Thesis for the degree of Master of Science

Unbiased Transformation of Censored Survival Data

Lee Eunju

Department of Statistics
The Graduate School
Pusan National University

February 2023

Unbiased Transformation of Censored Survival Data **Lee Eunju** **Feb 2023**

Unbiased Transformation of Censored Survival Data

A Thesis submitted to the graduate school of
Pusan National University in partial fulfillment of
the requirements for the degree of Master of Science
under the direction of Kim Choongrak

The thesis for the degree of Master of Science
by Lee Eunju
has been approved by the committe members.

2022. 12. 28.

Chair	Lee Donghyuk	_____
Member	Yang Hojin	_____
Member	Kim Choongrak	_____

CONTENTS

page

List of Figures

page

List of Tables

page

Unbiased Transformation of Censored Survival Data

Lee Eunju

Department of Statistics
The Graduate School
Pusan National University

Abstract

Censored observations are inevitable in the survival time data, and the incomplete data due to censoring is not desirable in the sense of two reasons. First, some information is lost due to censoring, and secondly, statistical tools for analyzing censored data are restrictive. Therefore, it would be good if we transform the censored observations to uncensored observations. In this paper, we consider three kinds of unbiased transformation of censored survival times to uncensored ones; Method 1 is a Buckley-James type transformation, Method 2 is a Koul-Susarla-van Ryzin type transformation, and Method 3 is a mean imputation method. To compare the numerical performance of the three methods with the Kaplan-Meier estimator, we evaluate MISE (mean integrated squared error) of the distribution function estimates, MSE (mean squared error) of the median survival time estimates, and the c-index, estimate of the concordance between the observed and the predicted values. Conclusively, the mean imputation showed the best results under various censoring percentages. We can expect to extend the suggested methods to the regression problem in the censored survival times.

CHAPTER 1. INTRODUCTION

Survival analysis is an analysis of the time from the start of observation to the time of the event. Events can be various experiences that we are interested in, such as disease outbreak, recurrence, or death. Therefore, survival analysis is widely used in various fields such as medical, economy, and marketing. That is, the variable of interest in the survival analysis is survival time, T. The goal of survival analysis is to estimate and compare survival functions or to perform survival regression model with explanatory variables.

However, survival analysis deals with incomplete data due to censoring. The censored observation means that observation was stopped before the event of interest occurred. This is an important problem in survival analysis because some information for the purpose of the experiment is lost.

In previous studies, survival analysis was conducted by nonparametric approach. One of the most popular nonparametric estimator of the survival function is the Kaplan-Meier estimator (Kaplan and Meier, 1958). And the log-rank test (Mantel, 1966) is a hypothesis test to compare survival functions between the two groups. The Cox proportional hazards model (Cox, 1972) is the most popular regression model in survival analysis.

Nevertheless, the censored data have a difficulty in applying to various statistical analysis methods recently developed. This is a clear limitation in survival analysis. Accordingly, this paper tries to solve the core problem related to censored data. To this end, the approach we propose is to transform censored data into completed data.

There are three proposed methods for transforming censored data. The first method is Buckley-James type transformation based on the unbiased estimate for the mean of survival time. The second method is Koul-Susarla-van Ryzin type transformation to the approximate unbiased estimate of the mean of survival time. The last method is mean imputation method.

In the numerical experiments, we compared estimators of the distribution function and estimators of the median survival time with Kaplan-Meier estimate through the mean integrated squared error (MISE) and the mean squared error (MSE), respectively. We compared based on the concordance index (c-index) for the three methods. The results show that among the three methods, the mean imputation method has higher estimation performance than Kaplan-Meier estimator when censoring percentage is large.

This paper is organized as follows. In Chapter 2, we review the basic functions used in survival analysis and the method of estimating survival functions for censored data in previous studies. In Chapter 3, we propose three methods of transforming censored data into completed data. There are the results of the numerical analysis in Chapter 4, and the conclusion with future research in Chapter 5.

CHAPTER 2. SURVIVAL ANALYSIS

2.1 Basic Functions

Let T be a non-negative random variable representing the survival time with probability distribution function $f(t)$ and cumulative distribution function $F(t)$. Three basic functions are used in survival analysis.

First, survival function $S(t)$ is the probability of survival after time t, so that the sum with the cumulative distribution function which denotes the probability that the event occurs before time t is always 1. Second, hazard function $\lambda(t)dt$ is the conditional propability as the probability of death in $[t,\ t+dt)$ when surviving more than time t. Third, cumulative hazard function $\Lambda(t)$ is the integral of the hazard function. That is, the three functions are defined as follows.

$$
\begin{aligned}
S(t) &= P(T > t) = 1 - F(t) \\
\lambda(t) &= \lim_{dt \to 0} \frac{P(t \le T < t + dt \mid T \ge t)}{dt} = \frac{f(t)}{S(t)} \\
\Lambda(t) &= \int_0^t \lambda(u) du
\end{aligned}
$$

Let C be the censoring time. The response variable Y is the minimum of the survival time T and the censoring time C, and it is denoted as $Y = min(T, C)$ or (Y_i, δ_i), $i = 1, \cdots, n$ where $\delta_i = I(T_i \le C_i)$. That is, $(Y_i, 1)$ is uncensored data and $(Y_i, 0)$ is censored data. In addition, $T \sim F$ and $C \sim G$ are independent. Given the fact, the probability density function of Y is as follows.

$$
P(Y_i, \delta_i) = \begin{cases} P(Y_i = T_i, C_i > Y_i) = f(Y_i)(1 - G(Y_i)) & \text{if } \delta_i = 1 \\ P(Y_i = C_i, T_i > Y_i) = g(Y_i)(1 - F(Y_i)) & \text{if } \delta_i = 0 \end{cases} \tag{1}
$$

2.2 Kaplan-Meier Estimator

The Kaplan-Meier estimator is the nonparametric estimator of the survival function. Suppose data $(y_i,\ \delta_i),\ i = 1 \cdots n$ where $y_1 < y_2 < \cdots < y_n$. Let n_i denote the number of subjects at risk at y_i^- named risk set at y_i, which is $n - i + 1$.

The survival probability p_i is the conditional probability as the probability of surviving at y_i when alive at y_i^-. That is, p_i is $P(T > y_i | T > y_{i-1})$ and is estimated that 1 when $\delta_i = 0$ and $n_{i+1}/n_i = 1 - 1/(n - i + 1)$ when $\delta_i = 1$. The survival function at time t is the product of survival probabilities prior to time t, which is called Kaplan-Meier estimator and is as follws.

$$\hat{S}(t) = \prod_{i:y_i \leq t} (1 - \frac{\delta_i}{n - i + 1}) = \prod_{i:y_i \leq t} (1 - \frac{1}{n - i + 1})^{\delta_i}$$

2.3 Kernel Estimator

Assume that random variables $Y_1, \cdots Y_n$ follow $f(y)$ and $F(y)$. Kernel estimator estimates the probability density function and cumulative distribution function in a nonparametric method based on sample data $y_1, \cdots y_n$. The kernel estimator is proposed by Rosenblatt (1956) and Parzen (1962). The method to estimate is a sum of jumps at each data point.

$$\hat{f}(t) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{t - y_i}{h})$$

Kernel function $K(t)$ represents the shape of the bumps, that is weighting function. It is usually a symmetric function and should satisfy the following.

$$\int K(t)dt = 1, \int K(t)dt = 0, \int t^2 K(t)dt < \infty$$

The widely used kernel functions are Uniform kernel function , Gaussian kernel function or Epanechnikov kernel function.

$$\begin{aligned} \text{Uniform} : K(t) &= \frac{1}{2} I(|t| \leq 1) \\ \text{Gaussian} : K(t) &= \frac{1}{\sqrt{2\pi}} \exp^{-t^2/2} \\ \text{Epanechnikov} : K(t) &= \frac{3}{4}(1-t^2) I(|t| < 1) \end{aligned} \tag{2}$$

Bandwidth h is the amount of smoothing and is also called smoothing parameter. The method of choosing the smoothing parameter is mainly cross validation or generalized cross validation.

In survival analysis, the kernel density estimate $\hat{f}(t)$ and the kernel distribution function estimate $\hat{F}(t)$ are as follows.

$$\hat{f}(t) = \frac{1}{h} \sum_{i=1}^{n} s_i K(\frac{t - y_i}{h}) \tag{3}$$

$$\hat{F}(t) = \frac{1}{h} \sum_{i=1}^{n} s_i W(\frac{t - y_i}{h}) \tag{4}$$

where s_i is the jump size at y_i in Kaplan-Meier estimator and $W(t)$ is a cumulative kernel as follows.

$$W(t) = \int_{-\infty}^{t} K(x) dx$$

CHAPTER 3. UNBIASED TRANSFORMATION OF CENSORED SURVIVAL DATA

3.1 Method 1: Buckley-James type Transformation

In this chapter, we propose three methods of unbiased transformation of censored survival data. As shown in (1), the propability of y with $\delta = 0$ is $g(y)(1 - F(y))$, it can be shown that

$$Y^{(1)} = \frac{f(Y)}{g(Y)(1 - F(Y))}\ Y$$

is unbaised estimator of $E(T)$ when $f(y)$, $F(y)$ and $g(y)$ are known because

$$\begin{aligned} E(Y^{(1)}) &= \int_0^\infty \hat{u}^* g(u)(1 - F(u))du \\ &= \int_0^\infty \frac{f(u)}{g(u)(1 - F(u))} u\ g(u)(1 - F(u))du \\ &= \int_0^\infty uf(u)\ du \\ &= E(T) \end{aligned}$$

However, we do not know $f(y)$, $g(y)$ and $F(y)$, so these are estimated using kernel estimator. That is, $f(y)$ and $F(y)$ are estimated by (3) and (4), respectively. And for the estimation of $g(y)$, it is based on jump size s_i in Kaplan-Meier estimator using data $(y_i, 1 - \delta_i),\ i = 1, \cdots, n$.

$$\hat{g}(t) = \frac{1}{h} \sum_{i=1}^{n} s_i K(\frac{t - y_i}{h})$$

We focuse on Epanechnikov kernel function of (2) because it is the most widely used. The bandwidth is estimated using likelihood cross validation method. It selects h to maximize log-likelihood $CV(h)$.

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n} log\hat{f}_{-i}(y_i)$$

where $\hat{f}_{-i}(y_i)$ is the estimate at y_i with kernel density estimate using data that removed y_i.

$$\hat{f}_{-i}(y_i) = \frac{1}{h}\sum_{j\neq i} s_j K(\frac{y_i - y_j}{h})$$

The first method to transform of censored data to uncensored data is as follows.

$$\hat{y}_i^{(1)} = \delta_i y_i + (1-\delta_i)\frac{\hat{f}(y_i)}{\hat{g}(y_i)(1-\hat{F}(y_i))} y_i \tag{5}$$

3.2 Method 2: Koul-Susarla-van Ryzin type Transformation

The unbiased estimator for $E(T)$ is $E(T|T > Y)$ when $\delta = 0$. For proof, assume $Y^{(2)} = \delta Y + (1-\delta)\ E(T|T > Y)$. The mean of $Y^{(2)}$ is as follows by (1).

$$\begin{aligned}
E(Y^{(2)}) &= \int_0^\infty u(1-G(u))dF(u) + \int_0^\infty \{\int_u^\infty \frac{sdF(s)}{1-F(u)}\}(1-F(u))dG(u) \\
&= \int_0^\infty u(1-G(u))dF(u) + \int_0^\infty \{\int_u^\infty sdF(s)\}\ dG(u) \\
&= \int_0^\infty u(1-G(u))dF(u) + \int_0^\infty \{\int_0^s dG(u)\}\ sdF(s) \\
&= \int_0^\infty u(1-G(u))dF(u) + \int_0^\infty G(s)\ sdF(s) \\
&= \int_0^\infty udF(u) \\
&= E(T)
\end{aligned}$$

However, $E(T|T > Y)$ is unknown, so we approximately estimate it.

$$E(T|T > Y) = \int_y^\infty \frac{tdF(t)}{1 - F(y)}$$

This is replaced by using uncensored data after the time point and the Kaplan-Meier estimator of the survival function. In this case, s_k is the jump size in the Kaplan-Meier estimator and $\hat{F}(y)$ is the kernel distribution function estimate by (4). And assume that y_n is uncensored data.

$$\hat{E}(T_i|T_i > Y_i) = \sum_{k:y_k > y_i} s_k y_k / \{1 - \hat{F}(y_i)\}$$

The second method to transform of censored data to uncensored data is as follows.

$$\hat{y}_i^{(2)} = \delta_i y_i + (1 - \delta_i) \sum_{k:y_k > y_i} s_k y_k / \{1 - \hat{F}(y_i)\} \tag{6}$$

3.3 Method 3: Mean Imputation Method

As stated in method 2, $E(T|T > Y)$ is the unbiased estimator of $E(T)$ when $\delta = 0$ and $E(T|T > Y)$ depends on the unknown density function. To avoid this difficulty, we propose to estimate $E(T|T > Y)$ by using the mean imputation method. To be specific, we replace y with $\delta = 0$ by the mean of uncensored observations which are larger than y as follows. And assume that y_n is uncensored data.

$$\hat{E}(T_i|T_i > Y_i) = \sum_{j=i+1}^{n} \delta_j y_j / \sum_{j=i+1}^{n} \delta_j$$

The third method to transform of censored data to uncensored data is as follows.

$$\hat{y}_i^{(3)} = \delta_i y_i + (1 - \delta_i) \sum_{j=i+1}^{n} \delta_j y_j / \sum_{j=i+1}^{n} \delta_j \tag{7}$$

CHAPTER 4. NUMERICAL ANALYSIS

To compare finite sample performance, we conducted numerical analysis by transforming censored data to uncensored data by method 1, method 2, and method 3. The data for simulation are assumed to be $T \sim \varepsilon(1), C \sim \varepsilon(\lambda)$, where λ is 1, 3/7, and 1/9 for 50%, 30%, and 10% censoring percentage, respectively. In addition, each data was generated with different data sizes of n = 30, 50, and 100, and this was repeated 1000 times. As metioned, the randomly generated uncomplete data were transformed according to method 1, method 2, and method 3 by (5), (7), and (8), respectively. The performance was confirmed from each point of view with three measures: MISE, MSE and, c-index. It was also compared with the Kaplan-Meier estimator.

4.1 Comparison between Estimators of Distribution Function

The mean integrated squared error is a measure comparing the estimated cumulative density function $\hat{F}$ with the true cumulative density function F. That is, $\text{MISE}(F_n) = \int E(F_n - F)^2 dx$ and based on r replications, it is estimated as follows.

$$\begin{aligned}
\widehat{MISE(F_n)} &= \frac{1}{r}\sum_{j=1}^{r}\sum_{i=1}^{I}(F_{nj}(x_i) - F(x_i))^2\Delta_n \\
&= \frac{1}{r}\sum_{j=1}^{r}\sum_{i=1}^{I}(F_{nj}(x_i) - \bar{F}_n(x_i))^2\Delta_n + \sum_{i=1}^{I}(\bar{F}_n(x_i) - F(x_i))^2\Delta_n \\
&= \widehat{IV} + \widehat{ISB},
\end{aligned}$$

where $\widehat{IV}$ denotes integrated variance and $\widehat{ISB}$ denotes integrated squared bias. In this paper, we estimated the smoothed cumulative density function from transformed data by each method through *kcde* package in R and compared with the true cumulative density function by dividing it into 0.0001 sizes from 0 to 5.

Table 1. The MISE (IV + ISB) of four estimators of the distribution function with three censoring percentages (10%, 30%, and 50%) when n=30.

Estimator	Censoring	MISE (IV + ISB)
Kaplan-Meier Estimator	10%	0.013 (0.011 + 0.002)
	30%	0.032 (0.009 + 0.023)
	50%	0.083 (0.007 + 0.076)
Buckley-James type Transformation	10%	0.014 (0.013 + 0.001)
	30%	0.031 (0.012 + 0.019)
	50%	0.102 (0.010 + 0.092)
Koul-Susarla-van Ryzin type Transformation	10%	0.015 (0.013 + 0.002)
	30%	0.033 (0.015 + 0.018)
	50%	0.079 (0.028 + 0.051)
Mean Imputation Method	10%	0.015 (0.014 + 0.001)
	30%	0.021 (0.017 + 0.004)
	50%	0.042 (0.019 + 0.023)

Table 2. The MISE (IV + ISB) of four estimators of the distribution function with three censoring percentages (10%, 30%, and 50%) when n=50.

Estimator	Censoring	MISE (IV + ISB)
Kaplan-Meier Estimator	10%	0.009 (0.007 + 0.002)
	30%	0.029 (0.005 + 0.024)
	50%	0.085 (0.004 + 0.081)
Buckley-James type Transformation	10%	0.010 (0.009 + 0.001)
	30%	0.027 (0.007 + 0.020)
	50%	0.100 (0.006 + 0.094)
Koul-Susarla-van Ryzin type Transformation	10%	0.011 (0.008 + 0.003)
	30%	0.031 (0.010 + 0.021)
	50%	0.077 (0.021 + 0.056)
Mean Imputation Method	10%	0.010 (0.009 + 0.001)
	30%	0.014 (0.010 + 0.004)
	50%	0.035 (0.011 + 0.024)

Table 3. The MISE (IV + ISB) of four estimators of the distribution function with three censoring percentages (10%, 30%, and 50%) when n=100.

Estimator	Censoring	MISE (IV + ISB)
Kaplan-Meier Estimator	10%	0.007 (0.004 + 0.003)
	30%	0.029 (0.003 + 0.026)
	50%	0.084 (0.002 + 0.082)
Buckley-James type Transformation	10%	0.006 (0.005 + 0.001)
	30%	0.023 (0.004 + 0.019)
	50%	0.094 (0.003 + 0.091)
Koul-Susarla-van Ryzin type Transformation	10%	0.008 (0.004 + 0.004)
	30%	0.031 (0.006 + 0.025)
	50%	0.072 (0.017 + 0.055)
Mean Imputation Method	10%	0.006 (0.005 + 0.001)
	30%	0.009 (0.005 + 0.004)
	50%	0.029 (0.006 + 0.023)

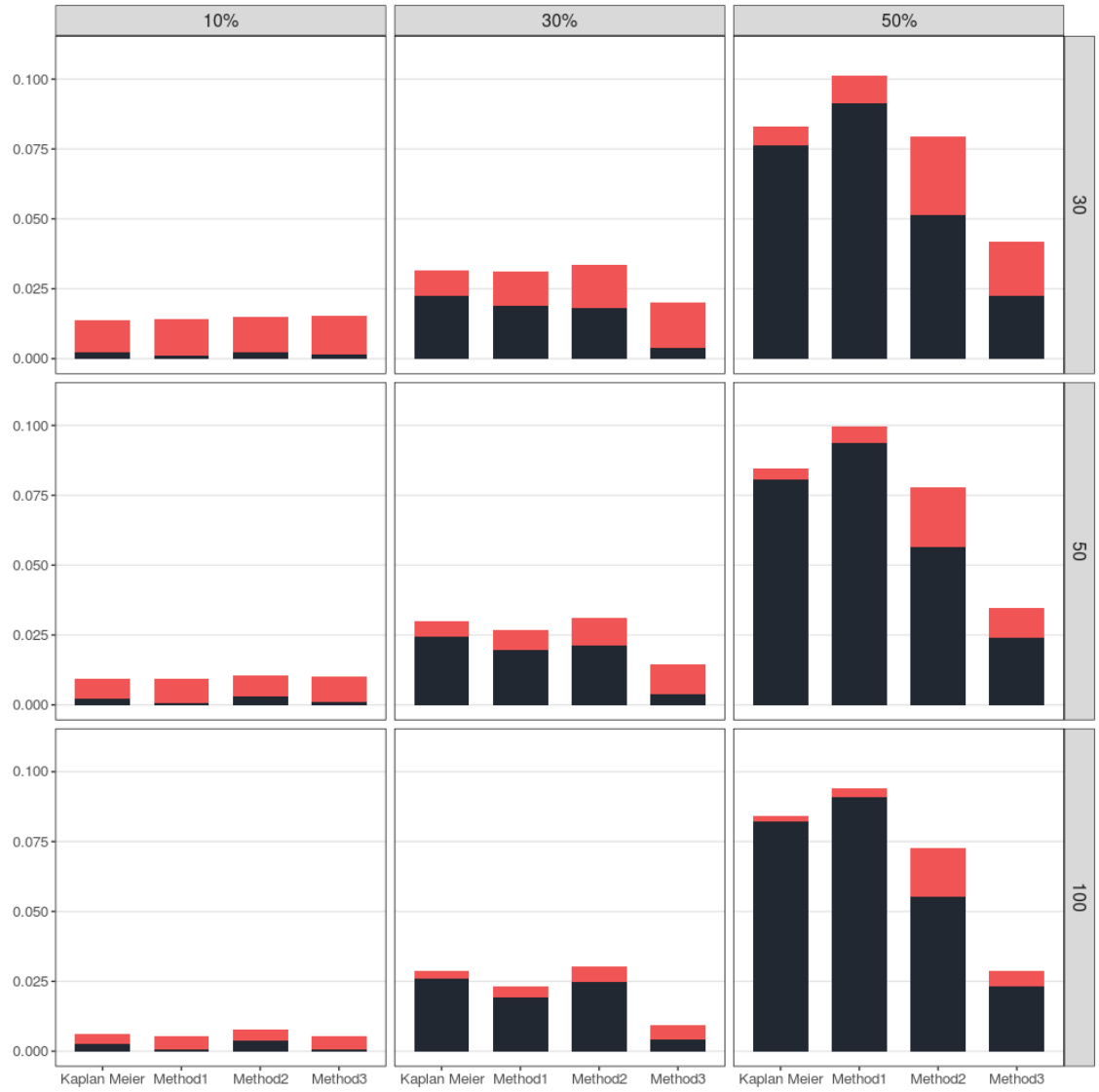

Figure 1. The MISE (IV + ISB) of four estimators of the distribution function with three censoring percentages (10%, 30%, and 50%) when n=30, 50, 100. The above and the below of each histogram denote IV and ISB, respectively.

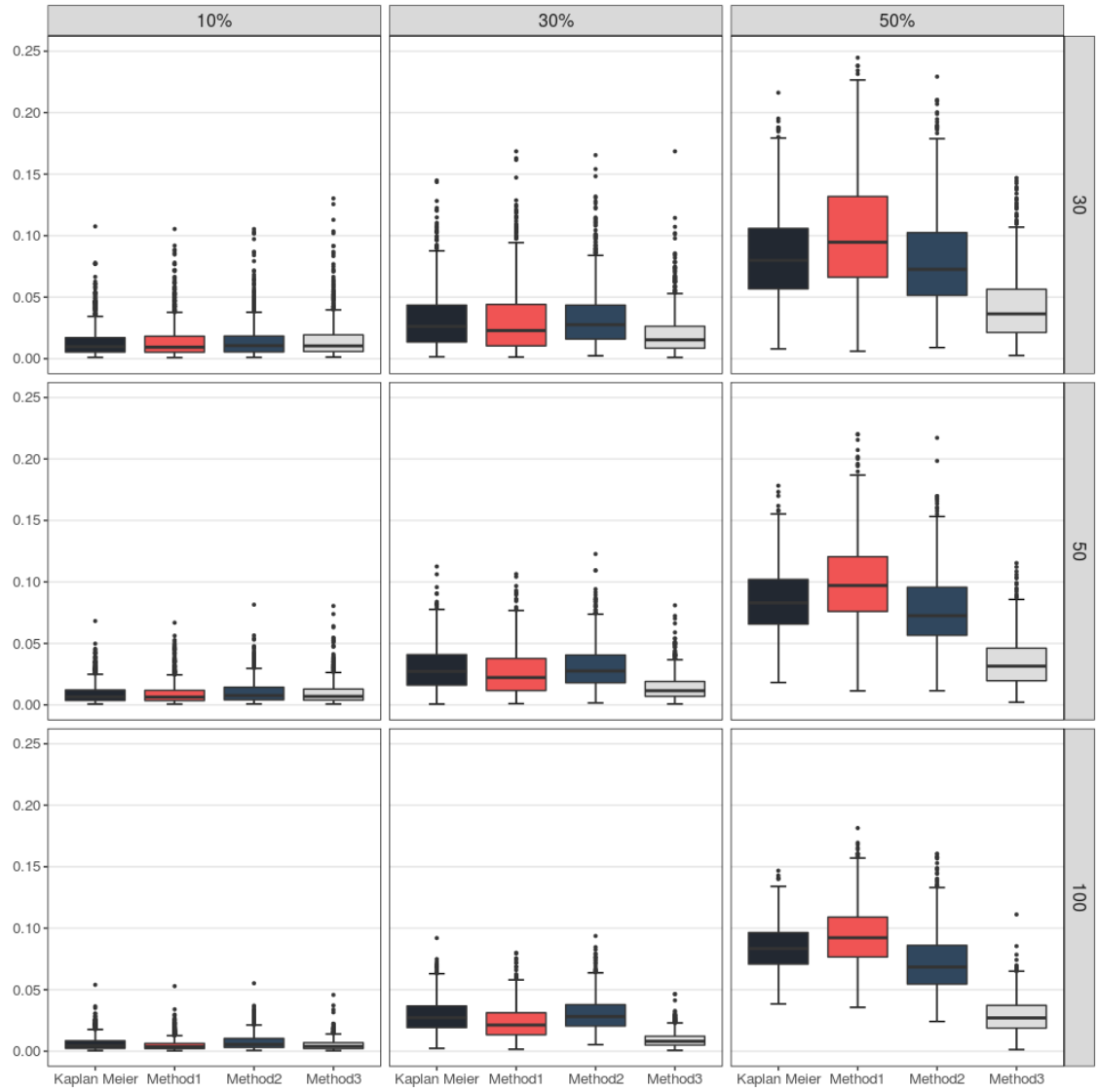

Figure 2. The Boxplot of MISE of four estimators of the distribution function with three censoring percentages (10%, 30%, and 50%) when n=30, 50, 100 for 1000 iterations.

Table 1, Table 2, and Table 3 demonstrate that the experiment was well conducted because the MISE become smaller as the number of data becomes larger and the censoring becomes smaller. The result of experiment first represents that the MISE of the first method is larger than the MISE of the Kaplan-Meier estimator when the censoring percentage is large. In particular, the ISB of the first method is large. The reason of this seems that the bias increases as the kernel density estimator result. However, if the kernel density estimator is estimated more accurately when the censoring percentage is small, it is better performance than the Kaplan-Meier estimator. The MISE of the second method shows similar to the Kaplan-Meier estimator. In the case of the third method, the result represents the best performance when the censoring percentage is large. For that, it is significant as a method of transforming censored data.

4.2 Comparison between Estimators of the Median Survival Time

θ is called median survival time when $\theta = S^{-1}(0.5)$. In other words, when the survival function is 0.5, the time is the parameter of interest. The mean squared error is used to check the performance of the estimate for the median survival time. Because $MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$, it is estimated as follows.

$$\begin{aligned}
\widehat{MSE(\theta_{0.5})} &= \frac{1}{r}\sum_{j=1}^{r}(\hat{\theta}_{0.5,j} - \theta_{0.5})^2 \\
&= \frac{1}{r}\sum_{j=1}^{r}(\hat{\theta}_{0.5,j} - \bar{\theta}_{0.5})^2 + (\bar{\theta}_{0.5} - \theta_{0.5})^2 \\
&= \widehat{V} + \widehat{SB},
\end{aligned}$$

where $\widehat{V}$ denotes variance and $\widehat{SB}$ denotes squared bias. In this paper, the median survival time estimated by the each method for n=30, 50, 100 and λ=1, 3/7, 1/9. After that, the MSE was calculated and compared each other.

Table 4. The MSE (V + SB) of four estimators of the distribution function with three censoring percentages (10%, 30%, and 50%) when n=30.

Estimator	Censoring	MSE (V + SB)
Kaplan-Meier Estimator	10%	0.023 (0.022 + 0.001)
	30%	0.041 (0.014 + 0.027)
	50%	0.105 (0.007 + 0.098)
Buckley-James type Transformation	10%	0.029 (0.027 + 0.002)
	30%	0.034 (0.019 + 0.015)
	50%	0.122 (0.008 + 0.114)
Koul-Susarla-van Ryzin type Transformation	10%	0.024 (0.024 + 0.000)
	30%	0.031 (0.029 + 0.002)
	50%	0.065 (0.051 + 0.014)
Mean Imputation Method	10%	0.036 (0.030 + 0.006)
	30%	0.037 (0.031 + 0.006)
	50%	0.028 (0.027 + 0.001)

Table 5. The MSE (V + SB) of four estimators of the distribution function with three censoring percentages (10%, 30%, and 50%) when n=50.

Estimator	Censoring	MSE (V + SB)
Kaplan-Meier Estimator	10%	0.014 (0.013 + 0.001)
	30%	0.041 (0.008 + 0.033)
	50%	0.122 (0.004 + 0.108)
Buckley-James type Transformation	10%	0.018 (0.017 + 0.001)
	30%	0.027 (0.010 + 0.017)
	50%	0.125 (0.005 + 0.120)
Koul-Susarla-van Ryzin type Transformation	10%	0.015 (0.015 + 0.000)
	30%	0.020 (0.019 + 0.001)
	50%	0.046 (0.037 + 0.009)
Mean Imputation Method	10%	0.025 (0.020 + 0.005)
	30%	0.026 (0.020 + 0.006)
	50%	0.015 (0.014 + 0.001)

Table 6. The MSE (V + SB) of four estimators of the distribution function with three censoring percentages (10%, 30%, and 50%) when n=100.

Estimator	Censoring	MSE (V + SB)
Kaplan-Meier Estimator	10%	0.010 (0.007 + 0.003)
	30%	0.044 (0.004 + 0.040)
	50%	0.117 (0.002 + 0.115)
Buckley-James type Transformation	10%	0.009 (0.009 + 0.000)
	30%	0.024 (0.006 + 0.018)
	50%	0.124 (0.003 + 0.121)
Koul-Susarla-van Ryzin type Transformation	10%	0.008 (0.008 + 0.000)
	30%	0.012 (0.011 + 0.001)
	50%	0.029 (0.027 + 0.002)
Mean Imputation Method	10%	0.012 (0.010 + 0.002)
	30%	0.015 (0.010 + 0.005)
	50%	0.009 (0.008 + 0.001)

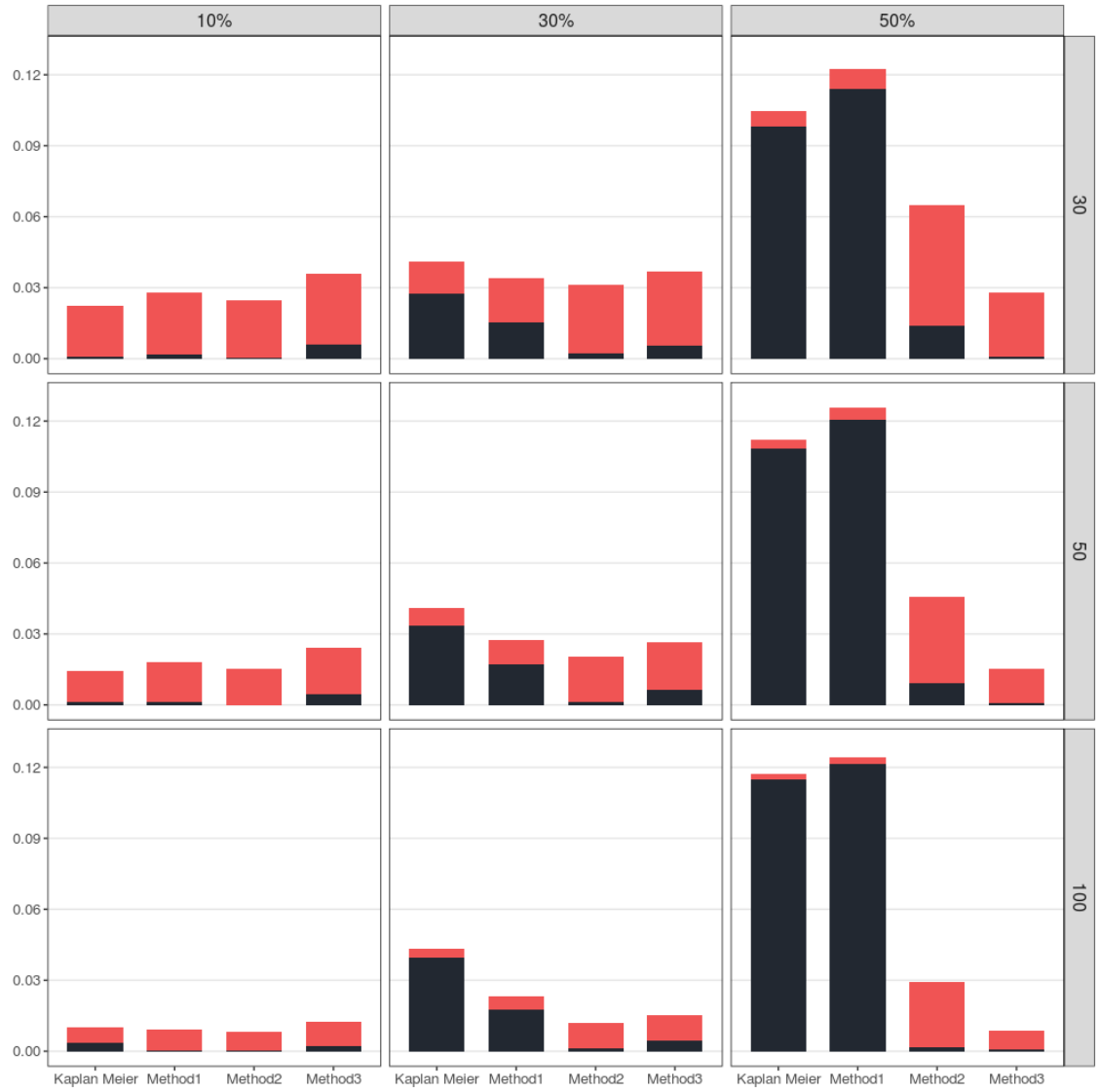

Figure 3. The MSE (V + SB) of four estimators of the median survival time with three censoring percentages (10%, 30%, and 50%) when n=30, 50, 100. The above and the below of each histogram denote V and SB, respectively.

As a result of the simulation, it shows that the MSE is also similar to the MISE. First, the SB of the median survival time estimated by the first method is large like MISE. However, the MSE estimated by the data transformed by the first method and the second method is small when the censoring percentage is small. In the case of the third method, the median survival time is better estimated than the Kaplan-Meier estimator because the MSE is the smallest when the censoring percentage is large.

4.3 Comparison based on the Concordance Index

The concordance index (Harrell et al., 1982) is one of the main measures of survival analysis. Suppose we have n observations $y_1, \cdots, y_n$ and n fitted values $\hat{y}_1, \cdots, \hat{y}_n$. The c-index is a ratio of concordant pairs among all possible pairs, i.e. c-index is defined as follows.

$$c = \frac{\sum_{i=1}\sum_{j\neq i} I(y_i > y_j) I(\hat{y}_i > \hat{y}_j)\delta_j}{\sum_{i=1}\sum_{j\neq i} I(y_i > y_j)\delta_j}$$

A pair of observations i and j is called concordant if $I(y_i > y_j) = I(\hat{y}_i > \hat{y}_j)$, and called discordant if $I(y_i > y_j) = I(\hat{y}_i < \hat{y}_j)$.

Table 7 and Figure 4 are confirmed that the result is similar to MISE and MSE. In particular, the third method has the largest the c-index as a whole. That is, the data transformed the censored data by the third method most corresponds to the true observation. In addition, the first method to transform is better performance than the censored data when the censoring percentage is small.

Table 7. The c-index with three censoring percentages (10%, 30%, and 50%) when n=30, 50, and 100.

Estimator	Censoring	n		
		30	50	100
Censored Data	10%	0.951	0.949	0.950
	30%	0.851	0.849	0.850
	50%	0.750	0.748	0.749
Buckley-James type Transformation	10%	0.954	0.955	0.957
	30%	0.855	0.860	0.868
	50%	0.724	0.729	0.736
Koul-Susarla-van Ryzin type Transformation	10%	0.943	0.941	0.938
	30%	0.836	0.835	0.831
	50%	0.732	0.731	0.728
Mean Imputation Method	10%	0.973	0.974	0.973
	30%	0.916	0.918	0.917
	50%	0.844	0.846	0.849

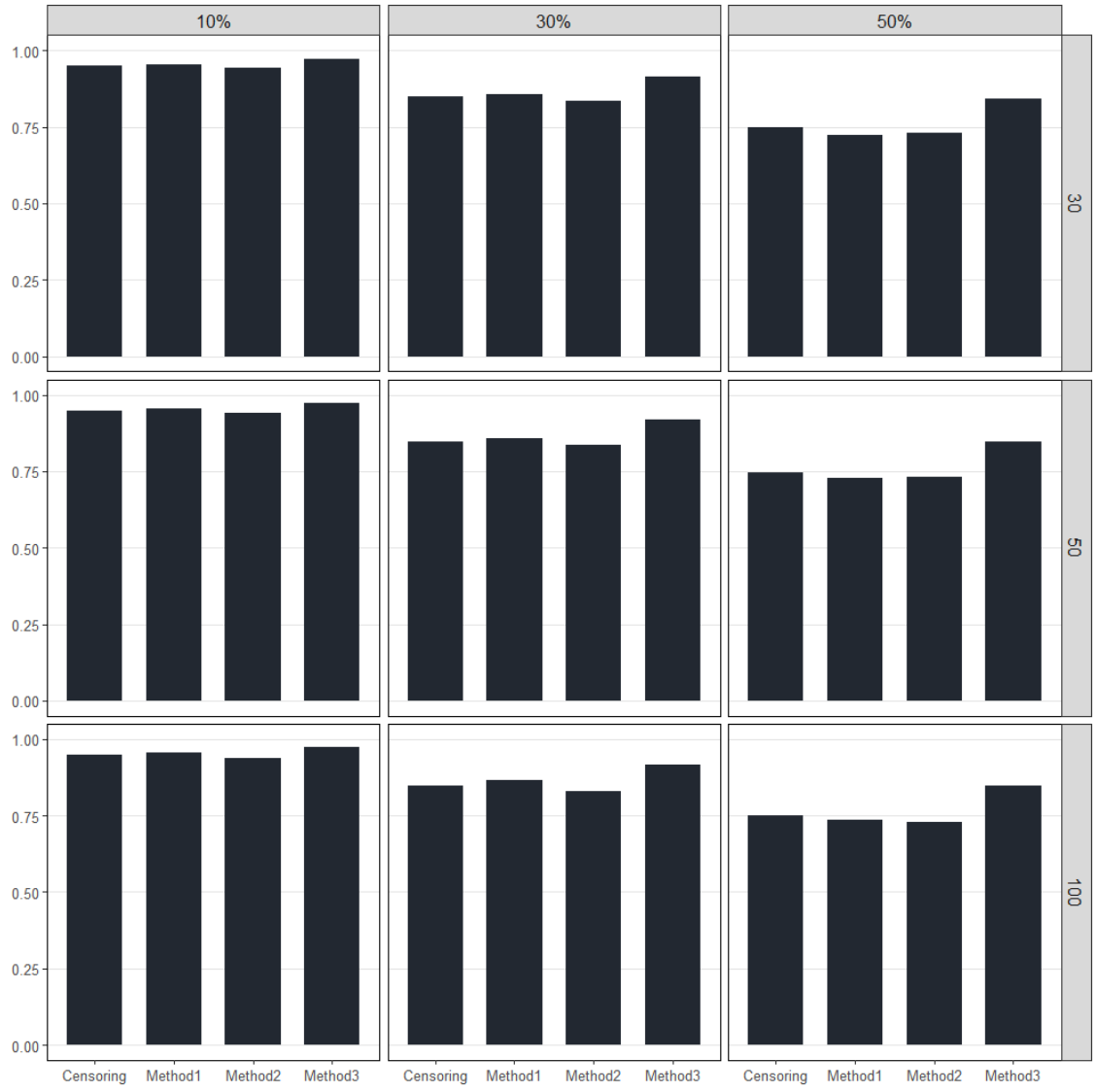

Figure 4. The c-index with three censoring percentages (10%, 30%, and 50%) when n=30, 50, 100.

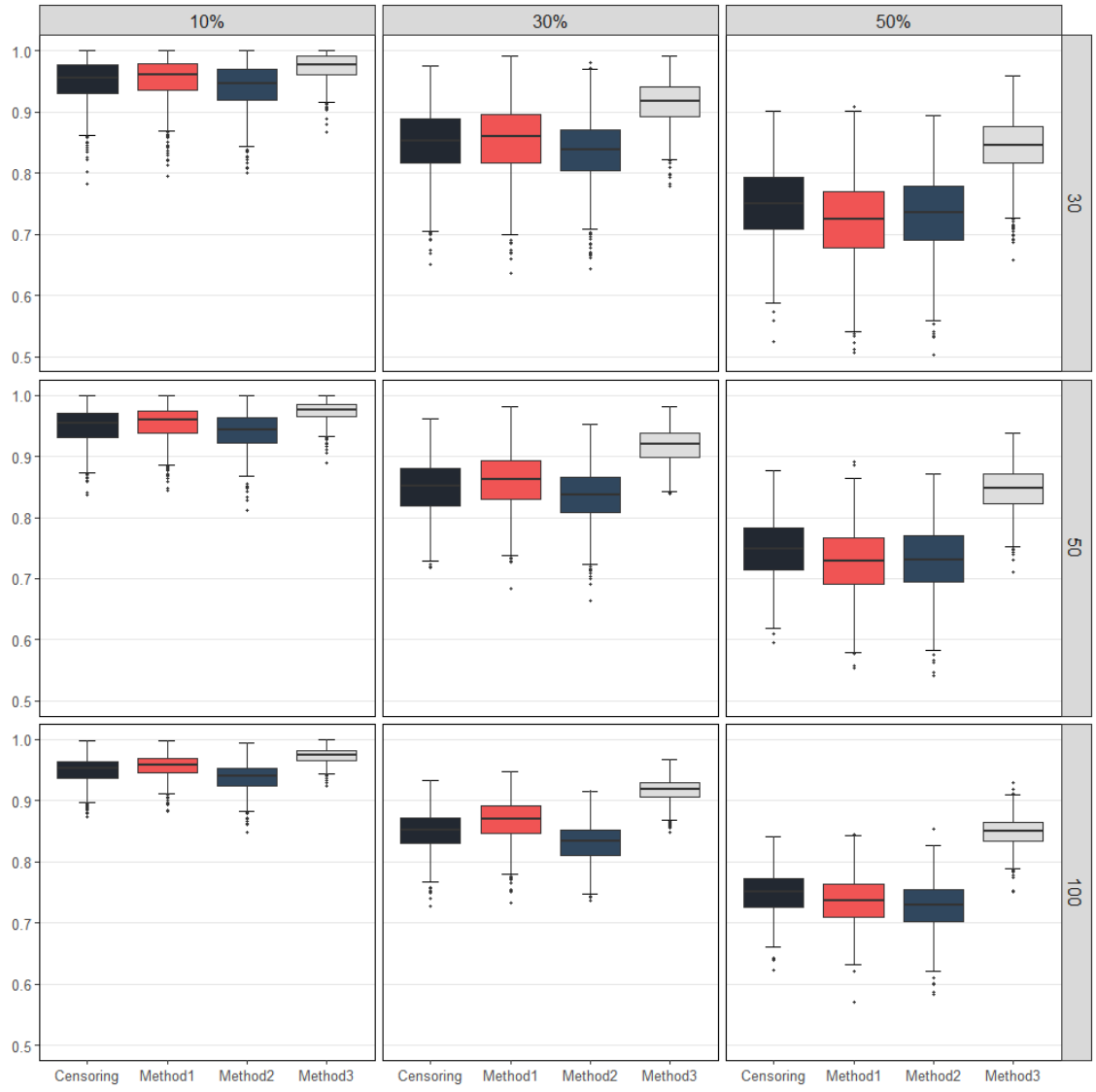

Figure 5. The Boxplot of c-index with three censoring percentages (10%, 30%, and 50%) when n=30, 50, 100 for 1000 iterations.

CHAPTER 5. CONCLUSION

As stated in the introduction, the research was conducted to solve the essential problem of survival analysis by transforming of censored data to uncensored data. To the end, we suggested the three unbiased transformation methods.

The first method is the Buckley-James type transformation to the unbaised estimator of the mean of survival time using the kernel density estimator and the kernel distribution function. The second method is the Koul-Susarla-van Ryzin type transformation to the approximate unbiased estimate. Finally, the third method is the mean imputation method using the mean of uncensored data after the time point. To check the performance of the three methods, we compared the distribution function, the median survival time, and the concordance.

As a result of the numerical analysis, there is a difference in performance according to the censoring percentage. In the case of small censoring percentage, the first method and the second method have good performance because the estimation of the kernel density estimate is well estimated. In the case of large censoring percentage, the mean imputation method has good performance overall. In particular, the mean imputation method shows better performance than the Kaplan-Meier estimator. The results of this study represents that it is significant to transform of censored data to uncensored data.

In the next research, we might extend the method of transforming censored data proposed in this paper to the multivariate regression problem in censored survival data.

References

Blum, J. and Susarla, V. (1980). Maximal deviation theory of density and failure rate function estimates based on censored data. *Multivariate analysis*, **5**, 213–222.

Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrika*, **66**, 429–436.

Chacón, J. E. and Duong, T. (2018). *Multivariate Kernel Smoothing and Its Applications*. Chapman; Hall/CRC.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, **34**, 187–202.

Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications*, **14**, 153–158.

Harrell Jr, F. E., Lee, K. L. and Mark, D. B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, **15**, 361–387.

Kalbfleisch, J. D. and Prentice, R. L. (2011). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457–481.

Koul, H., Susarla, V. and Van Ryzin, J. (1981). Regression analysis with randomly right-censored data. *The Annals of Statistics*, **9**, 1276–1288.

Lawless, J. F. (2011). *Statistical Models and Methods for Lifetime Data*. John Wiley & Sons.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother. Rep.*, **50**, 163–170.

Nadaraya, E. A. (1964). Some new estimates for distribution functions. *Theory of Probability & Its Applications*, **9**, 497–500.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, **33**, 1065–1076.

Rosenblatt, M. (1956). Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics*, **27**, 832–837.

Wand, M. P. and Jones, M. C. (1994). *Kernel Smoothing*. CRC press.

Watson, G. S. and Leadbetter, M. (1964). Hazard analysis II. *Sankhyā: The Indian Journal of Statistics, Series A*, **26**, 101–116.

중도 절단 자료의 변환

이은주

부산대학교 대학원 통계학과

요약

생존 시간 자료에서 중도 절단된 관측치는 불가피하며, 중도 절단으로 인한 불완전 자료는 두 가지 측면에서 바람직하지 않다. 첫 번째, 일부 정보가 중도 절단으로 인해 손실되고 두 번째, 중도 절단된 자료를 분석하기 위한 통계 도구가 제한적이다. 그러므로 중도 절단된 자료를 완전한 자료로 변환하는 것이 타당하다. 본 논문에서는 중도 절단된 생존 시간을 완전한 자료로 불편(unbiased) 변환하는 세 가지 종류를 고려한다; 첫 번째 방법은 Buckley-James 종류의 변환 방법이고, 두 번째 방법은 Koul-Susarla-van Ryzin 종류의 변환 방법이며, 세 번째 방법은 평균 대체 방법이다. 세 가지 방법과 카플란-마이어 추정치의 수치적 성능을 비교하기 위해 분포 함수 추정치의 적분 평균 오차 제곱 (MISE), 중간 생존 시간 추정치의 평균 오차 제곱 (MSE), 그리고 관측치와 예측치 사이의 일치 추정치인 c-index를 평가하였다. 결론적으로, 평균 대체 방법이 다양한 중도 절단율에서 가장 좋은 결과를 보였다. 향후 제안된 방법은 중도 절단된 생존 시간의 회귀 분석으로 확장하고자 한다.