,\beta_j}^{(0)}(t,\beta_j,\gamma_2)} \right\} dM_{ij}^{(2)}(t),$$

$$c_{1ij}(\gamma) = \int_0^\infty \left\{ Z_i(t) - \frac{s_{1,\gamma_1}^{(1)}(t,\gamma_1)}{s_{1,\gamma_1}^{(0)}(t,\gamma_1)} \right\} d\tilde{M_{ij}}^{(1)}(t),$$

and

$$c_{2ij}(\gamma) = \int_0^\infty \left\{ Z_i(t) - \frac{s_{2,\gamma_2}^{(1)}(t,\gamma_2)}{s_{2,\gamma_2}^{(0)}(t,\gamma_2)} \right\} d\tilde{M_{ij}}^{(2)}(t).$$

Then

$$\frac{1}{\sqrt{n}}U_\beta(\beta,\hat{\gamma}) \ = \ \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \alpha_i(\beta,\gamma) + o_p(1),$$

where $\alpha_i(\beta,\gamma) = \sum_{j=1}^{K}\{a_{1ij}(\beta,\gamma) + a_{2ij}(\beta,\gamma) + A_\gamma B^{-1}[c_{1ij}(\gamma) + c_{2ij}(\gamma)]\}$. It thus follows from the U statistic theory that $\frac{1}{\sqrt{n}}U_\beta(\beta_0,\hat{\gamma})$ converges in distribution to a zero mean normal random vector. The asymptotic covariance matrix of $\frac{1}{\sqrt{n}}U_\beta(\beta_0,\hat{\gamma})$ can be consistently estimated by

$$\hat{\Gamma} = \frac{1}{n}\sum_{i=1}^{n} \hat{\alpha}_i(\hat{\beta}, \hat{\gamma})\hat{\alpha}'_i(\hat{\beta}, \hat{\gamma}),$$

with $\hat{\alpha}_i(\hat{\beta}, \hat{\gamma}) = \sum_{j=1}^{K}\{\hat{a_{1ij}}(\hat{\beta}, \hat{\gamma}) + \hat{a_{2ij}}(\hat{\beta}, \hat{\gamma}) + \hat{A}_\gamma(\hat{\beta}, \hat{\gamma})\hat{B}(\hat{\gamma})[\hat{c}_{1ij}(\hat{\gamma}) + \hat{c}_{2ij}(\hat{\gamma})]\}$, where $\hat{a}_{1ij}, \hat{a}_{2ij}, \hat{c}_{1ij}$ and $\hat{c}_{2ij}$ are the estimates of $a_{1ij}, a_{2ij}, c_{1ij}$ and c_{2ij}, respectively.

BIBLIOGRAPHY

Aalen, O. O. (1978), Nonparametric inference for a family of counting process. *The Annals of Statistics*, **6**,701-726.

Andersen, P. K. and Ronn, B. B. (1995), A nonparametric test for comparing two samples where all observations are either left- or right-censored. *Biometrics*, **51**, 323-329.

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

Chen, L. and Sun, J. (2010), Multiple imputation to regression analysis of interval-censored failure time data with linear transformation models. *Far East Journal of Theoretical Statistics* **33**, 41C55

Chen, M., Tong, X. and J., S. (2007), The proportional odds model for multi-variate interval-censored failure time data. *Statistics in Medicine* **26**, 5147-5161.

Chen, M., Tong, X., and Sun, J. (2009), A frailty model approach for regression analysis of multivariate current status data. *Statistics in Medicine* **28**, 3424-3436.

Fang, H., Sun, J. and Lee, M-L T. (2002), Nonparametric survival comparison for interval-censored continuous data. *Statistica Sinica* **12**, 1073-1083.

Fay, M. and Shih, J. (2012), Weighted logrank tests for interval censored data when assessment times depend on treatment. *Statistics in Medicine* **31**, 3760-3772.

Ferreira, M. U., Cardoso, M. A., Santos, A. L. S., Ferreira, C. S., and Szarfarc, S. C. (1996), Rapid epidemiologic assessment of breastfeeding practices: probit analysis of current status data. *Journal of Tropical Pediatrics* , **42(1)**, 50-53.

Fleming, T. R. and Harrington, D. P. (1991), *Counting Process and Survival Analysis.* John Wiley: New York.

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

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

Finkelstein D. M., Goggins W. B., Schoenfeld D. A. (2002), Analysis of failure time data with dependent interval censoring. *Biometrics*,**58**, 298C304.

Frydman, H. and Szarek, M. (2009), Nonparametric estimation in a Markov 'Illness-Death' process from interval censored observations with missing intermediate transition status. *Biometrics*,**65**, 143-151.

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 subjects 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 Mathematics*, 1648, Springer-Verlag, Berlin.

Groeneboom, P. and Wellner, J. A. (1992), *Information bounds and nonparametric maximum likelihood estimation*, DMV Seminar, Band 19, Birkhauser, New York.

Grummer-strawn, LM (1993), Regression analysis of current-status data: an application to breast-feeding. *Journal of American Statistical Association*, **88(423)**, 758-65.

Hens, N., Wienke, A., Aerts, M. and Molenberghs, G. (2009), The corre-lated and shared gamma frailty model for bivariate current status data: An illustration for cross-sectional serological data. *Statistics in Medicine* **28**, 2785C2800.

Huang, J. and Wellner, J. A. (1995), Asymptotic normality of the NPMLE of linear functionals for interval censored data, case I. *Statistica Neerlandica* **49**, 153-163.

Huang, J., Lee, C. and Yu, Q. (2008), A generalized log-rank test for interval-censored failure time data via multiple imputation. *Statistics in Medicine* **27**, 3217-3226.

Jewell, N.P. and Lann, M. (1995), Generalizations of current status data with applications. *Lifetime Data Analysis*, **1**,101-109.

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

Kim, Y., Kim, J., Nam, C. and Kim, Y. (2012), statistical analysis of dependent current status data. *Interval censored time to event data*, Tylor and Francis Group,113-148.

Lin, D.Y., Robins, J.M. and Wei, L.J. (1996), Comparing two failure time distributions in the presence of dependent censoring. *Biometrika*, **83,2**, 381-393.

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

Nielsena, J. and Parner, E. (2010), Analyzing multivariate survival data using composite likelihood and flexible parametric modeling of the hazard functions. *Statistics in Medicine*, **29**, 2126-2136.

Oller, R. and Gomez, G. (2012), A generalized Fleming and Harrington's class of tests for interval-censored data. *Canadian Journal of Statistics*, **40**, 1-16.

Pan, W. (2000), A two-sample test with interval censored data via multiple imputation. *Statistics in Medicine*, **19**, 1-11.

Peto, R. and Peto, J. (1972), Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society* A **135**, 185-207.

Self, S. G. and Grossman, E. A. (1986), Linear rank tests for interval-censored data with application to PCB levels in adipose tissue of transformer repair workers. *Biometrics*. **42**, 521-530.

Schick, A. and Yu, Q. (2000), Consistency of the GMLE with mixed case interval censored data. *Scandinavian Journal of Statistics*, **27**, 45-55.

L. Sun, L. Wang, and J. Sun (2006), Estimation of the Association for Bivariate Interval-censored Failure Time Data. *Scandinavian Journal of Statistics*, **33**, 637-649.

Sun, J. (1996), A nonparametric test for interval-censored failure time data with application to AIDS studies. *Statistics in Medicine* **15**, 1387-1395.

Sun, J. and Kalbfleisch, J. D. (1993), The analysis of current status data on point processes. *Journal of American Statistical Association*, **88**, 1449-1454.

Sun, J. (1995), Empirical estimation of a distribution function with truncated and doubly interval censored data and its application to AIDS studies *Biometrics*, **51**, 1096-1104.

Sun, J. (1999), A nonparametric test for current status data with unequal Censoring. *J. R. Statist. Soc.* B **61**, 243-250.

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

Sun, J. (2001), Nonparametric tests for doubly interval-censored failure time data. *Lifetime Data Analysis*,**7**, 363-375.

Sun, J. (2006), *Statistical Analysis of Interval-censored Failure Time Data*. New York: Springer.

Sun, J., Zhao, Q. and Zhao, X. (2005), Generalized long-rank tests for interval-censored failure time data. *Scandinavian Journal of Statistics* **32**, 49-57.

Tian, L. and Cai, T. (2006), On the accelerated failure time model for current status and interval censored data. *Biometrika*, **93 (2)**, 329-342.

Tong, X., Chen, M. and Sun, J. (2008), Regression analysis of multivariate interval-censored failure time data with application to tumorigenicity experiments. *Biometrical Journal*, **50**, 364-374.

Turnbull, B. W. (1976), The empirical distribution with arbitrarily grouped censored and truncated data. *J. R. Statist. Soc.* B **38**, 290-295.

van der Vaart, A. W. and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes*. New York: Springer.

Wang, C., Sun, J., Sun, L., Zhou, J. and Wang, D. (2012), Nonparametric estimation of current status data with dependent censoring *Lifetime Data Analysis* **18**, 434-445.

Wang, L., Sun, J. and Tong, X. (2010), Regression analysis of case II interval censored failure time data with the additive hazards model. *Statistica Sinica* **20**, 1709-1723.

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

Zeng, D., Cai, J. and Shen, Y. (2006), Semiparametric additive risks model for interval censored data *Statisitics Sinica*,**16**, 287-302.

Zhao, Q. and Sun, J. (2004), Generalized log-rank test for mixed-censored failure time Data. *Statistics in Medicine* **23**, 1621-1629.

Zhao, X., Zhao, Q., Sun, J. and Kim, J. S. (2008), Generalized log-rank tests for partly interval-censored failure time data. *Biometrical Journal* **50**, 375-385.

Zhang, Y., Liu, W. and Zhan, Y. (2001), A nonparametric two sample test of the failure function with interval censoring case 2. *Biometrica*, **88**, 677-686.

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

Zhang, Z., Sun, L., Sun, J. and Finkelstein, D. (2007), Regression analysis of failure time data with informative interval censoring. *Statistics in Medicine*, **26**, 2533-2546.

Zhu, C., Yuen, K. and Sun, J. and Zhao, X. (2008), A nonparametric test for interval censored failure time data with unequal censoring. *Communications in Statistics-Theory and Methods* **37**, 1895-1904.

Table 2.1: Estimated Size and Power Based on Simulated Data from Exponential Distribution

Censoring percentages	ρ	γ	β 3	2	1.5	0	-1.5	-2	-3
(20% , 20% , 60%)	0	0	0.478	0.158	0.085	0.057	0.260	0.838	1
	0	0.5	0.397	0.118	0.075	0.059	0.365	0.921	1
	0.5	0	0.708	0.242	0.103	0.043	0.212	0.804	1
	0.5	0.5	0.628	0.214	0.095	0.044	0.320	0.893	1
(17% , 16% , 67%)	0	0	0.489	0.171	0.118	0.047	0.262	0.844	1
	0	1	0.335	0.122	0.084	0.059	0.530	0.962	1
	1	0	0.855	0.397	0.172	0.040	0.199	0.793	1
	1	1	0.770	0.291	0.140	0.046	0.399	0.923	1

Table 2.2: Estimated Size and Power Based on Simulated Data from Gamma Distribution

Censoring			β						
percentages	ρ	γ	3	2	1.5	0	-1.5	-2	-3
(12% , 12% , 76%)	0	0	0.997	0.946	0.692	0.047	0.854	0.996	1
	0	0.5	0.993	0.923	0.686	0.053	0.927	1	1
	0.5	0	1.000	0.966	0.719	0.041	0.852	1	1
	0.5	0.5	1.000	0.960	0.708	0.042	0.919	1	1
(10% , 15% , 75%)	0	0	0.997	0.930	0.704	0.041	0.853	0.999	1
	0	1	0.989	0.907	0.695	0.053	0.946	1	1
	1	0	1.000	0.958	0.714	0.043	0.817	0.998	1
	1	1	0.999	0.948	0.705	0.040	0.936	1	1

Table 2.3: Results on the Analysis of AIDS Clinical Trial

	On Blood Shedding time						
(ρ, γ)	$(0, 0)$	$(0, 0.5)$	$(0.5, 0)$	$(0.5, 0.5)$	$(0, 1)$	$(1, 0)$	$(1, 1)$
S_0	0.00022	0.00029	0.00011	0.00036	0.00031	0.00023	0.00055
p-value	0.019	0.022	0.013	0.024	0.022	0.019	0.030
	On Urine Shedding time						
(ρ, γ)	$(0, 0)$	$(0, 0.5)$	$(0.5, 0)$	$(0.5, 0.5)$	$(0, 1)$	$(1, 0)$	$(1, 1)$
S_0	1.72e+15	3.17e+15	5.60e+17	1.75e+16	1.35e+16	7.11e+17	3.61e+16
p-value	3.06e-08	2.25e-08	1.70e-09	9.60e-09	1.09e-08	1.50e-09	6.70e-09

Table 3.1: Empirical Power and Size for Exponential Distribution

			β						
n	r_1	r_2	-4	-3.5	-2.5	0	2.5	3.5	4
75	0	0	0.337	0.330	0.204	0.047	0.405	0.740	0.794
	0.1	0	0.343	0.322	0.199	0.045	0.426	0.751	0.807
	0.2	0	0.339	0.324	0.207	0.059	0.466	0.762	0.821
	0	0.1	0.385	0.375	0.226	0.051	0.452	0.752	0.802
	0	0.2	0.432	0.417	0.235	0.05	0.467	0.761	0.828
	0.1	0.1	0.390	0.381	0.229	0.051	0.461	0.773	0.823
100	0	0	0.361	0.345	0.221	0.047	0.408	0.762	0.826
	0.1	0	0.357	0.348	0.227	0.049	0.432	0.774	0.843
	0.2	0	0.361	0.348	0.235	0.054	0.469	0.795	0.848
	0	0.1	0.392	0.383	0.262	0.055	0.461	0.781	0.834
	0	0.2	0.455	0.444	0.291	0.056	0.475	0.793	0.857
	0.1	0.1	0.412	0.397	0.263	0.059	0.472	0.815	0.852
200	0	0	0.392	0.373	0.250	0.047	0.435	0.781	0.832
	0.1	0	0.389	0.375	0.264	0.051	0.441	0.795	0.841
	0.2	0	0.392	0.379	0.255	0.053	0.483	0.817	0.855
	0	0.1	0.421	0.412	0.273	0.049	0.466	0.805	0.864
	0	0.2	0.481	0.472	0.335	0.052	0.486	0.814	0.886
	0.1	0.1	0.435	0.424	0.280	0.056	0.512	0.833	0.858

Table 3.2: Empirical Power and Size for Gamma Distribution

			β						
n	r_1	r_2	-4	-3.5	-2.5	0	2.5	3.5	4
75	0	0	0.316	0.307	0.252	0.058	0.990	0.990	0.991
	0	0.1	0.317	0.315	0.264	0.060	0.995	0.996	0.996
	0	0.2	0.318	0.316	0.285	0.061	0.994	0.996	0.996
	0.1	0	0.368	0.355	0.302	0.056	0.992	0.994	0.995
	0.2	0	0.395	0.392	0.318	0.058	0.993	0.996	0.996
	0.1	0.1	0.361	0.367	0.307	0.058	0.997	0.997	0.998
100	0	0	0.342	0.318	0.308	0.056	0.996	0.996	0.997
	0	0.1	0.342	0.334	0.302	0.059	0.998	0.998	0.998
	0	0.2	0.350	0.341	0.304	0.044	1	1	1
	0.1	0	0.391	0.382	0.353	0.050	0.998	0.999	1
	0.2	0	0.422	0.409	0.356	0.052	0.999	0.999	1
	0.1	0.1	0.395	0.385	0.353	0.054	0.998	0.998	1
200	0	0	0.367	0.365	0.360	0.057	1	1	1
	0	0.1	0.386	0.367	0.365	0.058	1	1	1
	0	0.2	0.392	0.382	0.374	0.047	1	1	1
	0.1	0	0.461	0.431	0.383	0.051	1	1	1
	0.2	0	0.505	0.472	0.416	0.047	1	1	1
	0.1	0.1	0.453	0.425	0.420	0.057	1	1	1

Table 3.3: Empirical Size for Exponential Distribution.

n	r_1	r_2	Proposed Test	Generalized Log-rank Test
100	0	0	0.053	0.042
	0.1	0	0.051	0.023
	0.2	0	0.058	0.010
	0	0.1	0.048	0.021
	0	0.2	0.047	0.013
	0.1	0.1	0.059	0.032

Table 4.1: Simulation Result for Estimation of β.

		Ture	Value		β_1				β_2			
n	γ_1	γ_2	β_1	β_2	BIAS	SSE	SEE	CP	BIAS	SSE	SEE	CP
100	1	1	0	0	0.017	0.118	0.140	0.947	0.029	0.070	0.099	0.953
			0	0.5	0.018	0.108	0.109	0.955	0.062	0.104	0.111	0.962
			0.5	0	0.017	0.106	0.109	0.961	0.019	0.068	0.097	0.938
			0.5	0.5	0.018	0.117	0.119	0.947	0.024	0.065	0.067	0.968

Table 4.2: Simulation Result for Estimation of γ.

		Ture	Value		γ_1				γ_2			
n	γ_1	γ_2	β_1	β_2	BIAS	SSE	SEE	CP	BIAS	SSE	SEE	CP
100	1	1	0	0	0.017	0.029	0.026	0.957	0.027	0.034	0.032	0.950
			0	0.5	0.016	0.029	0.025	0.961	0.009	0.031	0.031	0.951
			0.5	0	0.016	0.029	0.026	0.966	0.009	0.031	0.030	0.947
			0.5	0.5	0.022	0.020	0.021	0.963	0.004	0.019	0.021	0.969

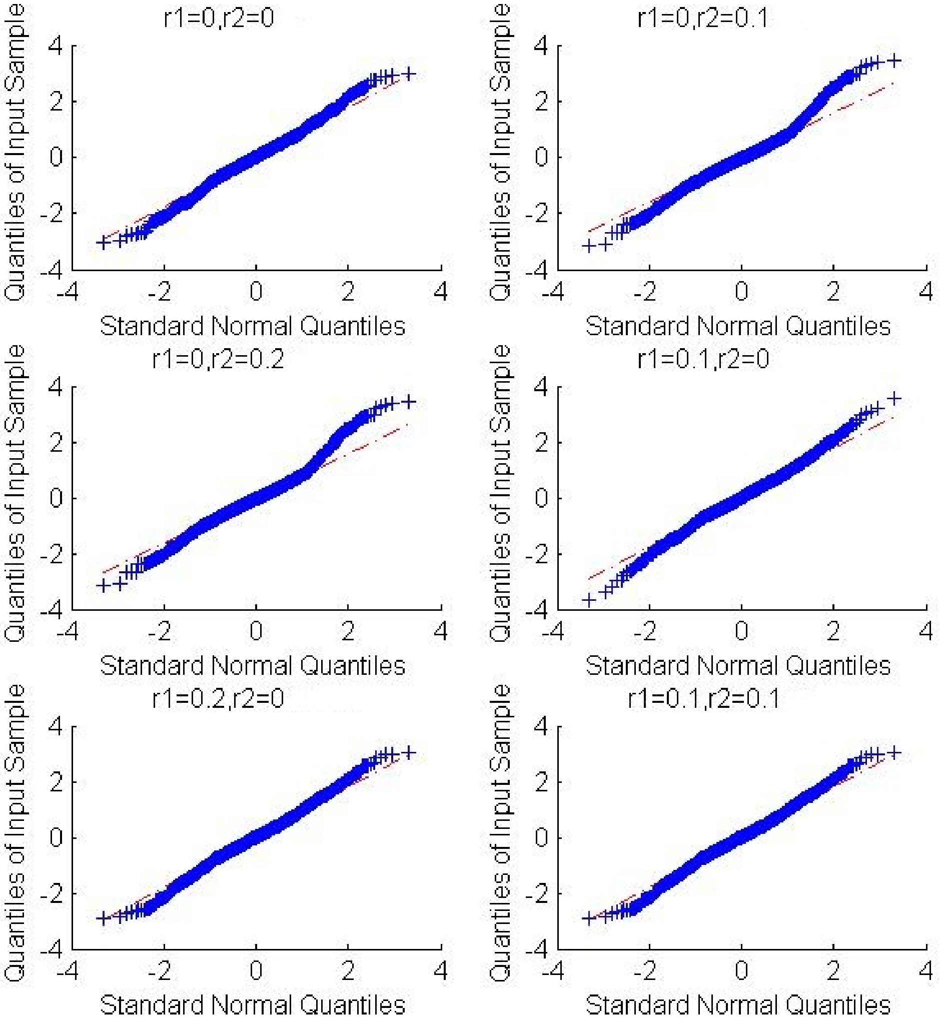

Figure 3.1: QQ-plot of Test Statistic for Exponential Distribution

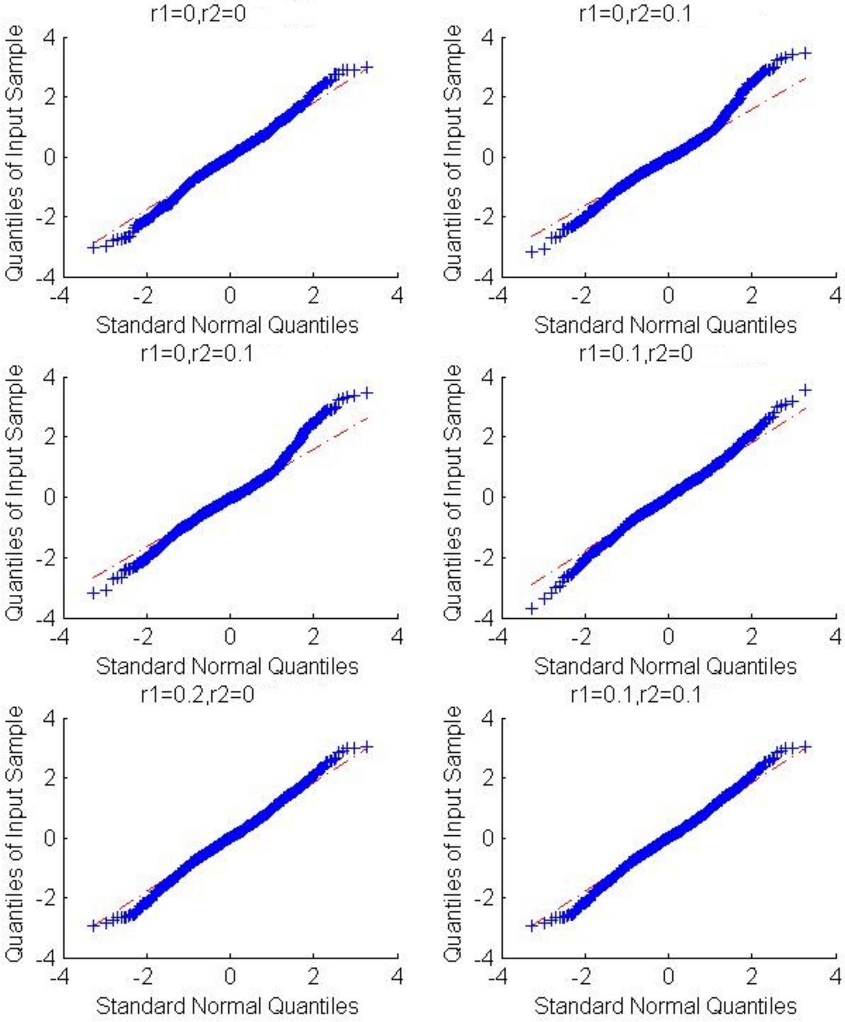

Figure 3.2: QQ-plot of Test Statistic for Gamma Distribution

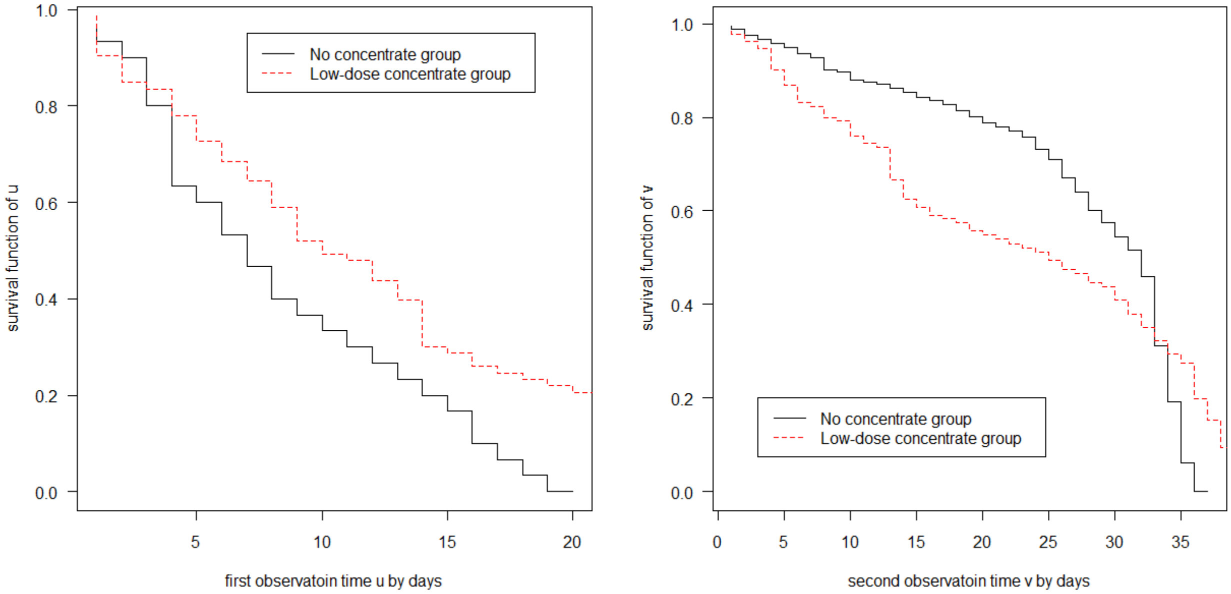

Figure 3.3: Kaplan-Meier estimate of survival function

VITA

Ran Duan was born on 1988 in Taiyuan, Shanxi Province, China. She received her B.S.in mathematics and statistics from Beijing Normal University in 2009. Then she joined the Ph.D program in the Department of Statistics at the University of Missouri in August 2009. She is going to graduate in 2013.