Summer 8-1-2011

Analysis of Dependently Truncated Sample Using Inverse Probability Weighted Estimator

Yang Liu

ANALYSIS OF DEPENDENTLY TRUNCATED SAMPLE USING INVERSE PROBABILITY WEIGHTED ESTIMATOR

by

YANG LIU

Under the Direction of Dr. Xu Zhang

ABSTRACT

Many statistical methods for truncated data rely on the assumption that the failure and truncation time are independent, which can be unrealistic in applications. The study cohorts obtained from bone marrow transplant (BMT) registry data are commonly recognized as truncated samples, the time-to-failure is truncated by the transplant time. There are clinical evidences that a longer transplant waiting time is a worse prognosis of survivorship. Therefore, it is reasonable to assume the dependence between transplant and failure time. To better analyze BMT registry data, we utilize a Cox analysis in which the transplant time is both a truncation variable and a predictor of the time-to-failure. An inverse-probability-weighted (IPW) estimator is proposed to estimate the distribution of transplant time. Usefulness of the IPW approach is demonstrated through a simulation study and a real application.

INDEX WORDS: Left truncation, Dependent, Inverse probability weighting, Cox regression model, SAS programming

ANALYSIS OF DEPENDENTLY TRUNCATED SAMPLE USING INVERSE PROBABILITY WEIGHTED ESTIMATOR

by

YANG LIU

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

ANALYSIS OF DEPENDENTLY TRUNCATED SAMPLE USING INVERSE PROBABILITY WEIGHTED ESTIMATOR

by

YANG LIU

Committee Chair:	Dr. Xu Zhang
Committee:	Dr. Gengsheng Qin
	Dr. Yichuan Zhao

Electronic Version Approved:

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

ACKNOWLEDGEMENTS

I gratefully acknowledge to all those who made it possible for me to complete this thesis.

My first, and most earnest, acknowledgment must go to my advisor, Dr. Xu Zhang, for her constant guidance during my thesis time. Her perpetual energy and enthusiasm motivated me to go through the difficult times in my research. She pays attention to the every detail problem and always tries the new method. Without her patient direction and valuable suggestions, this work would have been impossible.

It is a great honor for me that Dr. Gengsheng Qin and Dr. Yichuan Zhao would like to be my thesis committee members. Thanks for their valuable advice, heartfelt encouragement and generous accessibility during my graduate career. Their excellent teaching also opened a door for me to this fantastic world of statistics. I was extraordinarily fortunate for having these wonderful professors in Statistical department.

A sincere and sweet thank-you goes to my loving wife Xiaoling, son Kevin and my parents. No words can nearly express the full measure of my appreciation to their deep love, which enabled me to complete this work.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

Data truncation is a problem in scientific investigation. Truncation is one type of incompleteness which occurs when the incomplete nature of the observation is due to a systematic selection process inherent to the study design. A truncated sample includes realizations of (L, T) subject to the constraint $L \leq T$. Within the scope of life science, T is often the failure time, and L is the entrance time indicating that the subject enters the study. Two types of truncation, left truncation and right truncation, coexist in a truncated sample. In terms of the failure time and entrance time, left truncation occurs if a failure time is greater than the entrance time. T is left truncated by L. As a consequence, left truncation is also known as late entrance. The other type of truncation targets at the variable L, that is, L is right truncated by T.

To illustrate left truncation, a survival study of residents in a retirement center is considered. Age at death T and the age of entrance L are recorded. An individual must survive to a certain age, for example, 65 years, to enter the retirement center. If an individual died early and was not old enough to enter the center, he/she has no chance to be included in the study. Therefore, the ages of death, using the survival data collected at the retirement center, are left-truncated by the ages at entrance. A right truncation example is the HIV virus latent time in AIDS studies. It is mandatory to report AIDS cases to CDC. If patients were infected with HIV virus from blood transfusion, researchers can track back the calendar date of blood transfusion. Suppose that the closing date of a study was set at Dec. 31, 2010, then infected individuals who developed AIDS after this date were not included in the sample. In another word, the subjects can be possibly included only if the HIV virus latent time less than the time between blood transfusion date and study closing date. Thus, the latent time is right truncated by the time between blood transfusion date and the closing date of study.

The major study interest with a truncated sample is to find the marginal distributions of T and L (Woodroofe, 1985). So far most statistical methods for truncated data make the key assumption of quasi-independence between T and L. Truncated version of Kaplan-Meier estimators have been widely used for estimating the distribution function and their asymptotic properties were studied by Woodroofe (1985), Keiding and Gill (1990) and Wang et al. (1986).

However, the quasi-independence assumption is questionable in many instances. Tsai (1990) proposed a conditional Kendalls Tau test to test the quasi-independence for a truncated sample. He explained that independence between T and L cannot be nonparametrically verified in the quadrant $T < L$. Keiding (1992) also showed that when failure time T and truncation time L were dependent, the standard delayed entry method, based on the quasi-independence assumption, yielded a biased estimate of the failure hazard.

Estimation of the marginal distribution of T and L is rarely studied for a dependently truncated sample. To relax the independence assumption, Emma and Konno (2010) presented a bivariate normal distribution method of fitting a parametric model on (L, T), which could easily incorporate the dependence structure on the truncation mechanisms. They used the maximum likelihood estimation method to find the estimates of the parameters. However, it is difficult to extend this method to the context when multiple predictors are associated with the failure time. Chaieb et al. (2006) developed a nonparametric estimator of the time-to-event distribution in the presence of dependent truncation using a copula, while implementation of this method requires the user to specify a copula from an Archimedian family. There is a growing interest in finding more simple estimation methods when T and L are dependent.

For complete time to event data, it is well known that the empirical estimator can be used to estimate the distribution function of the failure time. In an empirical estimator, all observations contribute equally for estimating the distribution function. The observation should be weighted when we want to have the empirical estimator form for estimating the distribution function of a truncated variable.

Since truncated data represent a nonrandomly screened subset of a population, analytical methods must account for the biased selection nature of the sample. A commonly used method to correct biased selection for truncated data is inverse-probability-weighting (IPW) technique (Wang, 1989; Shen, 2003,2006). The concept of IPW is first proposed by Horvitz and Thompson (1952). The principle is to weight an observation by the reciprocal of its selection probability. Satten and Datta (2001)showed that the Kaplan-Meier estimator can be expressed as an IPW estimator for randomly censored data. Shen (2003) presented the IPW estimator for independently truncated samples and proved that the IPW estimator is evaluated the same as Kaplan-Meier estimator. Wang (1989) studied the IPW estimator when the parametric distribution of the truncation variable was known. Using the parametric information of the truncation variable, she proved the asymptotic efficiency of the IPW estimator compared to its analogue of Kaplan-Meier estimator. IPW estimator was more powerful for bias correction and efficiency improvement. The weighted average form is convenient for various statistical problems such as causal inference and missing data (Robins and Finkelstein, 2000).

To identify and quantify the effect of prognostic factors which is related to the course of a disease is the primary goal of survival analysis. To predict the outcome of a patient based on a series factors, several regression models, including Cox model (1972), accelerated failure time model (Kalbfleisch, 1980) and additive hazards models (Aalen, 1989; Lin and Yin, 1994; Mckeague and Sasieni, 1994) have been proposed. The Cox proportional hazard model is probably the most commonly used method when analyzing the impact of covariates on continuous survival time. The main advantage of the Cox model is the possibility to estimate the regression parameters without any assumption on the distribution of the duration variable. That is, there are no parametric restrictions on the functional form of the baseline hazard function. In a Cox model, the specification for the hazard function is given as:

$$\lambda(t|z) = \lambda_0(t) \exp(\beta^T z), \tag{1.1}$$

where $\lambda_0(t)$ is the unspecified baseline hazard function, β is the vector of regression coefficients and z is the vector of covariates. The estimation of the regression parameters can be carried out using the partial likelihood function (Cox, 1975). In its classical form, the Cox model was introduced in the setting of right censored observation. However, in practice the structure of data might be more complex and different sampling schemes are frequently encountered. These motivate the new extensions of Cox model allowing for truncation and recurrent event. For example, to compare chemotherapy and Bone Marrow Transplant (BMT) on treating leukemia patients, Klein and Zhang (1996) advocated the left-truncated version Cox model to analyze the pooled samples of chemotherapy and BMT with satisfied results. However, the effect of transplant is assumed to be constant regardless of the transplant time. They assumed the independence between failure time T and transplant time L in their model. In recent year, there is considerable interest in the differential effect of transplant time on the future survival of leukemia patient. It is hence necessary to modify the above Cox model to reflect the effects of various transplant times, so that such effect can be tested and evaluated.

The study cohorts obtained from BMT registry data are commonly recognized as truncated samples, the truncation time is the transplant time. Some clinical results show strong evidence that the longer waiting time in BMT regimen is associated with a worse prognosis (Balduzzi et al., 2008; Davies and Mehta, 2010). Thus, it is reasonable to assume the dependence between L and T in BMT study. The current analytical methods on the pooled samples include the matched pairs analysis and the Cox analysis assuming a constant effect for transplant. However, the effect of the transplant waiting time cannot be evaluated using these analytical approaches (Galimberti et al., 2002). In this thesis, we consider a Cox analysis and the transplant time is both a truncation variable and a predictor of the time-to-failure. We propose an IPW estimator to estimate the marginal distribution of L. Simulation studies have been conducted to investigate the performances of the proposed IPW estimator and variance estimators.

The structure of this thesis is organized as follows. In chapter 2, we first briefly describe the Kaplan-Meier estimators and IPW estimators for the distribution functions in a truncated sample. Second, we present the truncated version Cox model for analyzing the effect of covariates on continuous survival time. In Chapter 3, we introduce the new methodology for dependently truncated samples. In Chapter 4, the simulation study is performed to show the performances. In Chapter 5, a real BMT registry data is analyzed to illustrate the proposed method. Finally, the concluding remarks are given in Chapter 6.

Chapter 2

METHODOLOGY REVIEW

2.1 Kaplan-Meier estimator for truncated data

The Kaplan-Meier estimator is commonly used to find the crude survival curve for right censored time-to-event data. It is known that Kaplan-Meier estimator is NPMLE. For a truncated sample, suppose $F(t)$ and $G(t)$ are the distribution functions for the failure time T and the truncation time L, respectively. If one assume independence between T and L, the truncated version of Kaplan-Meier estimators can be used for the distribution functions of T and L. For the sample $(L_i, T_i), i = 1, \cdots, n$, define $\bar{Y}(t)$ as the number of individuals who entered the study prior to time t and remained under study at t, then $\bar{Y}(t) = \sum_{i=1}^{n} I(L_i \leq t \leq T_i)$, also let $t_{(1)} < \cdots < t_{(N)}$ be the ordered distinct event times. Kaplan-Meier (1958) and Lynden-Bell (1971) proposed a nonparametric estimator for $F(t)$:

$$\hat{F}(t) = 1 - \prod_{i:t_{(i)} \leq t} \left[1 - \frac{d(t_{(i)})}{\bar{Y}(t_{(i)})} \right], 0 \leq t \leq \tau, \tag{2.1}$$

where $d(t)$ is the number of failures at t. Let $l_{(1)} < l_{(2)} < \cdots < l_{(M)}$ be distinct truncation times. The estimator of $G(t)$ is given by,

$$\hat{G}(t) = \prod_{k:l_{(k)}>t} \left[1 - \frac{s(l_{(k)})}{\bar{Y}(l_{(k)})}\right], 0 \leq t \leq \tau, \tag{2.2}$$

where $s(l) = \sum_{i=1}^{n} I(L_i = l)$.

2.2 The inverse probability weighted (IPW) estimator

Besides the truncated version Kaplan-Meier estimator, various estimation methods have been proposed to handle truncated data. Among them, inverse probability weighting technique is one powerful tool for bias correction and efficiency improvement (Wang, 1989). The concept of IPW is first proposed by Horvitz and Thompson (Horvitz, 1952). The key idea is both straightforward and intuitively attractive as shown in the following example. Suppose that we have the following data

Group	A			B			C		
Response	1	1	1	2	2	2	3	3	3

Then the average response is 2. However, if selection probability varies between groups and the observed values are shown in the following table:

Group	A			B			C		
Response	1	?	?	2	2	2	3	3	?

The intuitive average is $\frac{13}{6} = 2.16$, which departs from the true mean. IPW technique can be used to eliminate this kind of selection bias. The probability of being selected in the sample is $\frac{1}{3}$ in group A, 1 in group B and $\frac{2}{3}$ in group C, we can calculate a weighted average where each observation is weighted by the reciprocal of its probability of selection:

$$\frac{1\times\frac{3}{1}+(2+2+2)\times 1+(3+3)\times\frac{3}{2}}{\frac{3}{1}+1+1+1+\frac{3}{2}+\frac{3}{2}}=2$$

Thus, the selection bias has been corrected in this example using IPW method. Some recent researches have proven the equivalence between Kaplan-Meiler estimator and the inverse probability weighted average. For example, Satten (Satten and Datta, 2001) showed that the Kaplan-Meier estimator (product-limit estimator, PLE) can be expressed as an IPW average for randomly censored data. Shen (Shen, 2003) showed the truncation PLE and the censoring-truncation PLE can also be expressed as IPW averages. In this section, the complex mathematical proof is out of our interest and ignored; however, the useful conclusions about IPW estimators for truncated data will be focused and shown as in the following discussion. For a truncated sample $(L_i, T_i), i = 1, \cdots, n$, let $\hat{F}(t)$ and $\hat{G}(t)$ be the truncated version Kaplan-Meier estimators (see Chapter 2.1). since $G(T_i)$ is the selection probability of T_i, the IPW estimator of $F(t)$ is given by

$$\hat{F}^{\mathrm{IPW}}(t)=\left(\sum_{i=1}^{n}\frac{1}{\hat{G}(T_i)}\right)^{-1}\sum_{i=1}^{n}\frac{I(T_i\le t)}{\hat{G}(T_i)}. \tag{2.3}$$

By recognizing $F(L_i)$ to be the selection probability of L_i, we can give the IPW estimator of $G(t)$,

$$\hat{G}^{\mathrm{IPW}}(t)=\left(\sum_{i=1}^{n}\frac{1}{1-\hat{F}(L_i-)}\right)^{-1}\sum_{i=1}^{n}\frac{I(L_i\le t)}{1-\hat{F}(L_i-)}. \tag{2.4}$$

The above IPW estimators are essentially the same as the truncated version Kaplan-Meier estimators (Shen, 2003). One useful application of the IPW estimator was studied by Wang (1989) for the context that the parametric distribution of the truncation variable was already known. $G(T_i;\theta)$, the given parametric distribution probability at T_i, was used as the selection probability of T_i. Wang (1989) proposed the following IPW estimator

$$\hat{F}^{\mathrm{IPW}}(t;\hat{\theta}) = \left(\sum_{i=1}^{n} \frac{1}{\hat{G}(T_i;\hat{\theta})}\right)^{-1} \sum_{i=1}^{n} \frac{I(T_i \leq t)}{\hat{G}(T_i;\hat{\theta})}, \tag{2.5}$$

where $\hat{\theta}$ and $\hat{G}$ are MLE's. Wang illustrated in a simulation study that the IPW estimator is more efficient than the truncated version Kaplan-Meier estimator.

2.3 The Cox model for truncated data

Regression analysis on the failure time T based on a truncated sample has been identified to be practically important (Karlsson and Laitila, 2008; Shen, 2010). The solution is simple if T and L are independent. For the hazard based regression models, the only modification one needs to implement in estimation procedure is to use truncation time to adjust the risk set. Let $t_{(1)} < t_{(2)} < \cdots < t_{(N)}$ denote the ordered event times. $\bar{Y}(t)$ is the risk set contains the subjects who enter the study before t and are still alive at t. Let $d_{(i)}$ be the total number of failures at $t_{(i)}$, $D_{(i)}$ be the set of all subjects who fail at time $t_{(i)}$. Let $s_{(i)}$ be the sum of the covariate values over all subjects in the set $d_{(i)}$, that is $s_{(i)} = \sum_{i \in D_{(i)}} z_i$. The MLE of β, maximizes the partial likelihood of β (Breslow, 1974) is given as

$$L(\beta) = \prod_{i=1}^{N} \frac{exp(\beta^T s_{(i)})}{[\sum_{i \in \bar{Y}(t_{(i)})} exp(\beta^T z_i)]^{d_{(i)}}}. \tag{2.6}$$

Let $\Lambda_0(t)$ be the cumulative baseline hazard function, $\Lambda_0(t) = \int_0^t \lambda_0(u)du$. Breslow estimator is routinely used for estimating $\Lambda_0(t)$, and it has the form,

$$\hat{\Lambda}_0(t) = \sum_{i:t_{(i)}\leq t} \frac{d_{(i)}}{\sum_{i\in\bar{Y}(t_{(i)})} exp(\beta^T z_i)}. \tag{2.7}$$

Given covariate z, the survival probability at t is

$$S(t; z) = exp(-\Lambda_0(t)e^{\beta^T z}). \tag{2.8}$$

One can use a plug-on estimator shown below:

$$\hat{S}(t; z) = exp(-\hat{\Lambda}_0(t)e^{\hat{\beta}^T z}). \tag{2.9}$$

The product limit estimator is an alternative. The explicit expression can be found in Klein and Zhang (2003). Estimation of covariate effects in a Cox model with a truncated sample has been implemented in the statistical software such as SAS and S-plus. The SAS procedure PHREG can be used to give us the estimation result. One simple example is illustrated below to show how to implement left truncated version of Cox model using the PHREG procedure. Suppose that a truncated sample has been saved as a SAS data set “sample”. In the SAS data set, the truncation time and the failure time are saved in the variables “Ltime”and “Xtime” respectively. The variable “event” takes the value 1 if the failure time is observed (patient is died), and takes the value 0 if the follow-up time is observed (patient is censored). Two factors, age and gender, are considered. The data set “sample” includes the continuous variable “age”and the binary variable “male” (1 if the gender is male, 0 otherwise). We can use the following statements to have the covariate effects estimated:

```
Proc PHREG data=sample;
Model (Ltime, Xtime)*event(0)= age male;
Run;
```

The left truncated version Cox model requires the condition of quasi-independence between the failure time variable and the truncation variable. The validity of the model has not been studied for a dependently truncated sample.

Chapter 3

NEW METHODOLOGY

3.1 The Cox model with truncation variable included as a covariate

The major study interest with a truncated sample is to find the marginal distributions of L and T. Many researches have been done based on the key assumption of quasi-independence between L and T. The real applications may yield dependently truncated samples. For example, in BMT studies, there is evidence for the association between the failure and the transplant waiting time. The longer waiting time in BMT regimen will be associated with a worse prognosis. It is reasonable to assume the dependence between L and T. Li (2010) employed a Cox analysis for the dependently truncated sample. The key idea is to use the transplant time as a predictor for the occurrence of the failure time. Inclusion of the transplant time explains the association between the transplant time L and the failure time T. For more general use, the regressor should also include other covariates z. The hazard of the failure time can be specified as follows

$$\lambda(t|L,Z) = \begin{cases} \lambda_0(t)exp(\alpha^T z) & \text{if } t < L \\ \lambda_0(t)exp(\gamma\kappa(L) + \alpha^T z) & \text{if } t > L \end{cases}, \tag{3.1}$$

where $\lambda_0(t)$ is the unspecified baseline hazard, γ and α are the regression coefficients, $\kappa(.)$ is a known function.

Estimation of regression coefficients in a Cox model, with the presence of right censoring and left truncation, has been well established. Suppose that the truncated sample is summarized as $\{L_i, X_i, \Delta_i, Z_i\}$, $i = 1, \cdots, n$, where $X_i = \min(T_i, C_i)$, $\Delta_i = I(T_i \leq C_i)$, Z_i and C_i are the covariate vector and the censoring time for the ith subject. Define the following processes, $N_i^T(t) = I(X_i \leq t, \Delta_i = 1)$, $\bar{N}^T(t) = \sum_{i=1}^n N_i^T(t)$ and $Y_i(t) = I(L_i \leq t \leq X_i)$. Since the transplant time is treated as predictor of failure event, we define the covariate vector, $\tilde{Z}_i^T = \left\{\kappa(L_i) \quad Z_i^T\right\}$. We also combine the regression parameters, γ and α, into one vector, $\beta^T = \left\{\gamma \quad \alpha^T\right\}$. Define

$$S^{(p)}(\beta, t) = n^{-1} \sum_{i=1}^n Y_i(t) \tilde{Z}_i^{\otimes p} \exp(\beta^T \tilde{Z}_i), \quad p = 0, 1, 2,$$

where $a^{\otimes 2} = aa^T$. The partial likelihoods (Cox, 1972) can be constructed for Model (3.1), yielding the following score estimation equation,

$$\mathcal{U}(\beta) = \sum_{i=1}^n \int_0^\infty \left(\tilde{Z}_i - \frac{\sum_{j=1}^n Y_j(t) \tilde{Z}_j \exp(\beta^T \tilde{Z}_j)}{\sum_{j=1}^n Y_j(t) \exp(\beta^T \tilde{Z}_j)} \right) dN_i^T(t).$$

Let $\hat{\beta}$ be the solution to $\mathcal{U}(\beta) = 0$, and it is hence the MLE. More explicitly, $\hat{\beta}^T = \left\{\hat{\gamma} \quad \hat{\alpha}^T\right\}$. The variance-covariance matrix of $\hat{\beta}$ can be estimated by the inverse of the estimated information matrix,

$$\hat{\mathcal{I}}(\hat{\beta}) = \sum_{i=1}^n \int_0^\infty \left[\frac{S^{(2)}(\hat{\beta}, t)}{S^{(0)}(\hat{\beta}, t)} - \left(\frac{S^{(1)}(\hat{\beta}, t)}{S^{(0)}(\hat{\beta}, t)} \right)^{\otimes 2} \right] dN_i^T(t)$$

One can use the Breslow-type estimator to estimate the cumulative baseline hazard function,

$$\hat{\Lambda}_0(t) = \sum_{i=1}^{n} \int_0^t \frac{dN_i^T(u)}{\sum_{j=1}^{n} Y_j(u) \exp(\hat{\beta}^T \tilde{Z}_i)}. \tag{3.2}$$

This type of Cox analysis has been implemented in a few statistical packages such as SAS and R. We use the example described in Chapter 2.3 to present the SAS syntax, assuming that the truncation variable should enters the regressor of Cox model. Suppose that the logarithm of "Ltime" is the proper form to be added into the regression. We can create a variable "logtime", which is the logarithm of the truncation time variable. We shall employ the following syntax to specify the model:

```
Proc PHREG data=sample;
model (Ltime, Xtime)*event(0)= logtime age male;
run;
```

3.2 A new IPW estimator

For BMT registry data, the truncation variable is the transplant time, which is dominantly determined by the donor searching process. For treating leukemia patients, the information about the amount of time normally spent on the donor searching is crucial for policy makers to efficiently allocate resources to assist patients in finding donors. It is challenging to estimate the distribution function of L, given Model (3.1). The truncated version Kaplan-Meier estimator is not applicable due to the dependence nature between transplant time and failure time. The bias of such an estimator is demonstrated in the simulation results included in Chapter 4. Li (2010) proposed an algorithm to estimate the distribution of L. However, Li only used the resampling approach to assess the precision of the estimates. In this section, we introduce an IPW estimator for the distribution of L, and give the analytical formula for variance estimation.

To find the form of the IPW estimator, it is important to identify the selection probability for individual observation L_i. In BMT operations, the donor searching can be viewed as a random process, independent from the failure event. The waiting time alters the future survivorship when the transplant is operated. We introduce a latent random variable $T_{0,z}$. Its underlying counting process is associated with the intensity process $Y_i(t)\lambda_0(t)e^{\alpha^T z}$. Since we assume that donor searching is a random process, the variables L and $T_{0,z}$ are independent. The selection probability of L_i is recognized as $P(T > L_i | L_i = 0, Z_i)$. For presentation simplicity, we assume no dies observed among the truncation times and observed times. Let $\{\tilde{Z}_i^0\}^T = \{0 \ Z_i^T\}$ and $S(L_i; \tilde{Z}_i^0) = P(T > L_i | L_i = 0, Z_i)$. The reciprocal of $S(L_i; \tilde{Z}_i^0)$ is the weight pertaining to the observation L_i. $S(L_i; \tilde{Z}_i^0)$ can be estimated by

$$\hat{S}(L_i; \tilde{Z}_i^0) = \exp\left(-\hat{\Lambda}_0(t)e^{\hat{\alpha}^T Z_i}\right).$$

To estimate the distribution function of L, we suggest the following IPW estimator,

$$\hat{G}(t; \hat{\beta}) = \hat{P}(\hat{\beta})n^{-1}\sum_{i=1}^{n}\frac{I(L_i \leq t)}{\hat{S}(L_i; \tilde{Z}_i^0)},$$

where $P(\beta) = P(L \leq T|\text{data})$ and

$$\hat{P}(\hat{\beta}) = \left(n^{-1}\sum_{i=1}^{n}\frac{1}{\hat{S}(L_i; \tilde{Z}_i^0)}\right)^{-1}.$$

The asymptotic distribution of $\sqrt{n}(\hat{G}(t; \hat{\beta}) - G(t))$ is given as follows.

Assume that:

(1) The regularity conditions needed for asymptotic properties of the estimators in Cox analysis. There exists $s^{(0)}, s^{(1)}, s^{(2)}$ such that

$$\sup_{\beta, t \in [0,\tau]} ||S^{(p)}(\beta, t) - s^{(p)}(\beta, t)|| \to_{\mathcal{P}} 0.$$

$s^{(0)}(\beta, t)$ is bounded away from zero. Define $e = s^{(1)}/s^{(0)}$, $v = s^{(2)}/s^{(0)} - e^{\otimes 2}$ and

$$\Sigma = \int_0^\infty v(\beta, t) s^{(0)}(\beta, t) \lambda_0(t) dt.$$

(2) Let $\tilde{z}^0(u)$ be the covariate value of the subject with transplant time u. Then $S(u; \tilde{z}^0(u))$ is the selection probability for the transplant time u. Suppose that $S(u; \tilde{z}^0(u))$ and $G(u)$ are the continuous functions defined on $[0, \infty)$, and the following condition is satisfied,

$$\int_0^\infty \frac{1}{S(t; \tilde{z}^0(t))} dG(t) < \infty.$$

Given t, the IPW estimator $\sqrt{n}(\hat{G}(t; \hat{\beta}) - G(t))$ converges in distribution to a normal random variable with mean zero and variance

$$\begin{aligned}
\sigma^2(t) \;=\;& P(\beta) \int_0^t S(u; \tilde{z}^0(u))^{-1} dG(u) + P(\beta) G(t)^2 \int_0^\infty S_0(u; \tilde{z}^0(u))^{-1} dG(u) \\
& -2P(\beta) G(t) \int_0^t S(u; \tilde{z}^0(u))^{-1} dG(u) \\
& + \int_0^\infty \{\eta(u, t) - P(\beta) G(t) \psi(u)\}^2 \frac{d\Lambda_0(u)}{s^{(0)}(\beta, u)} \\
& + \{\rho(t) - P(\beta) G(t) \times \pi\}^T \Sigma^{-1} \{\rho(t) - P(\beta) G(t) \times \pi\},
\end{aligned}$$

where

$$\eta(u,t) = \lim_{n\to\infty} n^{-1} P(\beta) \sum_{i=1}^{n} S(L_i; \tilde{Z}_i^0)^{-1} I(u \leq L_i \leq t) e^{\alpha^T Z_i},$$

$$\psi(u) = \lim_{n\to\infty} n^{-1} P(\beta) \sum_{i=1}^{n} S(L_i; \tilde{Z}_i^0)^{-1} I(u \leq L_i) e^{\alpha^T Z_i},$$

$$\rho(t) = \lim_{n\to\infty} n^{-1} P(\beta) \sum_{i=1}^{n} S(L_i; \tilde{Z}_i^0)^{-1} I(L_i \leq t) h(L_i; Z_i),$$

$$\pi = \lim_{n\to\infty} n^{-1} P(\beta) \sum_{i=1}^{n} S(L_i; \tilde{Z}_i^0)^{-1} h(L_i; Z_i).$$

$$h(t; z) = \int_0^t e^{\alpha^T z} \left(\left\{ \begin{array}{c} 0 \\ z \end{array} \right\} - e(\beta, u) \right) \lambda_0(u) du.$$

Here we present a brief description of our derivation. The variation of our IPW estimator can be explained by two sources: the variation of an IPW estimator using known weight functions, and the variation due to estimated weight. We define an interim term

$$\hat{G}(t; \beta) = \hat{P}(\beta) n^{-1} \sum_{i=1}^{n} \frac{I(L_i \leq t)}{S(L_i; \tilde{Z}_i^0)},$$

$$\hat{P}(\beta) = \left(n^{-1} \sum_{i=1}^{n} \frac{1}{S(L_i; \tilde{Z}_i^0)} \right)^{-1}.$$

Essentially, $\hat{G}(t; \beta)$ is an IPW estimator using known weight functions. Then,

$$\sqrt{n}\left(\hat{G}(t; \hat{\beta}) - G(t)\right) = \sqrt{n}\left(\hat{G}(t; \hat{\beta}) - \hat{G}(t; \beta)\right) + \sqrt{n}\left(\hat{G}(t; \beta) - G(t)\right)$$

First, we consider weak convergence of $\sqrt{n}\left(\hat{G}(t; \beta) - G(t)\right)$. Note that $\hat{G}(t; \beta)$ is an IPW estimator with known weight functions. Vardi (1985) studied the problem of estimating a distribution function when sampling weights are known. The proposed weighted estimator based on known weight functions was proved to be MLE, and the weak convergence result was sketched in his paper. Wang (1989) studied an IPW estimator for an independently truncated sample, when the parametric distribution of the other variable is known. She explicitly split the variation of the IPW estimator into two sources, Varid's result was used for the variation for the estimator with known weight functions. There is a high level of similarity between our IPW estimator and Wang's IPW estimator. According to Vardi (1985, Section 8) and Wang (1989, Lemma 3.3), we have the following convergence result. $\sqrt{n}\left(\hat{G}(t;\beta) - G(t)\right)$ converges in distribution to a normal variate with mean zero and variance

$$\begin{aligned}\sigma_1^2(t) &= P(\beta)\int_0^t S(u;\tilde{z}^0(u))^{-1}dG(u) + P(\beta)G(t)^2\int_0^\infty S(u;\tilde{z}^0(u))^{-1}dG(u) \\ &\quad -2P(\beta)G(t)\int_0^t S(u;\tilde{z}^0(u))^{-1}dG(u)\end{aligned}$$

Some notations should be defined for studying weak convergence of $\sqrt{n}\left(\hat{G}(t,\hat{\beta}) - \hat{G}(t;\beta)\right)$. Define

$$E(\beta,t) = \frac{S^{(1)}(\beta,t)}{S^{(0)}(\beta,t)},$$

$$V(\beta,t) = \frac{S^{(2)}(\beta,t)}{S^{(0)}(\beta,t)} - E(\beta,t)^{\otimes 2},$$

$$M_i(t) = N_i^T(t) - \int_0^t Y_i(u)\lambda_0(u)\exp(\beta^T\tilde{Z}_i)du, \quad i = 1,\cdots,n,$$

In the following context, $\approx$ means asymptotic equivalence. We have

$$\begin{aligned}
\sqrt{n}\left[\hat{G}(t,\hat{\beta})-\hat{G}(t;\beta)\right] &= \sqrt{n}\left[\hat{P}(\hat{\beta})\sum_{i=1}^{n}\frac{I(L_i\le t)}{\hat{S}(L_i;\tilde{Z}_i^0)}-\hat{P}(\beta)\sum_{i=1}^{n}\frac{I(L_i\le t)}{S(L_i;\tilde{Z}_i^0)}\right]\\
= & \sqrt{n}\hat{P}(\hat{\beta})\left[\sum_{i=1}^{n}\frac{I(L_i\le t)}{\hat{S}(L_i;\tilde{Z}_i^0)}-\sum_{i=1}^{n}\frac{I(L_i\le t)}{S(L_i;\tilde{Z}_i^0)}\right]+\sqrt{n}\left[\hat{P}(\hat{\beta})-\hat{P}(\beta)\right]\sum_{i=1}^{n}\frac{I(L_i\le t)}{S(L_i;\tilde{Z}_i^0)}\\
\approx & \sqrt{n}\hat{P}(\beta)\left[\sum_{i=1}^{n}\frac{I(L_i\le t)}{\hat{S}(L_i;\tilde{Z}_i^0)}-\sum_{i=1}^{n}\frac{I(L_i\le t)}{S(L_i;\tilde{Z}_i^0)}\right]\\
& -\hat{P}(\beta)^2\sqrt{n}\left[\sum_{i=1}^{n}\frac{1}{\hat{S}(L_i;\tilde{Z}_i^0)}-\sum_{i=1}^{n}\frac{1}{S(L_i;\tilde{Z}_i^0)}\right]\sum_{i=1}^{n}\frac{I(L_i\le t)}{S(L_i;\tilde{Z}_i^0)}\\
\approx & P(\beta)\frac{1}{n}\sum_{i=1}^{n}S(L_i;\tilde{Z}_i^0)^{-1}I(L_i\le t)\sqrt{n}\left[\hat{\Lambda}(L_i;\tilde{Z}_i^0)-\Lambda(L_i;\tilde{Z}_i^0)\right]\\
& -P(\beta)G(t)\frac{1}{n}\sum_{i=1}^{n}S(L_i;\tilde{Z}_i^0)^{-1}\sqrt{n}\left[\hat{\Lambda}(L_i;\tilde{Z}_i^0)-\Lambda(L_i;\tilde{Z}_i^0)\right]
\end{aligned}$$

Using the standard result of a Cox model (Andersen and Gill, 1982),

$$\begin{aligned}
\sqrt{n}\left[\hat{\Lambda}(L_i;\tilde{Z}_i^0)-\Lambda(L_i;\tilde{Z}_i^0)\right] \approx & \frac{1}{\sqrt{n}}\left[\sum_{j=1}^{n}\int_0^{L_i}e^{\alpha^T Z_i}\frac{dM_j(u)}{s^{(0)}(\beta,u)}\right.\\
& \left.+\,h(L_i;Z_i)\Sigma^{-1}\sum_{j=1}^{n}\int_0^{\infty}\left(Z_j-\frac{s^{(1)}(\beta,u)}{s^{(0)}(\beta,u)}\right)dM_j(u)\right].
\end{aligned}$$

The above equation can be further expressed as

$$\begin{aligned}
\sqrt{n}\left(\hat{G}(t,\hat{\beta})-\hat{G}(t;\beta)\right) \approx & \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\int_0^{\infty}\left\{\eta(u,t)-P(\beta)G(t)\psi(u)\right\}\frac{dM_j(u)}{s^{(0)}(\beta,u)}\\
& +\frac{1}{\sqrt{n}}\left\{\rho(t)-P(\beta)G(t)\times\pi\right\}^T\Sigma^{-1}\sum_{j=1}^{n}\int_0^{\infty}\left(Z_j-\frac{s^{(1)}(\beta,u)}{s^{(0)}(\beta,u)}\right)dM_j(u).
\end{aligned}$$

The standard result for the variation process of martingale can help us to find the variance. Using the martingale central limit theorem, $\sqrt{n}\left\{\hat{G}(t)-\hat{G}(t;\beta)\right\}$ converges in distribution to a zero-mean normal variate with variance

$$\begin{aligned}\sigma_2^2(t) &= \sum_{j=1}^{n}\int_0^\infty \{\eta(u,t)-P(\beta)G(t)\psi(u)\}^2\,\frac{\lambda_0(u)du}{s^{(0)}(\beta,u)} \\ &\quad +\{\rho(t)-P(\beta)G(t)\times\pi\}^T\,\Sigma^{-1}\,\{\rho(t)-P(\beta)G(t)\times\pi\}\,.\end{aligned}$$

Based on the arguments used in Wang's derivation, we have the independence between $\sqrt{n}\left(\hat{G}(t;\hat{\beta})-\hat{G}(t;\beta)\right)$ and $\sqrt{n}\left(\hat{G}(t;\beta)-G(t)\right)$. Therefore, $\sqrt{n}\left(\hat{G}(t;\hat{\beta})-G(t;\beta)\right)$ converges in distribution to a zero-mean normal random variable, with the variance $\sigma^2(t)=\sigma_1^2(t)+\sigma_2^2(t)$. The plug-in estimator can be used for the asymptotic variance of $\sqrt{n}(\hat{G}(t,\hat{\beta})-G(t))$. The explicit express has the form

$$\begin{aligned}\hat{\sigma}^2(t) &= \hat{P}(\hat{\beta})\int_0^t \hat{S}(u;\tilde{z}^0(u))^{-1}d\hat{G}(u)+\hat{P}(\hat{\beta})\hat{G}(t)^2\int_0^\infty \hat{S}(u;\tilde{z}^0(u))^{-1}d\hat{G}(u) \\ &\quad -2\hat{P}(\hat{\beta})\hat{G}(t)\int_0^t \hat{S}(u;\tilde{z}^0(u))^{-1}d\hat{G}(u) \\ &\quad +n^{-1}\int_0^\infty \left\{\hat{\eta}(u,t)-\hat{P}(\hat{\beta})\hat{G}(t)\hat{\psi}(u)\right\}^2\frac{d\bar{N}^T(u)}{\left[S^{(0)}(\hat{\beta},u)\right]^2} \\ &\quad +\left\{\hat{\rho}(t)-\hat{P}(\hat{\beta})\hat{G}(t)\times\hat{\pi}\right\}^T\hat{\Sigma}^{-1}\left\{\hat{\rho}(t)-\hat{P}(\hat{\beta})\hat{G}(t)\times\hat{\pi}\right\},\end{aligned}$$

where

$$\hat{\eta}(u,t)=n^{-1}\hat{P}(\hat{\beta})\sum_{i=1}^{n}\hat{S}(L_i;\tilde{Z}_i^0)^{-1}I(u\le L_i\le t)e^{\hat{\alpha}^T Z_i},$$

$$\hat{\psi}(u)=n^{-1}\hat{P}(\hat{\beta})\sum_{i=1}^{n}\hat{S}(L_i;\tilde{Z}_i^0)^{-1}I(u\le L_i)e^{\hat{\alpha}^T Z_i},$$

$$\hat{\rho}(t)=n^{-1}\hat{P}(\hat{\beta})\sum_{i=1}^{n}\hat{S}(L_i;\tilde{Z}_i^0)^{-1}I(L_i\le t)\hat{h}(L_i;Z_i),$$

$$\hat{\pi}=n^{-1}\hat{P}(\hat{\beta})\sum_{i=1}^{n}\hat{S}(L_i;\tilde{Z}_i^0))^{-1}\hat{h}(L_i;Z_i),$$

$$\hat{h}(t;z) = \int_0^t e^{\hat{\alpha}^T z} \left(\left\{ \begin{array}{c} 0 \\ \\ z \end{array} \right\} - E(\hat{\beta}, u) \right) d\hat{\Lambda}_0(u) \qquad \text{and} \qquad \hat{\Sigma} = n^{-1} \hat{\mathcal{I}}(\hat{\beta}).$$

Chapter 4

THE SIMULATION STUDY

Our goal is to evaluate the practical performance of the proposed IPW estimator and the variance estimator. The Kaplan-Meier estimator used for independently truncated sample is also reported for comparison. The simulation study in this section emphasizes on the scenario that L is one predictor of T and a fixed covariate z is also associated with T. We assume that regressor of Cox model contains a linear form of the truncation time.The underlying hazard rate function of T is given by

$$\lambda(t|L,z) = \begin{cases} \lambda_0(t)exp(\alpha^T z) & \text{if } t < L \\ \lambda_0(t)exp(\gamma L + \alpha^T z) & \text{if } t > L \end{cases} . \tag{4.1}$$

The truncation variable L was simulated from a Uniform distribution at the interval [0,80]. The baseline hazard rate in the above model has been set to a constant and we use different constants as the baseline hazard rate to control the censoring and truncation rates. Continuous covariate is generated from a standard normal distribution, restraining in the internal [-3, 3]. We use true value: $\alpha = 0.5, 1$. Discrete covariate is generated from a Bernoulli distribution with parameter value 0.5. We use true value: $\alpha = 1$. Settings with positive β value ($\beta = 0.02$) and negative β value ($\beta = -0.05$) were both generated. Positive β and negative β represent the escalated risk of failure rate or preventive effect of the truncation time variable, respectively.

We considered two levels for the truncation rate (25%, 50%) and two levels for the censoring rate (25%, 50%). Censoring time was generated from Uniform $[a,\ b]$. We adjusted the values of $a,\ b$ to control the censoring rate. For each setting, we generate 1000 samples with size 200. The bias is defined as the deviation between the average cumulative hazard estimate and the true value. For estimation on each parameter, we calculate bias, sample variance, estimated variance and 95% confidence interval coverage at different time points when $G(t)$ is evaluated to be 0.25, 0.5, 0.75. The following formulas are used to calculate the relative terms:

$$Bias = \bar{G}^{\mathrm{IPW}}(t) - G(t)$$

$$\bar{G}^{\mathrm{IPW}}(t) = \frac{1}{1000}\sum_{i=1}^{1000}\hat{G}_{(i)}(t;\hat{\beta})$$

$$var(\hat{G}(t;\hat{\beta})) = \frac{1}{1000-1}\sum_{i=1}^{n}(\hat{G}_{(i)}(t;\hat{\beta}) - \bar{G}^{\mathrm{IPW}}(t)$$

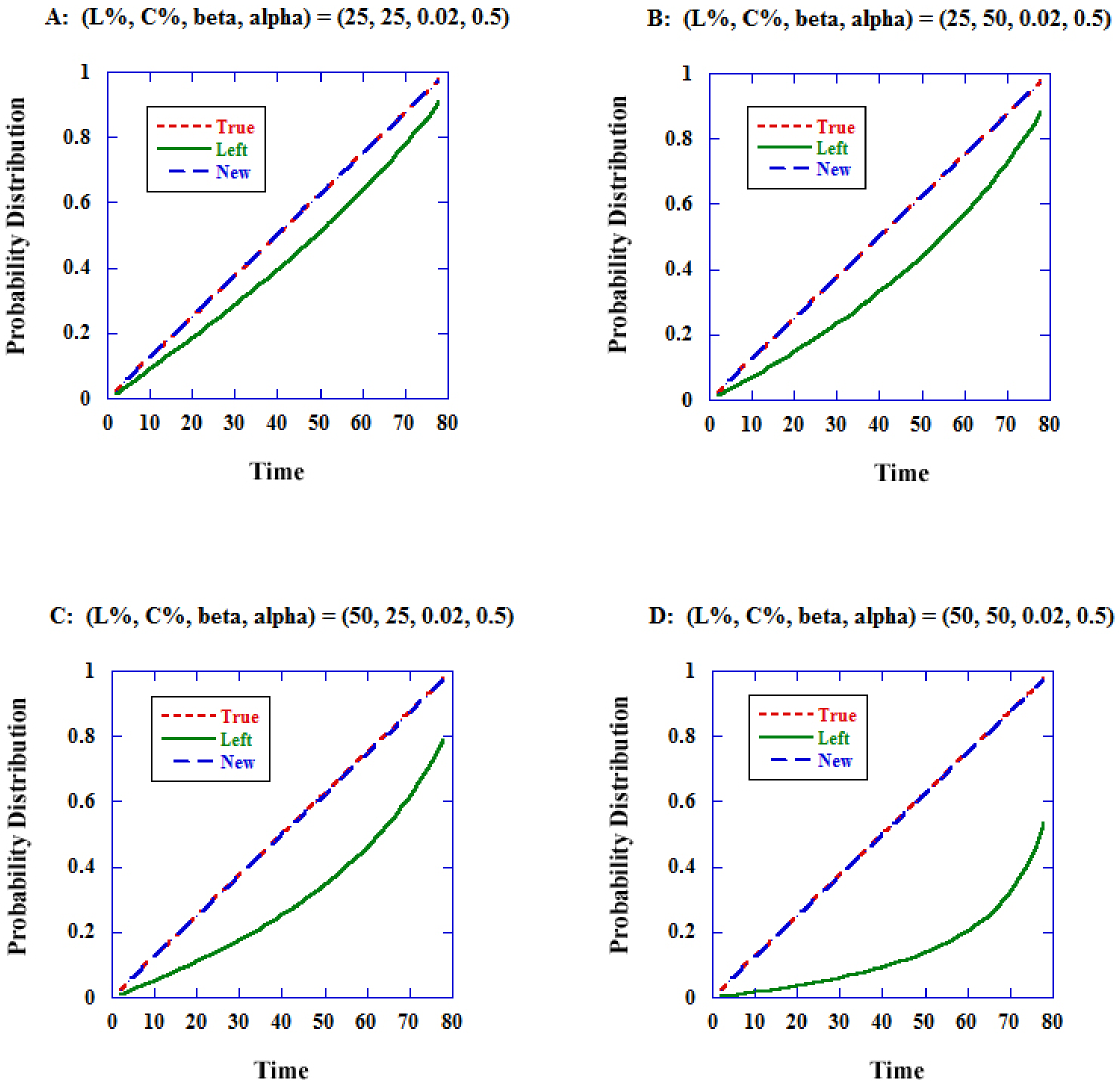
$$v\hat{a}r(\hat{G}(t;\hat{\beta})) = \frac{1}{1000}\sum_{i=1}^{1000}\hat{\sigma}^2_{(i)}(t)$$

where $\hat{G}_{(i)}(t;\hat{\beta})$ be the IPW estimate for the ith replicate at time t, which is discussed in Chapter 2.2. $\bar{G}^{\mathrm{IPW}}(t)$ be the average IPW estimate across 1000 replicates. $\hat{\sigma}^2_{(i)}(t)$ be the estimated variance of IPW estimate for the ith replicate using the the variance estimation result given in Chapter 3.2.

Regarding the distribution function of the truncation time L, we implement two methods: the Kaplan-Meier estimator for independently truncated sample given in (2.4) and the proposed IPW estimator. For each method, we find the average of the 1000 estimates at the predetermined times and plot the averages against the times. We also depict the true distribution function in each figure. Figure 4.1-4.4

and Figure 4.5-4.6 describe the estimation results for setting with continuous covariate and discrete covariate, respectively. The dotted line is the true value (“true”), the solid line is the naïve Kaplan-Meier estimation (“left”), the long dashed line is the proposed IPW estimator (“new”). We can see the bias clearly for the naïve Kaplan-Meier estimator, while the result from our new method closely matches the true function, indicating the distribution of L is precisely estimated using the proposed IPW estimator.

The simulation results for variance estimation and confidence interval coverage are given in Table 4.1-4.6. The tables show a good performance of the proposed variance estimator at at different time points when $G(t)$ is evaluated to be 0.25, 0.5, 0.75. The average of the estimated variances closely matches the variance pertaining to 1000 cumulative probability estimates. The actual coverage of the confidence intervals is very close to the nominal level, except for the settings with heavy censoring and truncation. A slight higher degree of departure is observed between sample variance and estimated variance for settings with continuous covariate when the truncation rate or censoring rate is heavy.



A: (L%, C%, beta, alpha) = (25, 25, 0.02, 0.5)

B: (L%, C%, beta, alpha) = (25, 50, 0.02, 0.5)

C: (L%, C%, beta, alpha) = (50, 25, 0.02, 0.5)

D: (L%, C%, beta, alpha) = (50, 50, 0.02, 0.5)

Figure 4.1. Estimated distribution function of L for the setting with a continuous covariate ($\beta = 0.02, \alpha = 0.5$).

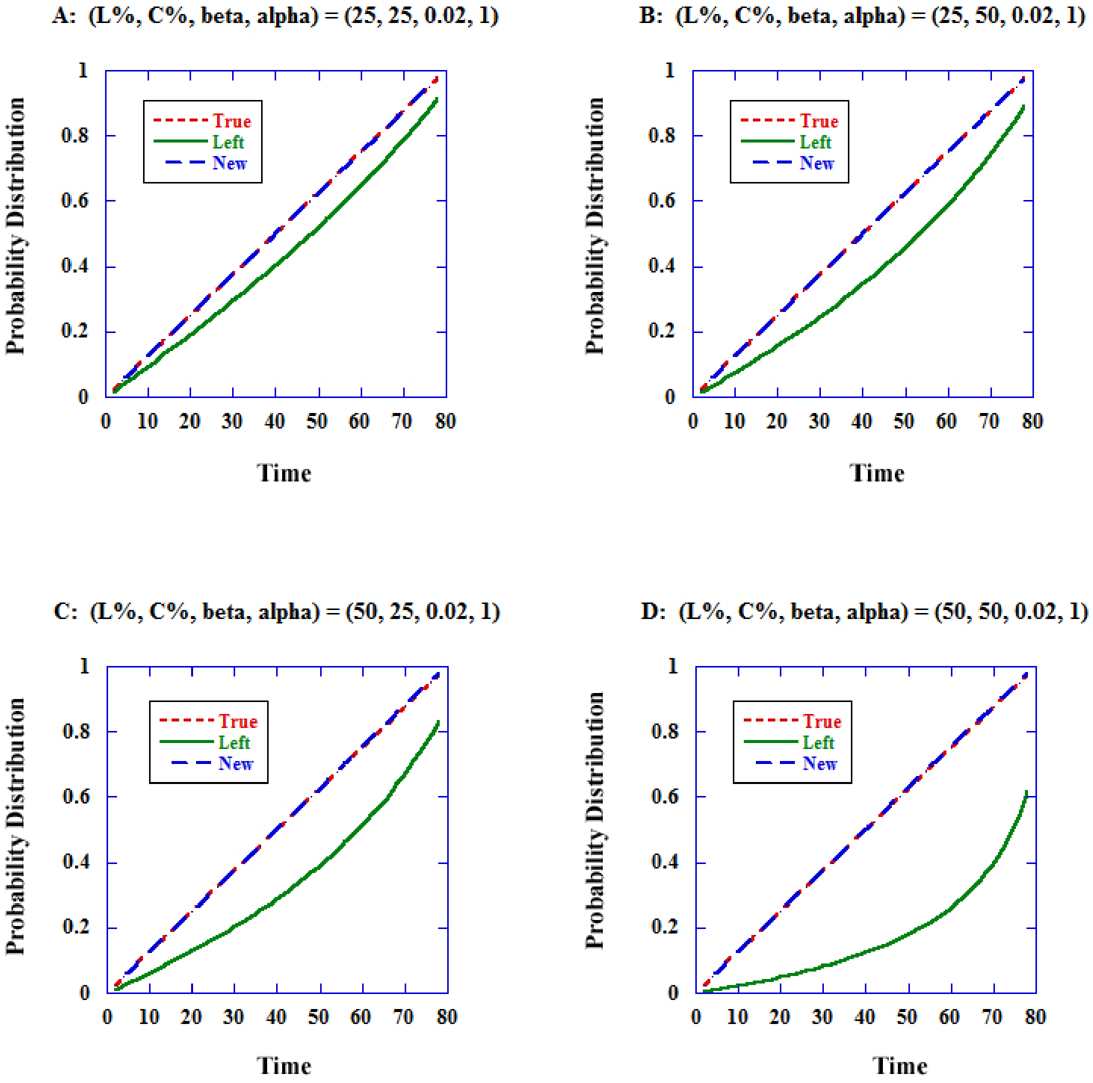


Figure 4.2. Estimated distribution function of L for the setting with a continuous covariate ($\beta = 0.02, \alpha = 1$).

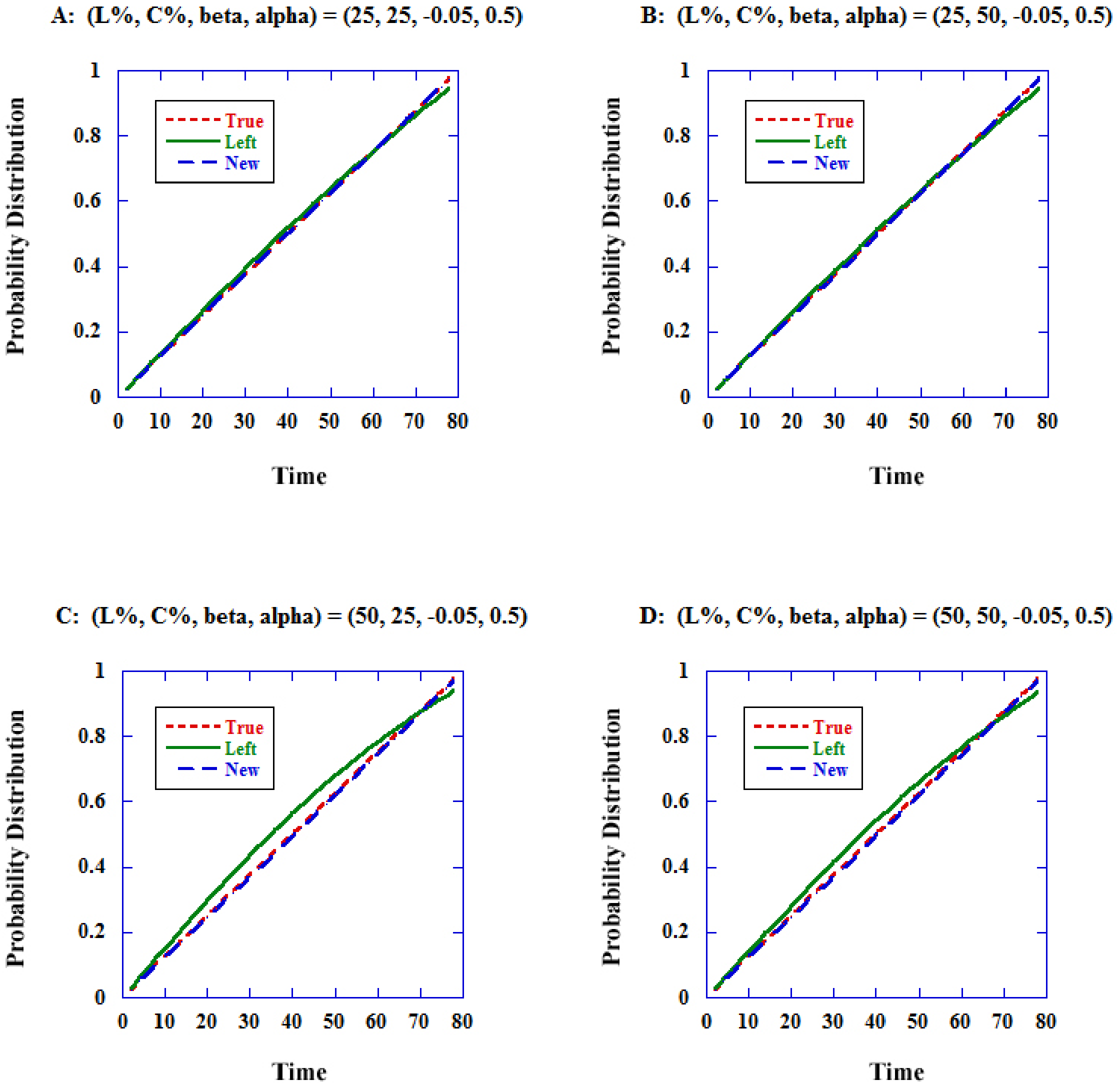


Figure 4.3. Estimated distribution function of L for the setting with a continuous covariate ($\beta = -0.05, \alpha = 0.5$).

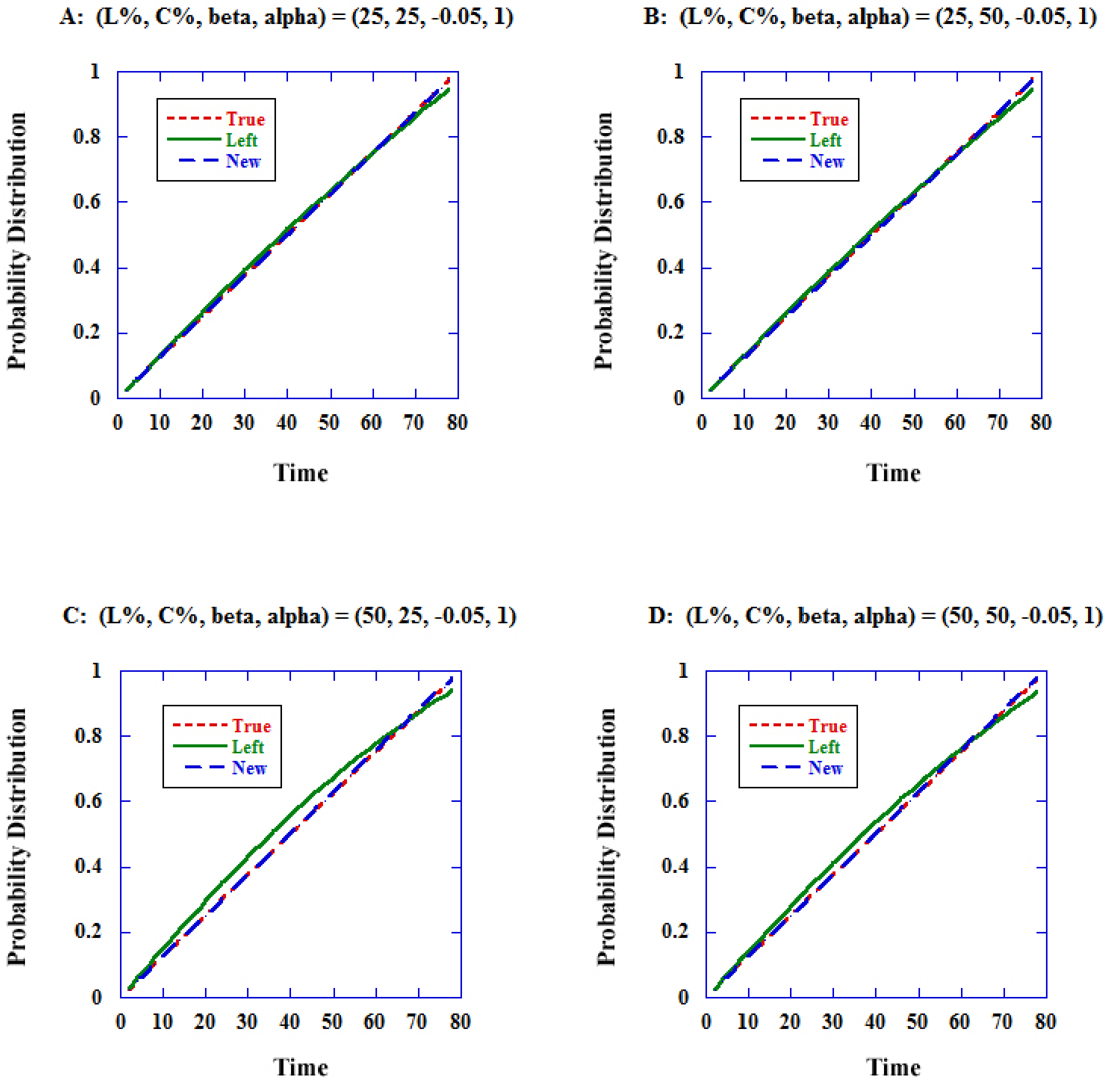


Figure 4.4. Estimated distribution function of L for the setting with a continuous covariate $(\beta = -0.05, \alpha = 1)$.

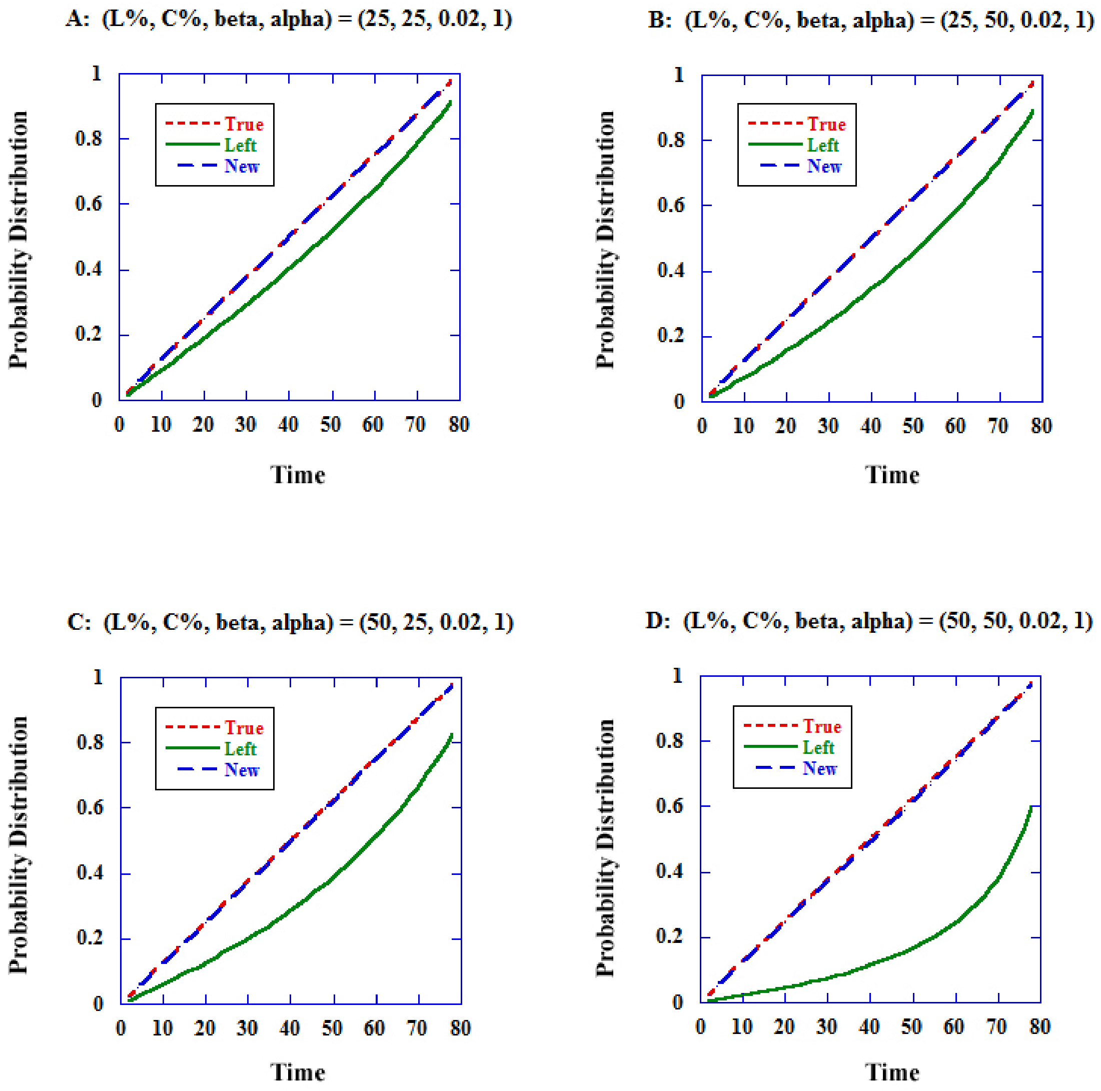


Figure 4.5. Estimated distribution function of L for the setting with a discrete covariate ($\beta = 0.02, \alpha = 1$).

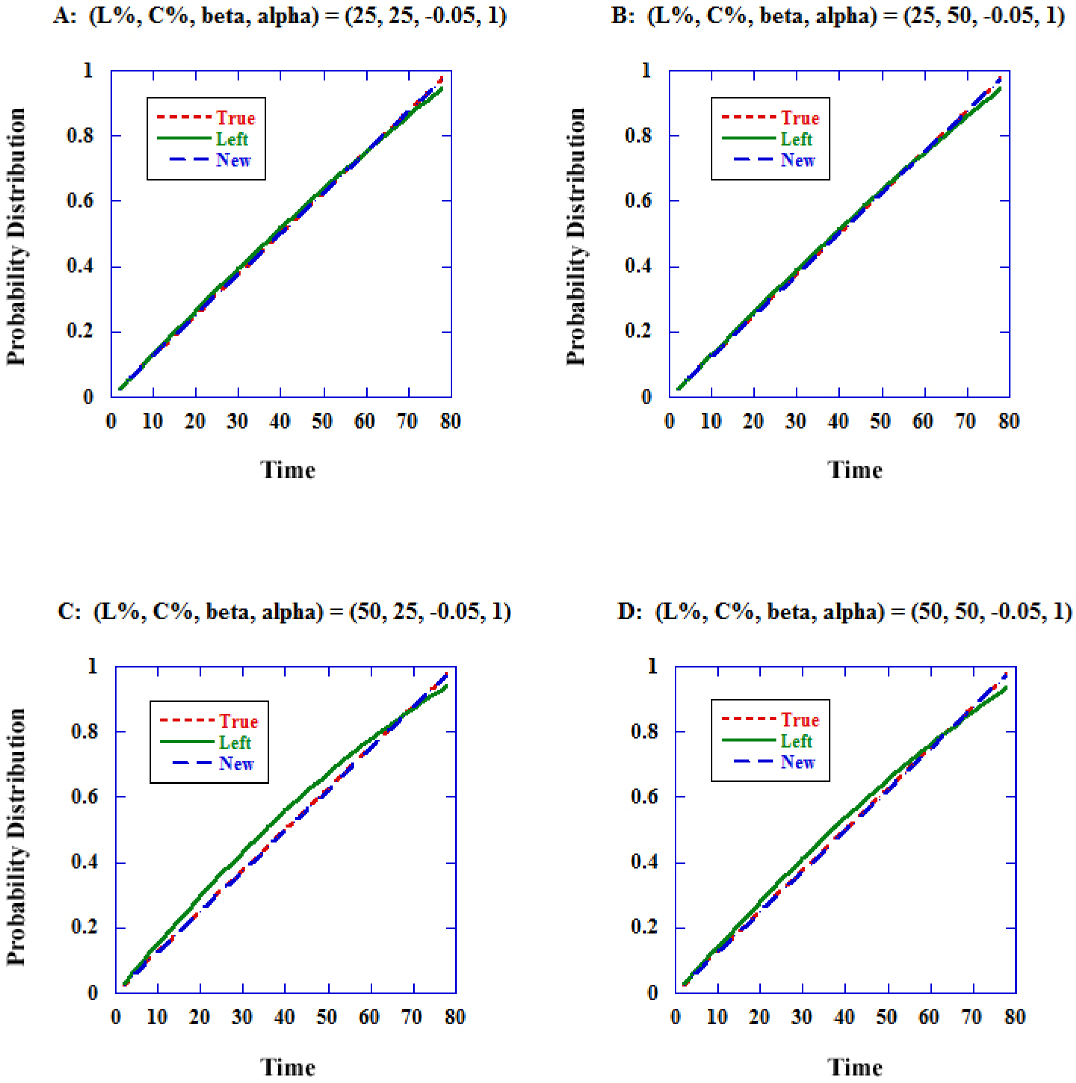


Figure 4.6. Estimated distribution function of L for the setting with a discrete covariate ($\beta = -0.05, \alpha = 1$).

Table 4.1. 95% confidence interval coverage for estimating $G(t)$ for the setting with a continuous covariate ($\beta = 0.02, \alpha = 0.5$).

(L%, C%)	$G(t)$	Bias	$var(\hat{G}(t;\hat{\beta}))$	$\hat{var}(\hat{G}(t;\hat{\beta}))$	95% CI cov.
(25, 25)	0.25	0.001	0.0009	0.0010	0.953
	0.50	0.001	0.0015	0.0015	0.955
	0.75	0.000	0.0013	0.0012	0.966
(25, 50)	0.25	-0.000	0.0010	0.0010	0.946
	0.50	0.000	0.0017	0.0016	0.942
	0.75	-0.001	0.0013	0.0013	0.966
(50, 25)	0.25	-0.001	0.0014	0.0014	0.952
	0.50	-0.001	0.0031	0.0030	0.935
	0.75	-0.004	0.0030	0.0031	0.948
(50, 50)	0.25	-0.000	0.0019	0.0023	0.970
	0.50	0.000	0.0042	0.0055	0.954
	0.75	-0.001	0.0041	0.0055	0.926

Table 4.2. 95% confidence interval coverage for estimating $G(t)$ for the setting with a continuous covariate ($\beta = 0.02, \alpha = 1$).

(L%, C%)	$G(t)$	Bias	$var(\hat{G}(t;\hat{\beta}))$	$v\hat{a}r(\hat{G}(t;\hat{\beta}))$	95% CI cov.
(25, 25)	0.25	0.001	0.0010	0.0010	0.942
	0.50	-0.000	0.0016	0.0016	0.947
	0.75	-0.001	0.0013	0.0014	0.968
(25, 50)	0.25	0.000	0.0010	0.0010	0.953
	0.50	-0.000	0.0017	0.0017	0.944
	0.75	-0.001	0.0014	0.0015	0.961
(50, 25)	0.25	0.000	0.0014	0.0017	0.969
	0.50	0.001	0.0031	0.0038	0.955
	0.75	0.004	0.0032	0.0039	0.917
(50, 50)	0.25	0.001	0.0018	0.0023	0.964
	0.50	0.001	0.0042	0.0057	0.950
	0.75	0.004	0.0039	0.0055	0.908

Table 4.3. 95% confidence interval coverage for estimating $G(t)$ for the setting with a continuous covariate ($\beta = -0.05, \alpha = 0.5$).

(L%, C%)	$G(t)$	Bias	$var(\hat{G}(t;\hat{\beta}))$	$\hat{var}(\hat{G}(t;\hat{\beta}))$	95% CI cov.
(25, 25)	0.25	0.001	0.0010	0.0010	0.947
	0.50	0.000	0.0017	0.0017	0.945
	0.75	-0.001	0.0014	0.0014	0.968
(25, 50)	0.25	-0.000	0.0011	0.0011	0.944
	0.50	-0.001	0.0016	0.0017	0.954
	0.75	-0.003	0.0014	0.0014	0.969
(50, 25)	0.25	-0.004	0.0013	0.0014	0.957
	0.50	-0.008	0.0027	0.0030	0.966
	0.75	-0.007	0.0026	0.0029	0.958
(50, 50)	0.25	-0.004	0.0015	0.0016	0.949
	0.50	-0.007	0.0031	0.0034	0.954
	0.75	-0.007	0.0031	0.0033	0.967

Table 4.4. 95% confidence interval coverage for estimating $G(t)$ for the setting with a continuous covariate ($\beta = -0.05, \alpha = 1$).

(L%, C%)	$G(t)$	Bias	$var(\hat{G}(t;\hat{\beta}))$	$v\hat{a}r(\hat{G}(t;\hat{\beta}))$	95% CI cov.
(25, 25)	0.25	0.001	0.0011	0.0011	0.937
	0.50	0.001	0.0019	0.0018	0.938
	0.75	-0.001	0.0015	0.0015	0.963
(25, 50)	0.25	0.001	0.0011	0.0011	0.937
	0.50	0.001	0.0019	0.0018	0.938
	0.75	-0.001	0.0015	0.0015	0.963
(50, 25)	0.25	0.000	0.0016	0.0018	0.948
	0.50	0.001	0.0034	0.0039	0.943
	0.75	0.007	0.0030	0.0037	0.924
(50, 50)	0.25	-0.000	0.0016	0.0019	0.953
	0.50	0.002	0.0033	0.0042	0.960
	0.75	0.006	0.0032	0.0040	0.920

Table 4.5. 95% confidence interval coverage for estimating $G(t)$ for the setting with a discrete covariate ($\beta = 0.02, \alpha = 1$).

(L%, C%)	$G(t)$	Bias	$var(\hat{G}(t;\hat{\beta}))$	$v\hat{a}r(\hat{G}(t;\hat{\beta}))$	95% CI cov.
(25, 25)	0.25	0.001	0.0010	0.0010	0.942
	0.50	-0.000	0.0016	0.0016	0.947
	0.75	-0.001	0.0013	0.0014	0.968
(25, 50)	0.25	-0.000	0.0010	0.0010	0.946
	0.50	0.000	0.0015	0.0016	0.955
	0.75	-0.000	0.0013	0.0014	0.966
(50, 25)	0.25	-0.001	0.0016	0.0016	0.947
	0.50	-0.003	0.0035	0.0037	0.955
	0.75	-0.001	0.0037	0.0040	0.934
(50, 50)	0.25	-0.003	0.0020	0.0024	0.957
	0.50	-0.008	0.0048	0.0060	0.942
	0.75	-0.007	0.0050	0.0065	0.925

Table 4.6. 95% confidence interval coverage for estimating $G(t)$ for the setting with a discrete covariate ($\beta = -0.05, \alpha = 1$).

(L%, C%)	$G(t)$	Bias	$var(\hat{G}(t;\hat{\beta}))$	$\hat{var}(\hat{G}(t;\hat{\beta}))$	95% CI cov.
(25, 25)	0.25	0.001	0.0011	0.0011	0.937
	0.50	0.001	0.0019	0.0018	0.938
	0.75	-0.001	0.0015	0.0015	0.963
(25, 50)	0.25	0.000	0.0010	0.0011	0.953
	0.50	-0.000	0.0018	0.0017	0.954
	0.75	-0.001	0.0016	0.0015	0.956
(50, 25)	0.25	0.000	0.0016	0.0017	0.962
	0.50	-0.002	0.0034	0.0036	0.953
	0.75	-0.001	0.0036	0.0036	0.927
(50, 50)	0.25	0.000	0.0016	0.0017	0.962
	0.50	-0.002	0.0034	0.0036	0.953
	0.75	-0.001	0.0036	0.0036	0.927

Chapter 5

A REAL EXAMPLE

5.1 Data Description

In this chapter, we analyze a transplant outcome data set from The Center for International Blood and Marrow Transplant Research (CIBMTR). The CIBMTR is comprised of clinical and basic scientists who confidentially share data on their blood and bone marrow transplant patients with CIBMTR Data Collection Center located at the Medical College of Wisconsin. The CIBMTR is a repository of information about results of transplants at more than 450 transplant centers worldwide. In our case, 376 children who received transplantation in second complete remission are selected. Since only the patients who received transplants are observed and patients who died while waiting for transplantation would not be included, the BMT group is a truncated sample.

The BMT sample, jointly with a sample of 540 children receiving chemotherapy, was analyzed by Barrett et al. (1994) to assess the treatment effect on the leukemia-free survival. They conducted Cox analysis on the BMT sample and identified the following significant risk factors for the leukemia-free survival at 0.10 levels: age ($>$ 10 yr, $\leq$ 10 yr), the T-cell phenotype (no, yes) and duration of the first remission ($\leq$ 18 months; $>$ 18 months). In his study, the effect of transplant time was not considered. Barrett's Cox analysis results were summarized in Table 5.1. We will compare this result with our new analysis which transplant time is included as a predictor.

Table 5.1. Regression coefficient estimates for the Cox model on the BMT sample.

Parameter	Barrett's Study		New analysis	
	Relative risk	P-value	Relative risk	P-value
Transplant time	-	-	1.357	0.0295
Age >10	1.51	0.003	1.374	0.0214
T cell phenotype	2.16	< 0.001	2.025	0.0003
Duration of the first remission ≤ 18	2.02	< 0.001	1.504	0.0043

5.2 Cox analysis

In Cox model 3.1, $k(\beta, L)$ indicates that a particular functional forms of L should be included in the regressor. The following simple forms $L, L^2 and \sqrt{L}$ were considered for the functional form of transplant time. We found that the quadratic form L^2 yielded the highest level of significance. Therefore, the quadratic transplant time was included in the Cox regression. A model-building procedure was used to search for the significant risk factors with p-value 0.05 as the threshold. Four risk factors, transplant time, age, duration of first remission, and T-cell phenotype were identified to be significant factors. As can be seen in Table 5.1, the relative risks of age, duration of first remission, and T-cell phenotype are all comparable to those in Barrett's study. The relative risks are estimated as 1.374 [95% CI (1.048, 1.800)] for patients with Age > 10, 2.025 [95% CI (1.387, 2.956) for patients with T cell phenotype and 1.504 [95% CI (1.136, 1.989)] for patients with duration of first remission in ≤ 18 months, respectively.

An important finding in our study is to find out the effect of transplant time. As can be seen in Table 5.1, the relative risk is 1.357 [95% CI (1.037, 1.786)]. This positive estimated regression coefficient for transplant time means that the long waiting time for transplant will lead to a higher rate of failure at future time. Suppose that there are two leukemia patients. One has the bone marrow transplant 6 months after diagnosis and the other one has the transplant 18 months after diagnosis. After their transplants, if both are alive at time t, then the patient who has the transplant one year later is 35.7% more likely to experence relapse or mortality. The finding that a longer waiting time is a poor prognosis of leukemia-free survival agrees well with the recent clinical observation (Balduzzi, 2008; Davies, 2010).

5.3 Distribution function of the transplant time

In BMT studies, the truncation time is the transplant time, which is dominantly determined by the donor search process. Since the transplant is the major surgical procedure and consequently dramatically alerts the pattern of survivorship, it is crucial to find the marginal distribution of transplant time. We propose an IPW estimator and use it to estimate the distribution function of L. The estimation result is plotted and 95% confidence intervals is also shown in Figure 5.1.

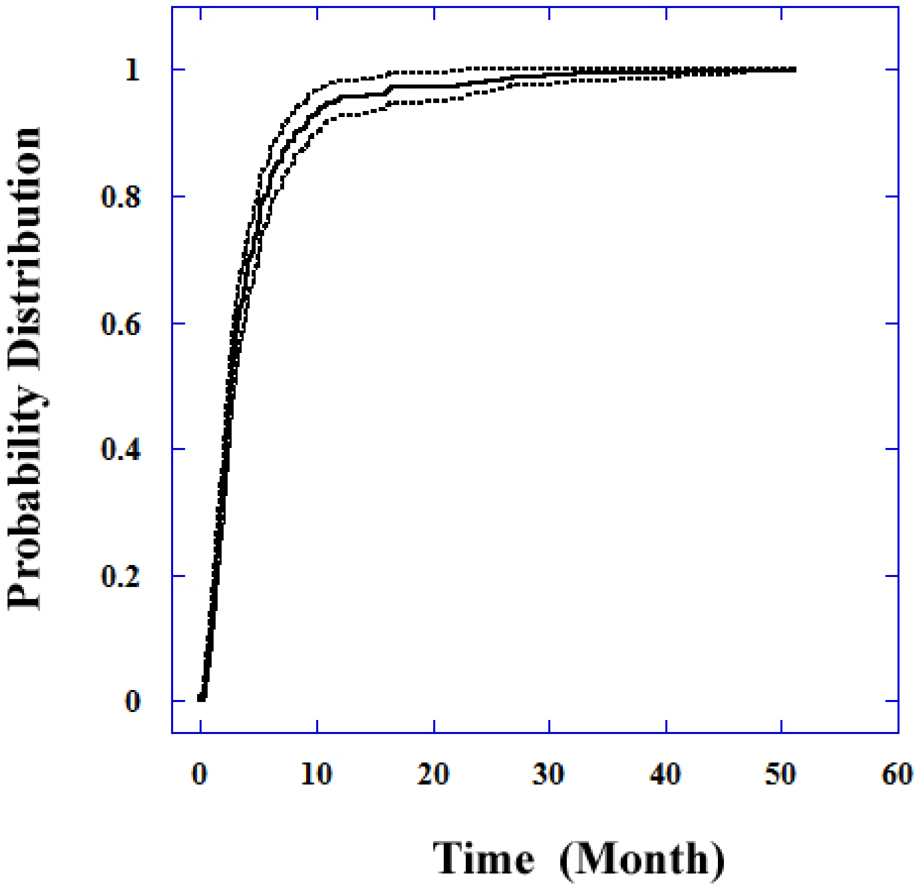


Figure 5.1. Estimated distribution function of the transplant time and 95% confidence intervals

Chapter 6

CONCLUSION

The study cohort obtained from Bone Marrow Transplant (BMT) registry data are commonly recognized as truncated samples, because the participating hospitals do not report data on patients died while waiting for transplants. The current analytical methods on the pooled samples include the matched pairs analysis and the Cox analysis assuming a constant effect for transplant. However, the effect of the transplant time cannot be evaluated using the above analytical approaches. In this thesis, we use a Cox model for analyzing the left-truncated data with the dependently truncation time L and failure time T. We also proposed an inverse probability weighted estimator to estimate the distribution of the transplant time. Simulation studies have been conducted to investigate the performances of the new IPW estimators and a variance estimator. A real data example was also applied to the proposed method.

The future direction of this work will be focused on the application of the new inverse probability weighting approach to more real data sets. For example, in BMT study, we can determine the effect of transplant time on different race groups based on our proposed method. This will provide valuable information on the survival of patients affected by the transplant time from different race groups. Our new inverse-probability-weighted approach will be more efficient since we consider the dependence of the truncation distribution on the covariate.

REFERENCES

[1] Aalen, O.O., A linear regression model for the analysis of life times, *Stat Med*, Vol. 8, pp. 907-925, 1989.

[2] Balduzzi, A., De Lorenzo, P., Schrauder, A., Conter, V., Uderzo, C., Peters, C., Klingebiel, T., Stary, J., Felice, M.S., Magyarosy, E., Eligibility for allogeneic transplantation in very high risk childhood acute lymphoblastic leukemia: the impact of the waiting time, *Haematologica*, Vol. 93, pp. 925-929, 2008.

[3] Barrett, A.J., Horowitz, M.M., Pollock, B.H., Zhang, M.J., Bortin, M.M., Buchanan, G.R., Camitta, B.M., Ochs, J., Graham-Pole, J., Rowling, P.A., Rimm, A.A., Klein, J.P., Shuster, J.J., Sobocinski, K.A., Gale, R.P., HLA-identical sibling bone marrow transplants versus chemotherapy for children with acute lymphoblastic leukemia in second remission, *The New England Journal of Medicine*, Vol. 331, pp. 1253-1258, 1994.

[4] Breslow, N.E., Covariance analysis of censored survival data, *Biometrics*, Vol. 30, pp. 579-594, 1974.

[5] Chaieb L.L., Rivest, L.P., Abdous, B., Estimating survival under a dependent truncation, *Biometrika*, Vol. 93, pp. 655-669, 2006.

[6] Cox, D.R., Regression models and life tables, *J Roy Statist Soc Ser B*, Vol. 34, pp. 187-220, 1972.

[7] Cox, D.R., Partial likelihood, *Biometrika*, Vol. 62, pp. 269-276, 1975.

[8] Davies, S.M., Mehta, P.A., Pediatric acute lymphoblastic leukemia: is there still a role for transplant? *Hematology / the Education Program of the American Society of Hematology American Society of Hematology*, pp. 363-367, 2010.

[9] Emma, T., Konno, Y., Multivariate normal distribution approaches for dependently truncated data, *Stat Papers*, DOI:10.1007/s00362-010-0321-x, 2010.

[10] Greenwood, M., The natural duration of cancer, *Repoerts on Public Health and Medical Subjects*, Vol. 33, pp. 1-26, 1926.

[11] Horvitz, D.G., Thompson, D.J., A generalization of sampling without replacement from a finite universe, *J Amer Statist Assoc*, Vol. 47, pp. 663-685, 1952.

[12] Kalbfleisch, J.D., Prentice, R.L., The statistical analysis of failure time data, *New York, John Wiley & Sons, Inc*, 1980.

[13] Kaplan, E., Meier, P., Nonparametric estimation from incomplete observations, *J Am Statist Assoc*, Vol. 84, pp. 360-372, 1958.

[14] Karlsson, M., Laitila, T., A semiparametric regression estimator under left truncation and right censoring, *Statist Probab Lett*, Vol. 78, pp. 2567-2571, 2008.

[15] Keiding, N., Independent delayed entry, *Boston, Kluwer*, 1992.

[16] Keiding, N., Gill, R.D., Random truncation models and Markov process, *The Annals of Statistics*, Vol. 66, pp. 382-392, 2010.

[17] Klein, J.P., Zhang, M.J., Statistical challenges in comparing chemotherapy and bone marrow transplantation as a treatment for leukemia, *Life data: models in reliability and survival analysis*, pp. 175, 1996.

[18] Li, J., COX model analysis with the dependently left truncated data, *Thesis of Georgia State University*, 2010.

[19] Lin, D.Y., Ying, Z., Semi-parametric analysis of the additive risk model, *Biometrika*, Vol. 81, pp. 61-71, 1994.

[20] Lynden-Bell, D., A method of allowing for known observational selection in small samples applied to 3CR quasars, *Mon Not R Astr Soc*, Vol. 155, pp. 95-118, 1971.

[21] McKeague, I.W., Sasieni, P.D., A partly parametric additive risk model, *Biometrika*, Vol. 81, pp. 501-514, 1994.

[22] Robins, J.M., Finkelstein, D., Correcting for Non-compliance and Dependent Censoring in an AIDS Clinical Trial with Inverse Probability of Censoring Weighted (IPCW) Log-rank Tests, *Biometrics*, Vol. 56, pp. 779-788, 2000.

[23] Satten, G.A., Datta, S., The kaplan-Meier estimator as an inverse-probability-of-censoring weighted average, *Amer Statist Ass*, Vol. 55, pp. 207-210, 2001. 55,

[24] Shen, P.S., The product-limit estimate as an inverse-probability-weighted average, *Communications in Statistics*, Vol. 32, pp. 1119-1133, 2003.

[25] Shen, P.S., An inverse-probability-weighted approach to estimation of the bivariate survival function under left-truncation and right censoring, *J Statist Plan Infer*, Vol. 136, pp. 4365-4384, 2006.

[26] Shen, P.S., Semiparametric estimation of survival function when data are subject to dependent censoring and left truncation, *Statist Probab Lett*, Vol. 80, pp. 161-168, 2010.

[27] Tsai, W.Y., Testing the assumption of the independence of truncation time and failure time, *Biometrika*, Vol. 77, pp. 169-177, 1990.

[28] Wang, M.C., a semiparametric model for randomly truncated data, *J Am Statist Assoc*, Vol. 84, pp. 742-748, 1989.

[29] Wang, M.C. Jewell, N.P., Tsai, W.Y., Asymptotic properties of the product limit estimate under random truncation, *The Annals of Statistics*, Vol. 14, pp. 1597-1605, 1986.

[30] Woodroofe, M., Estimating a distribution function with truncated data, *The Annals of Statistics*, Vol. 13, pp. 163-177, 1985.

[31] Vardi, Y., Empirical distributions in selection bias models, *The Annals of Statistics*, Vol. 13, pp. 178-203, 1985.