Unbiased Transformation of Censored Survival Data

Eunju Lee

Department of Statistics
Pusan National University

09 December, 2022

Table of contents

Survival Analysis

- What is survival analysis?
  - Survival analysis is an analysis of the time to event.
- Limitation of censoring
  - First, some information is lost due to censoring.
  - Secondly, statistical tools for analyzing censored data are restrictive.

Basic Functions

- Notation
  - $T_1, \cdots, T_n$: *iid* true survival times with cdf $F(\cdot)$ and pdf $f(\cdot)$
  - $C_1, \cdots, C_n$: *iid* censored times with cdf $G(\cdot)$ and pdf $g(\cdot)$

- We observe ordered data $(Y_i, \delta_i), i = 1, \cdots, n$, where $\delta_i = I(T_i \leq C_i)$ is censoring indicator and (T_i, C_i) are independent.

- Probability density function of Y

$$P(Y_i, \delta_i) = \begin{cases} P(Y_i = T_i, C_i > Y_i) = f_\theta(Y_i)(1 - G_\gamma(Y_i)) \text{ if } \delta_i = 1 \\ P(Y_i = C_i, T_i > Y_i) = g_\gamma(Y_i)(1 - F_\theta(Y_i)) \text{ if } \delta_i = 0 \end{cases}$$

Kaplan-Meier Estimator

- Notation
  - n_i : number of subjects at risk at y_i^-, which is $n - i + 1$
  - p_i : $P(T > y_i | T > y_{i-1})$ is estimated $n_{i+1}/n_i = 1 - 1/(n - i + 1)$ when $\delta_i = 1$

- Kaplan-Meier Estimator

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

Kernel Estimator

- Kernel density estimate and Kernel distribution function estimate

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

where K : kerenel function, W : cumulative kernel function

s_i : jump size at y_i in Kaplan-Meier estimator

h : smoothing parameter, bandwidth

- Kernel Function Type
  - Uniform : $K(t) = \frac{1}{2} \cdot I(|t| \leq 1)$
  - Gaussian : $K(t) = \frac{1}{\sqrt{2\pi}} \cdot \exp^{-t^2/2}$
  - Epanechnikov : $K(t) = \frac{3}{4} \cdot (1 - t^2) I(|t| < 1)$

- The Method of Choosing h
  - Generalized cross validation
  - Likelihood cross validation

Buckley-James type Transformation

- $\hat{y}_i^{(1)} = \delta_i y_i + (1 - \delta_i) \frac{\hat{f}(y_i)}{\hat{g}(y_i)(1-\hat{F}(y_i))} y_i$
  - Motivation : $E(Y^{(1)}) = E(T), \quad Y^{(1)} = \delta Y + (1 - \delta) \frac{f(Y)}{g(Y)(1-F(Y))} Y$
  - Estimate $\hat{f}(y)$, $\hat{F}(y)$, and $\hat{g}(y)$ by kernel estimator using Epanechnikov kernel function.
  - Calculate bandwidth h using likelihood cross validation.

Koul-Susarla-van Ryzin type Transformation

- $\hat{y}_i^{(2)} = \delta_i y_i + (1 - \delta_i) \sum_{k: y_k > y_i} s_k y_k / \{1 - \hat{F}(y_i)\}$
    - Motivation : $E(Y^{(2)}) = E(T), \quad Y^{(2)} = \delta Y + (1 - \delta)\ E(T|T > Y)$
    - Estimate $\hat{E}(T|T > y)$ as $\sum_{k: y_k > y_i} s_k y_k / \{1 - \hat{F}(y_i)\}$.
    - Estimate $F(y)$ by kernel estimator like Method 1.

Mean Imputation Method

- $\hat{\boldsymbol{y}}_i^{(3)} = \boldsymbol{\delta_i y_i + (1 - \delta_i) \sum_{j=i+1}^{n} \delta_j y_j / \sum_{j=i+1}^{n} \delta_j}$
  - Motivation : $E(Y^{(3)}) = E(T), \quad Y^{(3)} = \delta Y + (1 - \delta)\ E(T|T > Y)$
  - Estimate $\hat{E}(T|T > y)$ as $\sum_{j=i+1}^{n} \delta_j y_j / \sum_{j=i+1}^{n} \delta_j$.
  - Assume that y_n is uncensored data.

Numerical Analysis

- Data
  - $T \sim \varepsilon(1), C \sim \varepsilon(\lambda)$
  - λ : 1, 3/7, and 1/9 for 50%, 30% and 10% censoring percentage
  - $n =$ 30, 50, and 100
- Comparison
  - MISE to compare between estimators of distribution function
  - MSE to compare between estimators of the median survival time
  - ***c*-index** to compare the concordance

Numerical Analysis

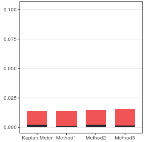
- ***c*-index**
  - $c = \frac{\sum\limits_{i=1}\sum\limits_{j \neq i} I(y_i > y_j) I(\hat{y}_i > \hat{y}_j)\delta_j}{\sum\limits_{i=1}\sum\limits_{j \neq i} I(y_i > y_j)\delta_j}$
  - The ratio of concordant pairs pairs among all possible pairs
    - A pair of observations i and j is called concordant if $I(y_i > y_j) = I(\hat{y}_i > \hat{y}_j)$
    - A pair of observations i and j is called discordant if $I(y_i > y_j) = I(\hat{y}_i < \hat{y}_j)$

Comparison of Distribution Function

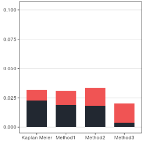
Table 1. The MISE of four estimators of the distribution function with three censoring percentages (10%, 30%, and 50%) when n=30, 50, and 100.

Estimator	n	MISE 10%	MISE 30%	MISE 50%
Kaplan-Meier	30	0.013	0.032	0.083
	50	0.009	0.029	0.085
	100	0.007	0.029	0.084
Method 1	30	0.014	0.031	0.102
	50	0.010	0.027	0.100
	100	0.006	0.023	0.094
Method 2	30	0.015	0.033	0.079
	50	0.011	0.031	0.077
	100	0.008	0.031	0.072
Method 3	30	0.015	0.021	0.042
	50	0.010	0.014	0.035
	100	0.006	0.009	0.029

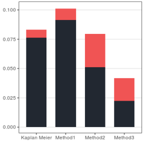
Comparison of Distribution Function



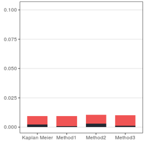
(a) n=30, 10%



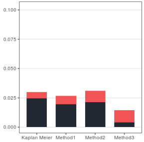
(b) n=30, 30%



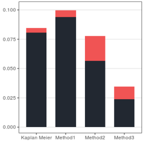
(c) n=30, 50%



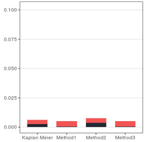
(d) n=50, 10%



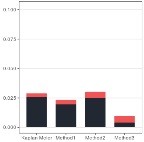
(e) n=50, 30%



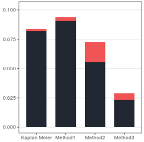
(f) n=50, 50%



(g) n=100, 10%



(h) n=100, 30%



(i) n=100, 50%

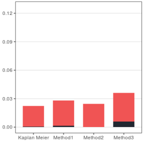
Figure 1. The MISE of four estimators of the distribution function with three censoring percentages (10%, 30%, and 50%) when n=30, 50, and 100. The above and the below of each historgram denote IV and ISB, respectively.

Comparison of the Median Survival Time

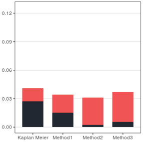
Table 2. The MSE (V + SB) of four estimators of the median survival time with three censoring percentages (10%, 30%, and 50%) when n=30, 50, and 100.

Estimator	n	MSE 10%	MSE 30%	MSE 50%
Kaplan-Meier	30	0.023	0.041	0.105
	50	0.014	0.041	0.122
	100	0.010	0.044	0.117
Method 1	30	0.029	0.034	0.122
	50	0.018	0.027	0.125
	100	0.009	0.024	0.124
Method 2	30	0.024	0.031	0.065
	50	0.015	0.020	0.046
	100	0.008	0.012	0.029
Method 3	30	0.036	0.037	0.028
	50	0.025	0.026	0.015
	100	0.012	0.015	0.009

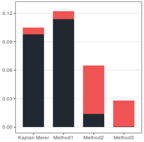
Comparison of the Median Survival Time



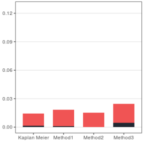
(a) n=30, 10%



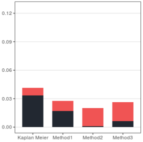
(b) n=30, 30%



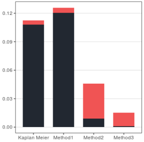
(c) n=30, 50%



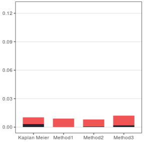
(d) n=50, 10%



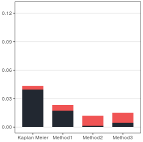
(e) n=50, 30%



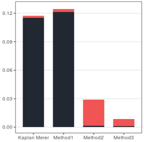
(f) n=50, 50%



(g) n=100, 10%



(h) n=100, 30%



(i) n=100, 50%

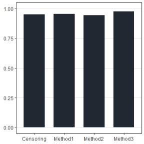
Figure 2. The MSE (V + SB) of four estimators of the median survival time with three censoring percentages (10%, 30%, and 50%) when n=30, 50, and 100. The above and the below of each historgram denote V and SB, respectively.

Comparison of the Concordance

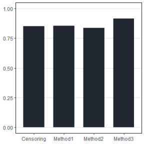
Table 3. The c-index of the censored data and the three transformed data with three censoring percentages (10%, 30%, and 50%) when n=30, 50, and 100.

Estimator	n	c-index 10%	c-index 30%	c-index 50%
Censoring	30	0.951	0.851	0.750
	50	0.949	0.849	0.748
	100	0.950	0.850	0.749
Method 1	30	0.954	0.855	0.724
	50	0.955	0.860	0.729
	100	0.957	0.868	0.736
Method 2	30	0.943	0.836	0.732
	50	0.941	0.835	0.731
	100	0.938	0.831	0.728
Method 3	30	0.973	0.916	0.844
	50	0.974	0.918	0.846
	100	0.973	0.917	0.849

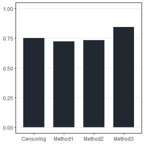
Comparison of the Concordance



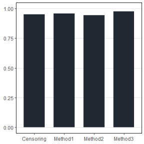
(a) n=30, 10%



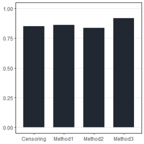
(b) n=30, 30%



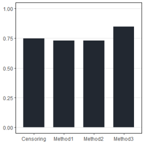
(c) n=30, 50%



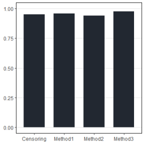
(d) n=50, 10%



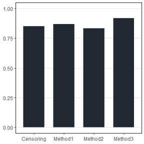
(e) n=50, 30%



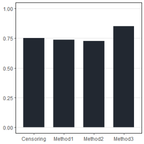
(f) n=50, 50%



(g) n=100, 10%



(h) n=100, 30%



(i) n=100, 50%

Figure 3. The c-index of the censored data and the three data transformed by the proposed methods with three censoring percentages (10%, 30%, and 50%) when n=30, 50, 100.

Conclusion

- In the case of small censoring percentage, the Buckley-James type method and the Koul-Susarla- Van Ryzin type 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 and shows especially better performance than the Kaplan-Meier estimator.
- As future research, we might extend the method of transforming censored data proposed in this paper to the multivariate regression problem in censored survival data.

Thank you