4-27-2011

Estimation of Hazard Function for Right Truncated Data

Yong Jiang

ESTIMATION OF HAZARD FUNCTION FOR RIGHT TRUNCATED DATA

by

YONG JIANG

Under the Direction of Dr. Xu Zhang

ABSTRACT

This thesis centers on nonparametric inferences of the cumulative hazard function of a right truncated variable. We present three variance estimators for the Nelson-Aalen estimator of the cumulative hazard function and conduct a simulation study to investigate their performances. A close match between the sampling standard deviation and the estimated standard error is observed when an estimated survival probability is not close to 1. However, the problem of poor tail performance exists due to the limitation of the proposed variance estimators. We further analyze an AIDS blood transfusion sample for which the disease latent time is right truncated. We compute three variance estimators, yielding three sets of confidence intervals. This work provides insights of two-sample tests for right truncated data in the future research.

INDEX WORDS: Right truncation, Kaplan-Meier method, Cumulative hazard, Two-sample test

ESTIMATION OF HAZARD FUNCTION FOR RIGHT TRUNCATED DATA

by

YONG JIANG

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Science

in the College of Arts and Sciences

Georgia State University

2011

ESTIMATION OF HAZARD FUNCTION FOR RIGHT TRUNCATED DATA

by

YONG JIANG

Committee Chair:	Dr. Xu Zhang
Committee:	Dr. Jiawei Liu
	Dr. Yuanhui Xiao

Electronic Version Approved:

Office of Graduate Studies
College of Arts and Sciences
Georgia State University
May 2011

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor, Dr. Xu Zhang, for all her dedicated guidance and support through my study at Georgia State University.

I would like to thank Dr. Jiawei Liu and Dr. Yuanhui Xiao for being my committee members, reading my thesis and providing valuable comments.

Finally I should thank my wife for her forever unconditional love. I could not have finished my work without her support. This thesis is dedicated to her.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS . . . . . **iv**

LIST OF TABLES . . . . . **vii**

LIST OF FIGURES . . . . . **viii**

Chapter 1 BACKGROUND INTRODUCTION **1**

Chapter 2 THEORETICAL FRAMEWORK **7**

2.1 Established Results 7

2.1.1 Survival quantities 7

2.1.2 Estimation of survival quantities with complete data 8

2.1.3 Estimation of survival quantities with right censored data . . . 9

2.1.4 Estimation of survival quantities with left truncated data . . . 10

2.1.5 Estimation of survival quantities with right truncated data . . 11

2.2 New Results with Right Truncated Data 14

2.2.1 Nonparametric estimators of the cumulative hazard function . 14

2.2.2 Variance estimation of the estimated cumulative hazard function 16

Chapter 3 SIMULATION STUDIES 19

Chapter 4 THE AIDS EXAMPLE 24

4.1 Data Description 24

4.2 Estimating the cumulative hazard function 26

Chapter 5 CONCLUSIONS 33

REFERENCES 35

LIST OF TABLES

3.1 The simulation results of $\Lambda_L(t)$ for t=0.1, 0.2, 0.3 21

3.2 The simulation results of $\Lambda_L(t)$ for t=0.4, 0.5, 0.6 22

3.3 The simulation results of $\Lambda_L(t)$ for t=0.7, 0.8, 0.9 23

4.1 AIDS transfusion data for children of age 1-4 years 26

4.2 Estimation of $\Lambda_L(t)$ using the naïve estimator 30

4.3 Estimation of $\Lambda_L(t)$ using the Nelson-Aalen type estimators 31

4.4 The 95% confidence intervals for the cumulative hazard function . . . 32

LIST OF FIGURES

4.1 The 95% Confidence intervals of the cumulative hazard rates 28

Chapter 1

BACKGROUND INTRODUCTION

In biomedical studies, one often need to analyze time-to-event data. One major characteristic of time-to-event data is incompleteness. Censored data gives partial information as events occurred to the right or left of a time boundary or within a time interval. It can be further classified into three categories: right censoring, left censoring and interval censoring. Truncation appears when a time to the event is only observed in a study if the time-to-event variable is greater or smaller than the truncation variable. We will describe in detail the following three types of incomplete data: right censored data, left truncated data, and right truncated data in the context of biomedical studies.

Right censoring occurs when a time-to-event is only known to be greater than a censoring time due to end of study, loss to follow-up, or patient's withdrawal. Let T be the failure time and C be the censoring time. In a right censored sample, the failure time is observed if $T \leq C$; otherwise, the censoring time is observed.

With right censored data, the standard nonparametric estimator of the survival function of the failure time was proposed by Kaplan and Meier (1958) using the maximum likelihood approach and assuming the independence between T and C. Let $t_1 < t_2 < \cdots < t_m$ be the distinct failure times. Let d_i be the number of events at time t_i, and Y_i be the size of risk set at t_i. The Kaplan-Meier estimator of the survival function of T is given by

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{Y_i}\right), \qquad t \leq \tau, \tag{1.1}$$

where τ is the largest failure time. The Kaplan-Meier estimates can be visualized as a step function with jumps at the observed event times. The variance of the Kaplan-Meier estimator can be estimated by the well-known Greenwood's formula,

$$\hat{V}[\hat{S}(t)] = \hat{S}(t)^2 \sum_{i:t_i \leq t} \frac{d_i}{Y_i(Y_i - d_i)} \tag{1.2}$$

In addition to the Kaplan-Meier estimator, Efron (1967) proposed a redistribution-to-the-right algorithm to estimate the survival function of the failure time. In Efron's algorithm, the mass of a censored item is equally redistributed to all items to its right. The estimates from Efron's algorithm agree with the Kaplan-Meier estimates. Efron's algorithm is the source of the inverse probability of censoring weighting (IPCW) technique, which has been widely utilized in various contexts with right

censored data. Robertson and Uppuluri (1984) proposed another mass redistribution algorithm by equally redistributing the mass of a censored item to only the failed items to its right. Their algorithm leads to the maximum entropy estimator of $S(t)$. However, this estimator is not commonly used.

Another type of incompleteness is left truncation, also known as late entrance. Let L be the truncation variable and T being the event time. A truncated sample includes realizations of (L, T) subject to the constraint $L \leq T$. Left truncation occurs if the failure time is only included in a sample if it is greater than the truncation time. Two examples are given here to illustrate left truncation. The first example is the lifetime analysis using data from a retirement community in California. Elder residents in certain area of California need to meet the age requirement to enter a retirement community. T is defined as the age of death. The event (death) can only be observed when it occurred after the age of entrance L. In this example, an individual can not be included in the study if he/she died early and was not old enough to enter the retirement center. Therefore, the ages of death, collected from the retirement center, are left-truncated by the ages at entrance. Another example of left truncation appears in bone marrow transplant (BMT) studies using registry data. Leukemia patients are often treated with BMT. A large number of participating hospitals routinely report the new BMT cases and the follow-up information of the previous BMT cases to the International Bone Marrow Transplant

Registry (IBMTR). Those patients who die while waiting for the transplants will not be reported to the IBMTR. The failure time T, defined as the time to failure since the initial diagnosis, can be possibly included in the registry if the failure occurred after the transplantation. Therefore, the samples obtained from IBMTR need to be treated as truncated samples. Please note that, with left truncated samples, the failure times are often also right-censored due to end of study or patient's withdrawal.

The Kaplan-Meier (1958) estimator of the survival function can naturally handle left truncated data by properly defining the risk set. The asymptotic properties of the left-truncated version of the Kaplan-Meier estimator have been studied by Woodroofe (1985), Keiding and Gill (1990) among others.

For a truncated sample (L, T) with the constraint $L \leq T$, right truncation occurs when the variable L is the study of interest and T is the truncation variable. Here the variable L is right truncated by T. A right truncated sample sometimes occur in the clinical research of the latent period of a disease. For example, the latent period of the acquired immune deficiency syndrome (AIDS) is defined as the time from HIV virus infection to the diagnosis of AIDS. Transfusion of contaminated blood is a source of HIV virus. If one received blood transfusion and later was diagnosed of AIDS, it can be traced back when the patient was infected. Let Y denote the calendar time of AIDS blood transfusion and L denote the latent period

of the the HIV virus. Let τ^* be the closing date of the study. A subject can be included in a sample only when the event, AIDS diagnosis, occurs before the study closing date. More specifically, the sample of reported AIDS patients satisfy the condition $Y + L \leq \tau^*$, or equivalently, $L \leq \tau^* - Y$. We therefore consider the AIDS latent time L being right truncated.

A right truncated variable can be converted to a left truncated variable if one reverses the time axis. The cumulative hazard function on the reversed time axis is known as the reverse-time hazard, which can be estimated by the Nelson-Aalen estimator. The statistical inferences on the reverse-time hazard can be easily developed because one may directly use the inferences about a cumulative hazard function of a left truncated variable.

The established inference procedures for a right truncated sample center on the reverse-time hazard. Lagakos et al. (1988) studied the product-limit estimator of the distribution function and proposed a two-sample log-rank test on the reverse-time hazard. Kalbfleisch and Lawless (1989) studies the Cox regression analysis on the reverse-time hazard. However, interpretation about a reverse-time hazard is awkward. Therefore, inferences on the regular forward-time survival quantities have gained increasing interests in recent researches. Chi et al. (2007) developed a two-sample test to compare the Kaplan-Meier estimates of the distribution functions. Shen (2010) proposed the forward-time cumulative hazard function, assuming a

parametric distribution in the truncation variable. However, Shen only suggested the resampling approach for estimating the variance of the test statistic.

The structure of this thesis is organized as follows. In chapter 2, we first review the estimators of the survival function and the cumulative hazard function for various incomplete data. Second, for a right truncated variable, we introduce two Nelson-Aalen type estimators of the cumulative hazard and propose three variance estimator. In Chapter 3, we present the simulation study result to show the performances of the proposed variance estimators. In Chapter 4, an AIDS data set, in which right truncation occurs to the disease latent time, is analyzed to illustrate the details in estimating the forward-time hazard. Finally, the concluding remarks are given in Chapter 5.

Chapter 2

THEORETICAL FRAMEWORK

2.1 Established Results

2.1.1 Survival quantities

For time-to-event data, the most fundamental quantities are the survival and distribution functions, $S(t) = P(T > t)$ and $F(t) = P(T \leq t)$, where T is the failure time variable. The instantaneous probability of failure at time t, given that one survives up to t, is measured by the hazard rate function $\lambda(t)$. It is defined as $\lambda(t) = dF(t)/P(T \geq t)$. The cumulative hazard function is given by

$$\Lambda(t) = \int_0^t \lambda(u)\, du. \tag{2.1}$$

For a continuous failure time variable T, the cumulative hazard function and

survival function have the following relationship:

$$S(t) = \exp[-\Lambda(t)].$$

2.1.2 Estimation of survival quantities with complete data

For a sample with all failure times observed, the sample can be summarized as $\{T_i; i = 1, 2, \cdots n\}$. The well-known empirical estimator can be used for estimating the distribution function,

$$\hat{F}(t) = \frac{1}{n} \sum I(T_i \leq t). \tag{2.2}$$

The cumulative hazard function is commonly estimated by the Nelson-Aalen estimator. It is more convenient to use counting process notation. Let $N_i(t) = I(T_i \leq t)$ and $Y_i(t) = I(T_i \geq t)$. $N_i(t)$ indicates whether the ith subject failed before t and $Y_i(t)$ indicates whether the ith subject remains in the risk set at time t. We further define $\bar{Y}(t) = \sum_{i=1}^{n} Y_i(t)$, which is the size of the risk set at time t. The Nelson-Aalen estimator is given by:

$$\hat{\Lambda}(t) = \sum_{i=1}^{n} \int_0^t \frac{dN_i(s)}{\bar{Y}(s)}. \tag{2.3}$$

2.1.3 Estimation of survival quantities with right censored data

Right censoring occurs when the failure time T is only known to be greater than the censoring time C. The major source of right censoring is end of study. The right censored data can be expressed as $\{(X_i, \Delta_i); i = 1, \cdots n\}$, where $X_i = \min(T_i, C_i), \Delta_i = I(X_i = T_i)$.

The Kaplan-Meier estimator, also known as the product-limit estimator, is routinely employed to estimate the survival function of T with right censored data. We define the following counting process notations: $N_i^C(t) = I(X_i \leq t, \Delta_i = 1)$, $Y_i^C(t) = I(X_i \geq t)$. Let $\bar{N}^C(t) = \sum_{i=1}^n N_i^C(t)$, which is the number of failures at time t. Let $\bar{Y}^C(t) = \sum_{j=1}^n Y_j^C(t)$, which is the size of the risk set at time t. The Kaplan-Meier estimator of $S(t)$ is given by

$$\hat{S}(t) = \prod_{s \leq t} \left(1 - \frac{d\bar{N}^C(s)}{\bar{Y}^C(s)}\right), \qquad 0 < t < \tau. \tag{2.4}$$

$\hat{S}(t)$ is updated to new values only at event times, creating a step function with jumps at the observed event times. The magnitude of a jump at a time depends on the number of events at time t, as well as the size of the risk set at this time.

The cumulative hazard function is estimated by the Nelson-Aalen estimator.

The explicit estimator is given by

$$\hat{\Lambda}(t) = \sum_{i=1}^{n} \int_0^t \frac{dN_i^C(s)}{\bar{Y}^C(s)}. \tag{2.5}$$

2.1.4 Estimation of survival quantities with left truncated data

Left truncation is also known as late entrance. Let L be the time that a subject enters the study. The failure time variable is only observed if $L \leq T$. Please note that, for the truncated variables (L, T) subject to the constraint $L \leq T$, T is left-truncated by L, while L is right-truncated by T.

A truncated sample can be summarized as $\{(L_i, T_i); i = 1, \cdots n\}$, with the constraint $L_i \leq T_i$. Now we define the counting processes, $N_i^T(t) = I(T_i \leq t)$ and $Y_i^L = I(L_i \leq t \leq T_i)$. Let $\bar{N}^T(t) = \sum_{i=1}^n N_i^T(t)$, indicating the number of failures at time t. Let $\bar{Y}^L(t) = \sum_{i=1}^n Y_i^L(t)$ and $\bar{Y}^L(t)$ is the size of the risk set at time t, while the risk set at t contains the subjects entered the study before t and are still under study at t.

Estimation of the survival function of T based on a left-truncated sample can be traced back to the paper by Kaplan and Meier (1958). The left truncated version

of the Kaplan-Meier estimator is given by

$$\hat{S}(t) = \prod_{s \leq t} \left(1 - \frac{d\bar{N}^T(s)}{\bar{Y}^L(s)}\right), \qquad 0 \leq t \leq \tau. \tag{2.6}$$

Using the new risk set, the Nelson-Aalon estimator of the cumulative hazard function is given by

$$\hat{\Lambda}(t) = \sum_{i=1}^{n} \int_0^t \frac{dN_i^T(s)}{\bar{Y}^L(s)}. \tag{2.7}$$

In practice, left truncated and right censored samples occur more frequently than left truncated samples. The above estimators only need to be slightly modified if the right censoring is also present. The details are omitted in the thesis.

2.1.5 Estimation of survival quantities with right truncated data

In Chapter 1, we used the contaminated blood transfusion data to explain right truncation. For the truncated variables (L, T) with the constraint $L \leq T$, L is now of study interest and is right truncated by T. Let τ be the largest observed time in the truncated sample. The transformed variable $L^* = \tau - L$ is left truncated by $\tau - T$. The cumulative hazard function of L^* is named as the reverse-time hazard because it is measured on the reversed time axis. The explicit definition is

$$\Lambda_L^*(t) = \int_t^\tau \frac{dH(s)}{P(L \leq s)}.$$

Note that the estimators used for a left truncated variable (Section 2.1.4) are directly applicable for estimating the survival and cumulative hazard functions of L^*. The distribution function of L, which is equivalent to the survival function of L^*, can be estimated by the Kaplan-Meier estimator. Let $H(t)$ be the distribution function of L. The Kaplan-Meier estimator of $H(t)$ is given by

$$\hat{H}(t) = \prod_{s>t} \left(1 - \frac{d\bar{N}^L(s)}{\bar{Y}^L(s)}\right), \tag{2.8}$$

where $\bar{N}^L(t) = \sum_{i=1}^n N_i^L(t)$ and $N_i^L(t) = I(L_i \leq t)$. The Nelson-Aalen estimator can be directly used to estimate $\Lambda_L^*(t)$. The explicit expression is

$$\hat{\Lambda}_L^*(t) = \sum_{i=1}^n \int_t^\tau \frac{dN_i^L(s)}{\bar{Y}^L(s)}. \tag{2.9}$$

Keiding and Gill (1990, Theorem 5.1) showed the weak convergence result of the Nelson-Aalen estimator for left truncated data. Their result is applicable to $\hat{\Lambda}_L^*(t)$ if we consider the reversed time axis. It can be shown that $n^{1/2}\{\hat{\Lambda}_L^*(t) - \Lambda_L^*(t)\} \longrightarrow W_t$, where W_t is a Gaussian martingale with zero mean and variance σ^2. They suggested a few estimators of σ^2. We provide some details of variance estimators in the following context.

A naïve variance estimator is given by

$$\hat{V}^{(1)}(\hat{\Lambda}_L^*(t)) = \sum_{i=1}^{n} \int_t^{\tau} \frac{dN_i^L(t)}{\bar{Y}^L(s)^2}. \tag{2.10}$$

A variance estimator of survival estimate was studied by Klein (1991) for right censored data. The analogous variance estimator for $\hat{\Lambda}_L^*(t)$ is given by

$$\hat{V}^{(2)}(\hat{\Lambda}_L^*(t)) = \sum_{i=1}^{n} \int_t^{\tau} \frac{(\bar{Y}^L(s) - \Delta N_i^L(s))dN_i^L(s)}{\bar{Y}^L(s)^3}. \tag{2.11}$$

Compared to the naïve variance estimator, Klein's variance estimator tends to give a smaller estimate of variation for right censored data. Another variance estimator of $\hat{\Lambda}_L^*(t)$, based on the Greenwood's formula, is given by

$$\hat{V}^{(3)}(\hat{\Lambda}_L^*(t)) = \sum_{i=1}^{n} \int_t^{\tau} \frac{dN_i^L(s)}{\bar{Y}^L(s)(\bar{Y}^L(s) - \Delta N_i^L(s))}. \tag{2.12}$$

The established statistical inferences using the reverse-time hazard include two-sample log-rank test (Lagakos et al.,1988; Chi et al., 2007), and Cox regression model (Kalbfleisch and Lawless, 1972).However, the reverse-time hazard is not convenient to interpret. The statistical inferences about the regular forward-time cumulative hazard function is practically needed. Shen (2010) recently studied the two-sample test on the forward-time cumulative hazard, assuming some parametric distributions

for the truncated variable. However, Shen only suggested the resampling method for estimating the variance of the test statistic.

Let $\Lambda_L(t)$ be the forward-time cumulative hazard function of L, with the definition,

$$\Lambda_L(t) = \int_0^t \frac{dH(s)}{P(L \geq s)}. \tag{2.13}$$

In order to render simple presentation, we assume no ties in the truncated sample. According to its definition, a naïve estimator of $\Lambda_L(t)$ is given by

$$\hat{\Lambda}_L^{(1)}(t) = \int_0^t \frac{d\hat{H}(s)}{1 - \hat{H}(s-)}. \tag{2.14}$$

This estimator has been practically utilized (Lagakos et al., 1988). However, the inference about this estimator is scarce.

2.2 New Results with Right Truncated Data

2.2.1 Nonparametric estimators of the cumulative hazard function

In this section, we introduce two Nelson-Aalen type estimators of $\Lambda_L(t)$. Define a new at-risk indicator $Y_i^{L*}(t) = I(L_i \geq t)$. One estimator can be derived from

Geskus' work on the left-truncated variable (2010). The estimator is given by

$$\hat{\Lambda}_L^{(2)}(t) = \sum_{i=1}^{n} \int_0^t \frac{\hat{w}^{(1)}(L_i) dN_i^L(s)}{\sum_{j=1}^{n} \hat{w}^{(1)}(L_j) Y_j^{L*}(s)}, \tag{2.15}$$

where

$$\hat{w}^{(1)}(t) = \frac{1}{\hat{P}} \times \frac{1}{\hat{S}(t-)}, \qquad \hat{P} = \sum_{i=1}^{n} \frac{1}{\hat{S}(L_i-)}.$$

Zhang, Zhang and Fine (2009) studied the other type of weighing for left-truncated variable. Their method suggests the other estimator of $\Lambda_L(t)$,

$$\hat{\Lambda}_L^{(3)}(t) = \sum_{i=1}^{n} \int_0^t \frac{\hat{w}^{(2)}(L_i) dN_i^L(s)}{\sum_{j=1}^{n} \hat{w}^{(2)}(L_j) Y_j^{L*}(s)}. \tag{2.16}$$

where

$$\hat{w}^{(2)}(t) = \frac{\hat{H}(t)}{\bar{Y}^L(t)}.$$

The above two estimators yield the same numerical values as the naïve estimator. The relevant proofs can be found in Geskus (2010), Zhang, Zhang and Fine (2009) for a variable subject to left truncation. Based on the new estimators, one can easily identify the appropriate weights for the statistical problems with a right truncated sample. The weights can be utilized in various contexts regarding the inferences of the forward-time hazard of the right truncated variable. One potential application is the regression analysis on the forward-time hazard. In Chapter 4, all

three estimators are applied to an AIDS data set and it can be verified that the numerical results are the same.

2.2.2 Variance estimation of the estimated cumulative hazard function

Analyses on the right-truncated data often employ the reversed time scale. It can be easily seen that the relation between $H(t)$ and $\Lambda_L^*(t)$ is $H(t) = \exp(-\Lambda_L^*(t))$. Also note that $1 - H(t) = \exp(-\Lambda_L(t))$. One can further derive the following relation between $\Lambda_L(t)$ and $\Lambda_L^*(t)$,

$$\Lambda_L(t) = -log\left[1 - \exp(-\Lambda_L^*(t))\right]. \tag{2.17}$$

That is, the forward-time hazard is a function of the reverse-time hazard. Using the weak convergence result of $\hat{\Lambda}_L^*(t)$ and applying the delta method, We have the result

$$n^{1/2}\{\hat{\Lambda}_L(t) - \Lambda_L(t)\} \longrightarrow g(\Lambda_L^*(t))W_t.$$

where

$$g(\Lambda_L^*(t)) = -\frac{\exp(-\Lambda_L^*(t))}{1 - \exp(-\Lambda_L^*(t))} = -\frac{H(t)}{1 - H(t)}.$$

and W_t is a zero-mean normally distributed martingale. Thus, we have the variance of $\hat{\Lambda}_L(t)$ can be expressed as

$$V\left[\hat{\Lambda}_L(t)\right] \approx \left[\frac{H(t)}{1-H(t)}\right]^2 V[\hat{\Lambda}_L^*(t)]. \tag{2.18}$$

This variance has not been suggested in previous studies. In Section 2.1.5, we presented three different variance estimators for the estimated reverse-time hazard. Plugging in these variance estimators into the above formula, we can obtain three variance estimators for $\hat{\Lambda}_L(t)$. The explicit formulas of these three variance estimators are

$$\hat{V}^{(1)}\left[\hat{\Lambda}_L(t)\right] \approx \left[\frac{\hat{H}(t)}{1-\hat{H}(t)}\right]^2 \sum_{i=1}^{n} \int_t^\tau \frac{dN_i^L(s)}{\bar{Y}^L(s)^2}, \tag{2.19}$$

$$\hat{V}^{(2)}\left[\hat{\Lambda}_L(t)\right] \approx \left[\frac{\hat{H}(t)}{1-\hat{H}(t)}\right]^2 \sum_{i=1}^{n} \int_t^\tau \frac{(\bar{Y}^L(s) - \Delta N_i^L(s))dN_i^L(s)}{\bar{Y}^L(s)^3}, \tag{2.20}$$

and

$$\hat{V}^{(3)}\left[\hat{\Lambda}_L(t)\right] \approx \left[\frac{\hat{H}(t)}{1-\hat{H}(t)}\right]^2 \sum_{i=1}^{n} \int_t^\tau \frac{dN_i^L(s)}{\bar{Y}^L(s)(\bar{Y}^L(s) - \Delta N_i^L(s))}. \tag{2.21}$$

The variance of $\hat{\Lambda}_L(t)$ (Equation (2.18)) will increase dramatically as t approaches the largest observed time of L. This is the limitation of the given variance estimators. The simulation studies in Chapter 3 reveals poor tail performance.

The result can be easily extended to the two-sample test on the cumulative hazard function. Let $\Lambda_L^1(t)$ and $\Lambda_L^2(t)$ be the cumulative hazard functions of two independent truncated samples. Considering the hypotheses $H_0 : \Lambda_L^1(t) = \Lambda_L^2(t)$ vs $H_a : \Lambda_L^1(t) \neq \Lambda_L^2(t) \quad \forall t < \tau$. One can define the test statistic as

$$z = \frac{\hat{\Lambda}_L^1(t) - \hat{\Lambda}_L^2(t)}{\sqrt{\hat{V}[\hat{\Lambda}_L^1(t)] + \hat{V}[\hat{\Lambda}_L^2(t)]}}. \tag{2.22}$$

This test statistic asymptotically has a standard normal distribution.

Chapter 3

SIMULATION STUDIES

We conducted a simulation study to investigate the performances of three variance estimators of the estimated cumulative hazard function with right-truncated data. In this simulation, we estimate the cumulative hazard function, calculate the deviation from the true values, the variance estimators and the coverage probabilities.

In this study, the underlying distribution of the variable L was generated from a uniform distribution in the interval [0,1]. The cumulative hazard function of this uniform distribution is given by

$$\Lambda_L(t) = -log(1-t), \quad 0 \leq t \leq 1. \tag{3.1}$$

The truncation variable T was generated from an exponential distribution, where $F(t) = 1 - \exp(-\lambda t), t > 0$. The parameter λ was searched to yield the

predetermined truncation percentages in the sample. We considered two truncation percentages, 25% and 50%. For each truncation percentage, settings with sample size of 50, 100 and 200 were obtained and each setting contains 1000 replicates. Let $\hat{\Lambda}_{L,i}(t)$ be the cumulative hazard estimate for the ith replicate at t. Let $\bar{\hat{\Lambda}}_L(t)$ denote the average cumulative hazard estimate across 1000 replicates, where $\bar{\hat{\Lambda}}_L(t) = \sum_{i=1}^{n} \hat{\Lambda}_{L,i}(t)$. We report the estimation result at $t = 0.1, 0.2, ..., 0.9$.

The bias is defined as the deviation between the average cumulative hazard estimate and the true value. The sampling standard deviation (SSD) was calculated to reflect the degree of variability among 1000 cumulative hazard estimates. The estimated standard error (ESE) is obtained by finding the average of the standard error estimates in 1000 replicates. For a given variance estimator, the 95% confidence interval have been calculated for each sample and the actual coverage rate across 1000 samples was obtained. We also report the average length of the 95% confidence interval (Avg CIL). The explicit formulas for calculating the above quantities are given as follows,

$$\text{Bias} = \bar{\hat{\Lambda}}_L(t) - \Lambda_L(t), \tag{3.2}$$

$$\text{RBias} = \text{Bias}/\Lambda_L(t),$$

Table 3.1. The simulation results of $\Lambda_L(t)$ for t=0.1, 0.2, 0.3

						Naive			Greenwood			Klein		
t	n	L%	Bias	RBias	SSD	ESE	Coverage	Avg CIL	ESE	Coverage	Avg CIL	ESE	Coverage	Avg CIL
0.1	50	25	0.000	0.001	0.044	0.045	0.921	0.178	0.048	0.942	0.188	0.043	0.899	0.168
		50	-0.003	-0.026	0.042	0.042	0.908	0.163	0.043	0.923	0.170	0.040	0.898	0.157
	100	25	0.000	0.000	0.032	0.032	0.919	0.124	0.032	0.925	0.127	0.031	0.912	0.121
		50	-0.003	-0.024	0.029	0.029	0.917	0.115	0.030	0.922	0.117	0.029	0.914	0.113
	200	25	0.000	-0.003	0.022	0.022	0.947	0.087	0.022	0.948	0.088	0.022	0.945	0.086
		50	0.001	0.006	0.022	0.022	0.938	0.083	0.021	0.943	0.083	0.021	0.936	0.082
0.2	50	25	-0.001	-0.004	0.071	0.071	0.929	0.276	0.073	0.935	0.285	0.068	0.917	0.269
		50	-0.001	-0.003	0.072	0.072	0.920	0.279	0.073	0.927	0.287	0.069	0.915	0.272
	100	25	-0.001	-0.005	0.049	0.072	0.938	0.194	0.050	0.941	0.196	0.049	0.934	0.191
		50	-0.001	-0.009	0.050	0.050	0.939	0.196	0.051	0.941	0.198	0.049	0.937	0.193
	200	25	-0.001	-0.001	0.035	0.035	0.938	0.137	0.035	0.939	0.138	0.035	0.937	0.136
		50	-0.001	0.005	0.035	0.036	0.946	0.140	0.036	0.946	0.140	0.035	0.943	0.139
0.3	50	25	-0.001	-0.004	-0.097	0.096	0.946	0.375	0.098	0.950	0.384	0.094	0.937	0.367
		50	0.001	0.003	0.104	0.102	0.922	0.402	0.105	0.932	0.412	0.100	0.913	0.392
	100	25	-0.002	-0.005	0.066	0.067	0.936	0.263	0.068	0.937	0.265	0.066	0.935	0.260
		50	-0.002	-0.004	0.074	0.072	0.939	0.281	0.073	0.945	0.285	0.071	0.934	0.278
	200	25	0.000	-0.001	0.049	0.047	0.939	0.185	0.048	0.941	0.186	0.047	0.938	0.184
		50	0.002	0.004	0.050	0.051	0.950	0.199	0.051	0.950	0.201	0.051	0.948	0.198

$$\text{SSD} = \sqrt{\frac{1}{1000-1}\sum_{i=1}^{1000}(\hat{\Lambda}_{L,i}(t) - \bar{\hat{\Lambda}}_L(t))^2}, \tag{3.3}$$

$$\text{ESE} = \frac{1}{1000}\sum_{i=1}^{1000}\sqrt{\hat{V}^{(k)}[\hat{\Lambda}_L(t)]}, \qquad k \in 1, 2, 3. \tag{3.4}$$

Tables 3.1-3.3 show the simulation results of the Nelson-Aalen estimator of the cumulative hazard function and three variance estimators. According to these tables, the largest absolute value of relative biases are no more than 2.6%. In summary, all three variance estimators perform equally well. The estimated standard errors (ESE) using either variance estimator closely match the sampling standard deviation (SSD)

Table 3.2. The simulation results of $\Lambda_L(t)$ for t=0.4, 0.5, 0.6

						Naive			Greenwood			Klein		
t	n	L%	Bias	RBias	SSD	ESE	Coverage	Avg CIL	ESE	Coverage	Avg CIL	ESE	Coverage	Avg CIL
0.4	50	25	0.000	-0.001	0.124	0.124	0.941	0.486	0.126	0.946	0.495	0.122	0.940	0.477
		50	0.004	0.009	0.138	0.138	0.939	0.544	0.142	0.945	0.557	0.136	0.932	0.531
	100	25	0.001	0.002	0.084	0.084	0.950	0.339	0.087	0.953	0.343	0.086	0.947	0.336
		50	-0.001	0.002	0.097	0.097	0.948	0.380	0.098	0.954	0.384	0.096	0.947	0.375
	200	25	-0.000	0.001	0.062	0.062	0.946	0.238	0.061	0.947	0.239	0.061	0.945	0.237
		50	0.003	0.006	0.065	0.065	0.962	0.268	0.069	0.962	0.269	0.068	0.959	0.266
0.5	50	25	-0.002	-0.003	0.156	0.156	0.940	0.613	0.159	0.946	0.624	0.154	0.935	0.603
		50	0.004	0.006	0.179	0.182	0.947	0.713	0.186	0.951	0.730	0.178	0.939	0.697
	100	25	0.000	0.000	0.106	0.109	0.951	0.427	0.110	0.951	0.431	0.108	0.948	0.424
		50	-0.001	-0.001	0.125	0.126	0.955	0.496	0.128	0.956	0.501	0.125	0.953	0.490
	200	25	0.001	0.001	0.077	0.076	0.944	0.300	0.077	0.946	0.301	0.076	0.943	0.299
		50	0.002	0.002	0.085	0.089	0.952	0.348	0.089	0.953	0.350	0.088	0.951	0.346
0.6	50	25	-0.001	0.001	0.200	0.199	0.937	0.780	0.202	0.942	0.793	0.196	0.934	0.768
		50	0.009	0.009	0.236	0.240	0.945	0.943	0.246	0.953	0.965	0.235	0.943	0.921
	100	25	0.001	0.001	0.133	0.137	0.956	0.539	0.138	0.959	0.543	0.136	0.954	0.534
		50	0.001	0.001	0.161	0.165	0.963	0.648	0.167	0.964	0.655	0.163	0.961	0.640
	200	25	0.001	0.001	0.095	0.096	0.948	0.376	0.096	0.949	0.378	0.096	0.948	0.375
		50	0.004	0.004	0.108	0.115	0.958	0.453	0.116	0.958	0.455	0.115	0.957	0.450

and the coverages maintain the 95 % level. However, we observe a large deviation between ESE and SSD when t is beyond 0.8. The Greenwood variance estimator tends to yield the largest estimates of the variations. The Klein's estimator gives a slightly smaller estimates of the variations.

Across the three estimators, we also observed some general trends. First, SSD decreases systematically as the sample size increases. Second, we observed an increase of SSD with increases of truncation percentage for the same sample size. This increase due to increasing truncation percentage is amplified when the survival function approaches to 0 ($t = 1$).

Table 3.3. The simulation results of $\Lambda_L(t)$ for t=0.7, 0.8, 0.9

						Naïve			Greenwood			Klein		
t	n	L%	Bias	RBias	SSD	ESE	Coverage	Avg CIL	ESE	Coverage	Avg CIL	ESE	Coverage	Avg CIL
0.7	50	25	0.004	0.003	0.254	0.260	0.945	1.108	0.264	0.948	1.034	0.256	0.941	1.002
		50	0.012	0.010	0.320	0.328	0.961	1.284	0.335	0.966	1.315	0.320	0.957	1.255
	100	25	-0.001	-0.001	0.170	0.176	0.949	0.691	0.178	0.951	0.697	0.175	0.947	0.686
		50	-0.003	-0.003	0.207	0.219	0.956	0.859	0.221	0.961	0.868	0.216	0.956	0.848
	200	25	-0.002	0.002	0.121	0.123	0.948	0.481	0.123	0.948	0.483	0.122	0.947	0.480
		50	0.003	0.002	0.143	0.152	0.961	0.597	0.153	0.962	0.600	0.151	0.960	0.593
0.8	50	25	0.014	0.008	0.356	0.369	0.961	1.445	0.374	0.963	1.466	0.363	0.956	1.423
		50	0.032	0.020	0.467	0.497	0.953	1.948	0.509	0.956	1.995	0.485	0.950	1.903
	100	25	-0.001	-0.001	0.237	0.240	0.953	0.943	0.242	0.954	0.950	0.239	0.952	0.936
		50	-0.003	-0.003	0.290	0.311	0.954	1.221	0.315	0.960	1.235	0.308	0.948	1.206
	200	25	0.000	0.000	0.162	0.165	0.952	0.647	0.166	0.952	0.650	0.165	0.951	0.645
		50	0.006	0.004	0.202	0.212	0.954	0.832	0.214	0.956	0.837	0.211	0.953	0.828
0.9	50	25	0.004	0.002	0.523	0.650	0.956	2.546	0.659	0.960	2.583	0.640	0.955	2.510
		50	-0.003	-0.001	0.651	0.854	0.950	3.349	0.875	0.954	3.430	0.834	0.944	3.270
	100	25	0.001	0.000	0.364	0.398	0.960	1.560	0.401	0.961	1.572	0.395	0.959	1.549
		50	0.000	0.000	0.474	0.563	0.957	2.208	0.570	0.960	2.234	0.556	0.956	2.181
	200	25	0.002	0.001	0.253	0.261	0.952	1.023	0.262	0.952	1.026	0.260	0.951	1.019
		50	0.004	0.002	0.325	0.353	0.958	1.383	0.355	0.959	1.391	0.351	0.957	1.375

Chapter 4

THE AIDS EXAMPLE

4.1 Data Description

In this chapter, we analyze a data set obtained from 295 AIDS patients who were infected by HIV by blood or blood product transfusion. The source of data was Center for Disease Control (CDC) that administrates a national registry of AIDS patients. Several groups of researchers used this data set as an example of right truncation (Lagakos et al, 1988, Chi et al, 2007, Shen, 2010). The data set contains the sex and age of the patient, the date of diagnosis and the date of blood transfusion. The closing date of study was July 1, 1986. Patients diagnosed of AIDS after that date are excluded from the data set.

The variable of study interest is the latent period of HIV virus L, which is defined as the time from the transfusion to the diagnosis of AIDS. Since AIDS has a latent period that varies from months to years, these people who are not be

diagnosed of AIDS before the study terminates are therefore excluded from CDC's registry of AIDS patients. Only these patients diagnosed before the end of study are included in the sample. Therefore, the latent time of HIV virus is right truncated by the time between HIV virus infection and the end of study. More specifically, let Y be the calendar time of blood transfusion and τ be the study closing date. In order to be included in the study, Y plus the latent period L cannot exceed τ, that is $Y + L \leq \tau$. Therefore, L is right truncated by $\tau - Y$.

In this thesis, we focused on the subset of children. Table 4.1 shows the latent times and truncation times of 34 children aged at 1-4 years. In the table, the truncation time variable $\tau - Y$ is denoted as T. First, we implemented the following three estimators of cumulative hazard function: the naive estimator, the estimator using Geskus's weight(2010) and the estimator using the weight suggested by Zhang, Zhang and Fine (2009). Although these three estimators yield identical numerical values, we use this example to illustrate how to calculate the weights. Second, we calculated three variance estimators described in Chapter 2.2.1 and obtained three sets of confidence intervals.

Table 4.1. AIDS transfusion data for children of age 1-4 years

L	T	Age	L	T	Age	L	T	Age
28	80	4	21	48	3	17	26	2
14	64	2	8	31	1	8	16	1
10	57	1	33	54	3	11	19	4
10	54	1	13	34	2	15	22	2
23	63	2	8	26	1	10	17	1
13	52	2	20	37	2	4	11	1
12	49	2	37	53	4	32	38	3
37	71	4	20	35	2	23	29	3
6	38	2	18	33	2	32	33	3
4	35	1	8	22	1	10	13	1
13	40	2	27	40	4			
11	38	1	43	52	4			

4.2 Estimating the cumulative hazard function

In this section, we present some details of estimating the cumulative hazard function of a right truncated variable. Table 4.2 shows the estimation of $\Lambda_L(t)$ using the naïve estimators. The naïve estimator of $\Lambda_L(t)$ uses distribution function estimates. At an event time, the increase in $\Lambda_L(t)$ can be estimated by the increase in the distribution function divided by the estimate of $P(L \geq t)$. In Table 4.2, we included the numbers of death and sizes of risk sets at individual event times. These quantities are needed for estimating the distribution function. The table also shows

the estimated cumulative distribution probabilities, as well as the increases at event times. The estimated cumulative hazard function is included in the last column in the table, also plotted in Figure 4.1. The figure shows increase at a lower rate before 30 months but increase much faster after 30 months.

Table 4.3 shows the cumulative hazard function through the Nelson-Aalen type estimators, using two individual weights suggested from Section 2.2.1. The table includes the number of deaths, the size of risk set, estimated distribution function of disease latent time ($\hat{H}(t)$), and the estimated survival probability of the truncation variable prior to t ($\hat{S}(t-)$). Geskus' weight uses the reciprocal of $\hat{S}(t-)$, multiplying the normalization constant $\frac{1}{\hat{P}}$. Note that the calculation formula for $\hat{P}$ is given by $\hat{P} = \sum_{i=1}^{n} \frac{1}{\hat{S}(L_i-)}$. It is evaluated to be 47.848. Geskus' weight is shown as $w^{(1)}(t)$ in the table. The weight suggested by Zhang, Zhang and Fine can be directly calculated by diving $\hat{H}(t)$ by the size of risk set. The evaluated values are proved in $w^{(2)}(t)$ in the table, which coincides with Geskus' weights. The next step is to use the calculated weights to find the adjusted death and adjusted risk set. Finally, the Nelson-Aalen type estimator of $\Lambda_L(t)$ is evaluated using the the adjust death and risk set.

Table 4.4 shows the standard errors from three variance estimators and the 95% confidence intervals. Although three confidence intervals are very close, Klein's variance estimator yields the narrowest sets of confidence intervals and Greenwood

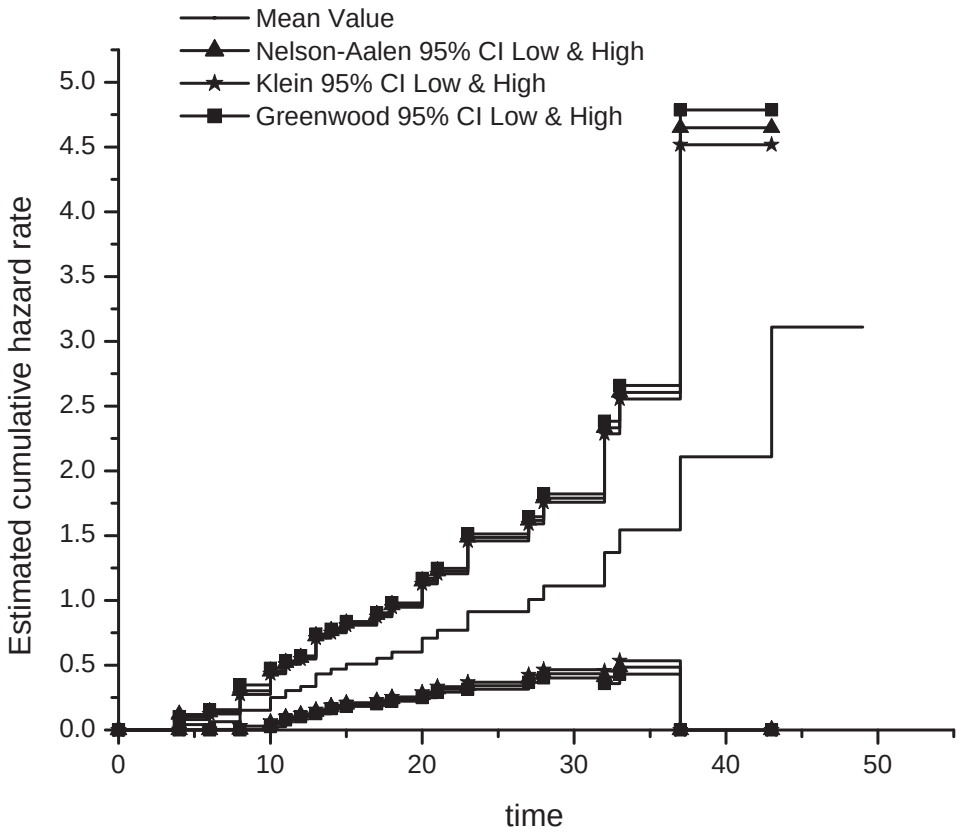

Figure 4.1. The 95% Confidence intervals of the cumulative hazard rates

variance estimator yields the widest confidence intervals. As time gets larger, the confidence intervals become much wider. When the time is beyond 35 months, the lower bound of cumulative hazard function will become negative, which in unrealistic. This reflects the limitation of the three variance estimators. The 95% confidence intervals are plotted in Figure 4.1. It can be easily seen from the figure that the estimated variances increase dramatically as time approached the largest observed

time. We suggest not to employ the proposed estimators for the end of study. Based on the simulation study, it is practically acceptable to use these variance estimators when the estimated survival probability is beyond 0.1.

Table 4.2. Estimation of $\Lambda_L(t)$ using the naïve estimator

t	#death	#at risk	$\hat{H}(t)$	Increase in $\hat{H}(t)$	$\hat{\Lambda}_L(t)$
4	2	2	0.042	0.042	0.042
6	1	3	0.063	0.021	0.064
8	4	7	0.146	0.083	0.153
10	4	11	0.230	0.084	0.251
11	2	13	0.273	0.043	0.305
12	1	13	0.294	0.021	0.336
13	3	16	0.362	0.068	0.432
14	1	16	0.386	0.024	0.470
15	1	17	0.411	0.025	0.510
17	1	17	0.436	0.025	0.553
18	1	17	0.463	0.027	0.601
20	2	18	0.521	0.058	0.709
21	1	19	0.550	0.029	0.770
23	2	19	0.615	0.065	0.914
27	1	18	0.651	0.035	1.008
28	1	19	0.688	0.037	1.112
32	2	19	0.768	0.080	1.371
33	1	20	0.809	0.041	1.545
37	2	17	0.917	0.108	2.109
43	1	12	1.000	0.083	3.109

Table 4.3. Estimation of $\Lambda_L(t)$ using the Nelson-Aalen type estimators

t	#death	#at risk	$\hat{S}(t-)$	$\frac{1}{\hat{S}(t-)}$	$\hat{w}^{(1)}(t)$	$\hat{w}^{(2)}(t)$	$\hat{H}(t)$	adjust death	adjust riskset	$\hat{\Lambda}_L(t)$
4	2	2	1.000	1.000	0.021	0.021	0.042	0.042	1.000	0.042
6	1	3	1.000	1.000	0.021	0.021	0.063	0.209	0.959	0.064
8	4	7	1.000	1.000	0.021	0.021	0.146	0.084	0.937	0.153
10	4	11	1.000	1.000	0.021	0.021	0.230	0.084	0.854	0.251
11	2	13	1.000	1.000	0.021	0.021	0.272	0.042	0.770	0.305
12	1	13	0.923	1.083	0.023	0.023	0.294	0.023	0.728	0.336
13	3	16	0.923	1.083	0.023	0.023	0.362	0.068	0.706	0.432
14	1	16	0.865	1.156	0.024	0.024	0.386	0.024	0.638	0.470
15	1	17	0.865	1.156	0.024	0.024	0.411	0.024	0.613	0.509
17	1	17	0.814	1.228	0.025	0.025	0.436	0.026	0.589	0.553
18	1	17	0.767	1.305	0.027	0.027	0.463	0.027	0.564	0.601
20	2	18	0.721	1.386	0.028	0.028	0.521	0.058	0.537	0.709
21	1	19	0.721	1.386	0.028	0.028	0.550	0.029	0.479	0.770
23	2	19	0.646	1.549	0.032	0.032	0.615	0.065	0.450	0.914
27	1	18	0.578	1.731	0.036	0.036	0.651	0.036	0.385	1.008
28	1	19	0.578	1.731	0.036	0.036	0.688	0.036	0.349	1.112
32	2	19	0.517	1.935	0.040	0.040	0.768	0.081	0.313	1.371
33	1	20	0.517	1.935	0.040	0.040	0.809	0.040	0.232	1.545
37	2	17	0.388	2.580	0.054	0.054	0.917	0.108	0.191	2.109
43	1	12	0.251	3.987	0.083	0.083	1.000	0.083	0.083	3.109

Table 4.4. The 95% confidence intervals for the cumulative hazard function

		Nelson-Aalen			Klein			Greenwood		
t	$\hat{\Lambda}_L(t)$	SE	CI Low	CI High	SE	CI Low	CI High	SE	CI Low	CI High
4	0.042	0.039	-0.035	0.119	0.020	0.003	0.081	0.031	-0.019	0.103
6	0.064	0.037	-0.010	0.137	0.030	0.004	0.123	0.048	-0.030	0.157
8	0.153	0.077	0.002	0.303	0.062	0.030	0.275	0.100	-0.043	0.348
10	0.251	0.103	0.049	0.453	0.093	0.066	0.434	0.115	0.025	0.476
11	0.305	0.109	0.091	0.519	0.103	0.102	0.508	0.116	0.078	0.532
12	0.336	0.114	0.113	0.559	0.108	0.125	0.547	0.120	0.101	0.571
13	0.432	0.148	0.141	0.723	0.141	0.157	0.708	0.157	0.125	0.739
14	0.470	0.150	0.177	0.764	0.143	0.189	0.751	0.156	0.164	0.777
15	0.510	0.160	0.196	0.823	0.153	0.209	0.809	0.167	0.182	0.837
17	0.553	0.172	0.217	0.889	0.164	0.232	0.874	0.180	0.201	0.905
18	0.601	0.185	0.239	0.963	0.176	0.256	0.947	0.194	0.222	0.981
20	0.709	0.224	0.271	1.148	0.214	0.291	1.128	0.235	0.249	1.170
21	0.770	0.232	0.314	1.226	0.222	0.335	1.205	0.243	0.293	1.247
23	0.914	0.292	0.342	1.486	0.278	0.369	1.459	0.306	0.314	1.514
27	1.008	0.311	0.398	1.618	0.297	0.425	1.591	0.326	0.369	1.647
28	1.112	0.346	0.434	1.789	0.329	0.466	1.758	0.363	0.401	1.822
32	1.371	0.491	0.408	2.333	0.467	0.456	2.286	0.517	0.358	2.383
33	1.545	0.541	0.484	2.606	0.516	0.535	2.556	0.568	0.431	2.659
37	2.109	1.295	-0.429	4.648	1.228	-0.298	4.517	1.366	-0.568	4.786
43	3.109	.	.	.	.	.	.	.	.	.

Chapter 5

CONCLUSIONS

In this thesis, we described the alternative methods to estimate the cumulative hazard function for a right truncated variable. Some current methods for problems with right truncated variable use the "reversed" time scale. The works on the forward-time cumulative hazard function are limited. A simulation was conducted to investigate actual performances of three variance estimators. The simulation results demonstrate that the estimated standard error closely matches the sampling standard deviation when an estimated survival probability is not close to 1.

We estimated the cumulative hazard function for an example of an AIDS study, where the HIV latent time is right truncated by the time interval between the blood transfusion date and study closing date. Three estimators were used to calculate the cumulative hazard functions, yielding identical values. Three variance estimators for the estimated cumulative incidence function were calculated and we obtained three

sets of confidence intervals.

The future directions of this work are discussed as follows. First, the weights explained in this thesis can be used in regression analyses. For example, the Cox model on the forward-time hazard rate function of a right truncated data set would require usage of weights. This work reveals the explicit forms of the weights. Second, the three variance estimators studied in this thesis can be directly extended to a two-sample test on the cumulative hazard function. Such a test can be treated as a special case of the two-sample test studied by Shen (2010). However, Shen only suggested the resampling approach to estimate the variance of the test statistic. We can derive the analytical result of the variance using the variance estimators described in this thesis. Finally, further consideration and investigation on the variance estimators is needed to solve the problem of poor tail performance.

REFERENCES

[1] Chi, Y., Tsai, W. Y. and Chiang, C. L. Testing the equality of two survival functions with right truncated data, *Statistics in Medicine*, Vol. 26, pp. 812-827, 2007.

[2] Efron, B. The Two Sample Problem with Censored Data, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 4, pp. 831-853, 1967.

[3] Geskus, R. B. Cause-specific cumulative incidence estimation and the Find and Gray model under both left truncation and right censoring, *Biometrics*, doi: 10.1111/j.1541-0420.2010.01420.x, 2010.

[4] Kalbfleisch, J. D. and Lawless, J. F. Inference based on retrospective Ascertainment: an analysis of the data on tranfusion-related AIDS, *Journal of the American Statistical Association*, Vol. 84, pp. 360-372, 1989.

[5] Kaplan, E. L. and Meier, P. Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association*, Vol. 53, pp. 457-481, 1958.

[6] Keiding, N. and Gill, R. D. Random truncation models and Markov process, *The Annals of Statistics*, Vol. 18, pp. 582-602, 1990.

[7] Lagakos, S. W., Barraj, L. M. and Gruttola, V. Nonparametric analysis of truncated survival data with applications to AIDS, *Biometrika*, Vol. 75, pp. 515-523, 1988.

[8] Qin, J. and Shen, Y. Statistical methods for analyzing right-censored length-biased data under cox model, *Biometrics*, Vol. 66, pp. 382-392, 2010.

[9] Robertson, J. B. and Uppuluri, V. R. R. A Generalized Kaplan-Meier Estimator, *The Annals of Statistics*, Vol. 12, pp. 366-371, 1984.

[10] Shen, P. A class of semiparametric rank-based tests for right-truncated data, *Statistics and Probability Letters*, Vol. 80, pp. 1459-1466, 2010.

[11] Woodroof, M. Estimating a distribution function with truncated Data, *The Annals of Statistics*, Vol. 13, pp. 163-177, 1985.

[12] Zhang, X., Zhang, M. J. and Fine, J. P. A mass redistribution algorithm for right-censored and left-truncated time to event data, *Journal of Statistical Planning and Inference*, Vol. 139, pp. 3329-3339, 2009.