The Korean Journal of Applied Statistics (2011) DOI: http://dx.doi.org/10.5351/KJAS.2011.24.5.871
24*(5), 871–881*

Logistic Regression Method in Interval-Censored Data

Eunyoung Yun[1] · Jinmi Kim[2] · Choongrak Kim[3]

[1]Department of Statistics, Pusan National University
[2]Department of Statistics, Pusan National University
[3]Department of Statistics, Pusan National University

(Received July 2011; accepted August 2011)

Abstract

In this paper we propose a logistic regression method to estimate the survival function and the median survival time in interval-censored data. The proposed method is motivated by the data augmentation technique with no sacrifice in augmenting data. In addition, we develop a cross validation criterion to determine the size of data augmentation. We compare the proposed estimator with other existing methods such as the parametric method, the single point imputation method, and the nonparametric maximum likelihood estimator through extensive numerical studies to show that the proposed estimator performs better than others in the sense of the mean squared error. An illustrative example based on a real data set is given.

Keywords: Cross validation, imputation method, Kaplan-Meier estimator, median survival time, nonparametric maximum likelihood estimation, survival function.

1. Introduction

Interval-censored data are incomplete data, where the exact time of the event is not available, however, the event is known to occur between two defined times. Interval censoring is a general concept of censoring that includes right censoring and left censoring as special cases. Interval-censored data are more difficult to deal with than the right-censored data because of the complexity and special structure of the interval censoring. Thus, existing methods in right-censored data cannot be directly applicable to the interval-censored data. For example, the Kaplan-Meier type estimator for the interval censored data is not analytically given. Among studies in interval-censored data, Peto (1973) derived survival curve for the interval-censored data, Finkelstein (1986) and Huang (1996) suggested a proportional hazards model for interval-censored data; in addition, Lindsey and Ryan (1998) provide an excellent review of this area. Recently, Fang *et al.* (2002) made nonparametric comparisons, Hudgens (2005) suggested a nonparametric maximum likelihood estimation with interval censoring and left truncation, and Lawless and Babineau (2006) proposed a simulation-based inference for

This work was supported by a National Research Foundation of Korea Grant funded by the Korean Government(2009-0071660).
[3]Corresponding author: Professor, Department of Statistics, Pusan National University, Pusan 609-735, Korea. E-mail: crkim@pusan.ac.kr

interval-censored data. An excellent book on the statistical analysis of interval-censored data is Sun (2006).

Here we consider the case II interval-censored data (a.k.a general case of interval-censored data) among four types of interval censoring (see Sun (2006) for details) because most of interval-censored data belong to the case II. As in the right-censored data, the estimation of survival function is of prime interest when we are given interval-censored data; other issues such as estimation of the median survival time, comparison of survival functions, relationship between survival times and the seemingly relevant covariates will be followed. For the parametric estimation of survival function a specific family of distributions such as Weibull is assumed a priori. There are many methods to estimate the survival function nonparametrically. Among them imputation method (Rubin, 1987) and nonparametric maximum likelihood estimator (Turnbull, 1976; Groeneboom and Wellner, 1992; Jongbloed, 1998) are often used.

In this paper we propose a method of estimating survival function and the median survival time using a logistic regression method in interval-censored data. The proposed method is motivated because the fact that the event of interest never occurred before the interval and it surely occurred after the interval. Therefore, we can augment arbitrary many samples both before and after the interval. If we regard the event before the interval failure and after the interval success, then by treating time as covariate a parametric logistic regression can be used to estimate the survival function. A cross-validation criterion to estimate the number of augmentation is proposed. Through simulation studies we compare the proposed method with imputation method and nonparametric maximum likelihood estimator in estimating the survival function and the median survival time.

2. Existing Methods

Let $T_1, T_2, \ldots, T_n$ be true survival times, then T_i is called interval-censored if instead of observing T_i exactly, only an interval $(L_i, R_i]$ is observed such that

$$T_i \in (L_i, R_i],$$

where $L_i \leq R_i$. Here, L_i and R_i indicate the left and right endpoints of the observed event time. Recall that $L_i = R_i$ means an exact observation, $R_i = \infty$ represents a right-censored observation, and $L_i = 0$ represents a left-censored observation.

Imputation method (Rubin, 1987) is originally intended for handling missing data problems, and the method can be adapted to the analysis of interval-censored data. The idea is to impute exact survival times from interval-censored data and to take advantage of many standard methods for right-censored data. For subject i, the underlying true failure time T_i is equal to a value within the observed interval $(L_i, R_i]$, $i = 1, \ldots, n$. Mean imputation is to let T_i be the middle point of the interval for a finite interval, *i.e.*, $R_i < \infty$. For intervals with $R_i = \infty$, they are regarded as right-censored observations. Taking T_i to be L_i or R_i is the left end point imputation or the right end point imputation, respectively. Then, the interval-censored data are transformed to the right-censored data, and therefore, many standard methods for right-censored data can be used. This imputation method is particularly called a single imputation method. On the other hand, multiple imputation methods (Tanner and Wong, 1987; Tanner, 1991) are also available in analyzing the interval-censored data.

For the right-censored data, the nonparametric maximum likelihood estimator(NPMLE) of a survival function is just the Kaplan-Meier estimator (Kaplan and Meier, 1958). For the interval-

censored data, the NPMLE does not have a closed form and can only be obtained numerically using iterative algorithms. There are three algorithms for computing the NPMLE of the survival function in interval-censored data. The first one is the self-consistency algorithm suggested by Turnbull (1976), and it is motivated by the self-consistency equation of Efron (1967). The second one is the iterative convex minorant(ICM) algorithm, originally introduced by Groeneboom and Wellner (1992) and later developed by Jongbloed (1998). The third one, called the EM-ICM algorithm, is a mixture of the self-consistency algorithm and the ICM algorithm, and it is suggested by Wellner and Zhan (1997). Among three algorithms ICM and EM-ICM are faster than the self-consistency algorithm, however, the self-consistency algorithm is still used quite often because it is simple and accurate, so that it.

3. Proposed Estimators

3.1. Motivation

Assume that we have n interval-censored data $(L_i, R_i]$, $i = 1, \ldots, n$, where R_i could be infinite for some i. Also, assume that there exists an interval $(0, w)$ such that

$$P(T > w) \simeq 0.$$

For example, if $T \sim \text{Exp}(1)$, then $P(T > 5) < 10^{-2}$. Since w is unknown in real data set we have to estimate it. Let d be the number of finite R_i's and let $D = \{i : R_i < \infty\}$ be the index set of finite R_i's. Also, let

$$A = \sum_{i \in D} \frac{R_i - L_i}{d}$$

be the average length of finite intervals. Here we propose to estimate w as

$$\hat{w} = A + \max\left(L_{(n)}, R_{(d)}\right), \tag{3.1}$$

where $L_{(n)}$ and $R_{(d)}$ are the largest order statistic of $L_1, \ldots, L_n$ and R_i, $i \in D$, respectively. That is to say we estimate w as the largest value among finite values of L_i and R_i plus the average length of finite intervals. This estimating scheme is based on the idea of Kim *et al.* (2003) in the right censored data.

First, we partition the interval $(0, w)$ into k subintervals with equal length w/k, *i.e.*, the j^{th} subinterval is given by

$$\left(\frac{(j-1)w}{k}, \frac{jw}{k}\right), \quad j = 1, \ldots, k.$$

The number of subintervals k must be given a priori or estimated by data, and this issue will be discussed in detail in Section 3.2.4. If

$$L_i \in \left(\frac{(l-1)w}{k}, \frac{lw}{k}\right) \tag{3.2}$$

for some $l = 1, \ldots, k$, then define, for each $i = 1, \ldots, n$,

$$X_{ij} = \frac{jw}{k}, \quad j = 1, \ldots, k$$

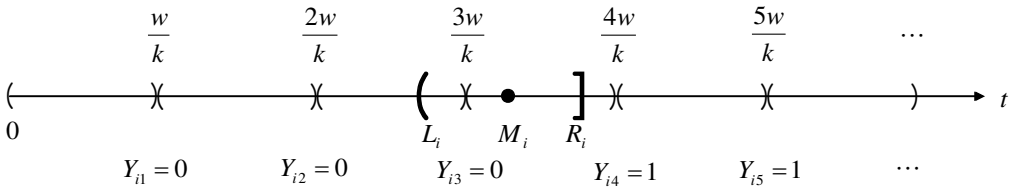

Figure 3.1. Illustration of constructing (X_{ij}, Y_{ij}), $j = 1, \ldots, k$ for each $i = 1, \ldots, n$.

and

$$Y_{ij} = 0, \quad j = 1, \ldots, l-1;$$
$$Y_{ij} = 1, \quad j = l, \ldots, k.$$

Therefore, for a given interval-censored observation $(L_i, R_i]$, we generate k pair of observations,

$$(X_{ij},\, Y_{ij}), \quad j = 1, \ldots, k$$

for each $i = 1, \ldots, n$. Hence, for n interval-censored data $(L_i, R_i]$, $i = 1, \ldots, n$, we generate kn observations

$$(X_{ij},\, Y_{ij}), \quad i = 1, \ldots, n;\ j = 1, \ldots, k.$$

Note that X_{ij} denotes a specific time jw/k and Y_{ij} denotes the corresponding status of the event of interest, *i.e.*, $Y_{ij} = 0$ implies the event does not occurred yet at time jw/k and $Y_{ij} = 1$ implies the event already occurred at time jw/k. Figure 3.1 illustrates the construction of $(X_{ij},\, Y_{ij})$, $j = 1, \ldots, k$ for each $i = 1, \ldots, n$.

Remark 3.1. In generating (X_{ij}, Y_{ij}), $i = 1, \ldots, n$; $j = 1, \ldots, k$, we see that in some cases the mean value

$$M_i = \frac{L_i + R_i}{2} \tag{3.3}$$

of the interval $(L_i, R_i]$ instead of L_i in (3.2) gives better results (see Section 4).

In defining M_i, we replace R_i by $\hat{w}$ if $R_i = \infty$.

3.2. Proposed estimators

We propose a method of logistic regression based on kn observations

$$(X_{ij}, Y_{ij}), \quad i = 1, \ldots, n;\ j = 1, \ldots, k$$

to estimate the survival function of r.v. T and to estimate the median survival time of r.v. T in interval-censored data. Let

$$P(Y = 1 | X = t) = \pi(t)$$

be the probability of success for a given covariate t, and we consider a simple logistic regression model defined as

$$\operatorname{logit} \pi \equiv \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 t.$$

3.2.1. Median survival time Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be MLEs of β_0 and β_1, respectively. The median survival time is estimated as a time $\hat{t}_M$ when $\hat{\pi} = 0.5$. Therefore, the estimator of median survival time is given by

$$\hat{t}_M = -\frac{\hat{\beta}_0}{\hat{\beta}_1}.$$

For the interval estimation of the median survival time, we need to evaluate the variance of $\hat{t}_M$, however, exact computation is not possible. Instead we compute approximate variance using the multivariate delta method. Since

$$\mathrm{Var}\left(\hat{t}_M\right) = \mathrm{Var}\left[g\left(\hat{\beta}_0, \hat{\beta}_1\right)\right],$$

where $g(x, y) = -x/y$, we have

$$\mathrm{Var}\left(\hat{t}_M\right) \simeq \nabla g^t \Sigma \nabla g,$$

where

$$\nabla g = \left(\frac{\partial g(\beta_0, \beta_1)}{\partial \beta_0}, \frac{\partial g(\beta_0, \beta_1)}{\partial \beta_1}\right) = \left(-\frac{1}{\beta_1}, \frac{\beta_0}{\beta_1^2}\right)$$

and

$$\Sigma = \begin{pmatrix} \mathrm{Var}\big(\hat{\beta}_0\big) & \mathrm{Cov}\big(\hat{\beta}_0, \hat{\beta}_1\big) \\ & \mathrm{Var}\big(\hat{\beta}_1\big) \end{pmatrix}.$$

Therefore,

$$\mathrm{Var}(\hat{t}_M) \simeq \frac{v_{00} - 2\hat{\rho}v_{01} + \hat{\rho}v_{11}^2}{\hat{\beta}_1^2},$$

where $v_{00} = \mathrm{Var}(\hat{\beta}_0)$, $\hat{\rho} = \hat{\beta}_0/\hat{\beta}_1$, $v_{01} = \mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1)$, $v_{11} = \mathrm{Var}(\hat{\beta}_1)$. Using this result, we can compute the approximate $100 \times (1-\alpha)\%$ confidence interval for the median survival time t_M as

$$\left(\hat{t}_M - z_{\frac{\alpha}{2}}\,\mathrm{s.e.}\left(\hat{t}_M\right),\ \hat{t}_M + z_{\frac{\alpha}{2}}\,\mathrm{s.e.}\left(\hat{t}_M\right)\right),$$

where $\mathrm{s.e.}(\hat{t}_M) = (\widehat{\mathrm{Var}}(\hat{t}_M))^{1/2}$ is the standard error of $\hat{t}_M$.

3.2.2. Survival function As an estimator of the survival function $S(t)$, consider the estimator of the probability of success at time t, *i.e.*,

$$\begin{aligned}\hat{\pi}(t) &= \hat{P}(Y = 1 | X = t) \\ &= \frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 t\right)}{1 + \exp\left(\hat{\beta}_0 + \hat{\beta}_1 t\right)}.\end{aligned}$$

One disadvantage of the estimator $\hat{\pi}(t)$ is that $\hat{\pi}(t) \to 1$ as $t \to 0$ is not guaranteed. If $t \to 0$, then $\hat{S}_L(t)$ is close to $1/(1 + e^{-\hat{\beta}_0})$. Therefore, we suggest an adjusted version, $\hat{S}_L(t)$, of $\hat{\pi}(t)$ as

$$\hat{S}_L(t) = \hat{\pi}(t) + (1 - \hat{\pi}(0))\,\frac{\hat{t}_M - t}{\hat{t}_M},$$

which satisfies $\hat{S}_L(0) = 1$.

3.2.3. Regression model One novel aspect of our proposed estimator is that the survival function and the median survival time can be estimated even when covariates exist. Assume that $\mathbf{z}$ represents p-dimensional covariates including variables like age and sex. Consider a multiple logistic regression model

$$\operatorname{logit} \pi = \beta_0 + \beta_1 t + \boldsymbol{\gamma}' \mathbf{z}.$$

Then, the estimator of median survival time is given by

$$\hat{t}_M = -\frac{\hat{\beta}_0 + \hat{\boldsymbol{\gamma}}' \mathbf{z}}{\hat{\beta}_1}$$

and the estimator of survival function is given by

$$\hat{S}_L(t) = \hat{\pi}(t) + (1 - \hat{\pi}(0)) \frac{\hat{t}_M - t}{\hat{t}_M}.$$

3.2.4. Estimation of the number of subintervals To estimate the number of subintervals k based on the data, we propose to use the cross-validation criterion. Let M_i be the mean value of the interval censored data $(L_i, R_i]$ defined in (3.3). Let $\hat{t}_{M,k(i)}$ be the estimate of t_M using k subintervals and $(n-1)$ observations after deleting the i^{th} observation. Then, define

$$\mathrm{CV}(k) = \sum_{i=1}^{n} \left\{ M_i - \hat{t}_{M,k(i)} \right\}^2 \tag{3.4}$$

and the estimate of k is given by

$$\hat{k} = \arg\min_k \mathrm{CV}(k).$$

The performance of the proposed cross-validation criterion is very good.

4. Numerical Studies

4.1. Design for numerical studies

We perform simulation studies to see numerical behavior of the proposed estimator. One difficulty in numerical studies under interval censoring is generating random intervals because convenient and useful techniques are not suggested so far in the relevant literature. While the generation of randomly right censored data are quite easy and straightforward with required censoring percentage, the generation of random intervals in interval censoring is difficult in several points of view. First, it is not easy to decide the length of interval which is deeply related with the percentage of censoring because if the length of interval is large it is almost censored and if the length of interval is small it is almost uncensored. Second, it is not easy to control the number of intervals with an infinite right-handed value *i.e.*, $(L, \infty]$. Here we suggest a method of generating random intervals $(L, R]$ as follows;

Step 1: Generate random number T from a specified distribution, Weibull(λ, γ), say.

Step 2: Let $L = \max(0, T - U_L)$ and $R = T + U_R$, where $U_L \sim U(0, a)$ and $U_R \sim U(0, b)$ for some positive constants a and b.

Step 3: After generating n random intervals based on Step 1 and Step 2, replace randomly $\alpha \times 100\%$, where $0 < \alpha < 1$, of R_i's by ∞ among intervals $(L_i, R_i]$, $i = 1, \ldots, n$ satisfying L being larger than the median of T (*i.e.*, 0.69 for Weibull$(1, 1)$ and 1.67 for Weibull$(1/2, 2)$, respectively).

REMARK 4.1. In real data sets, the interval with $R = \infty$ often occurs when the corresponding L is large compared to others. Step 3 reflects this phenomenon when the percentage of infinity is less than 50%. Note that the percentage of infinity in interval censored data may be viewed as the censoring percentage in the right-censored data.

4.2. Numerical results

Sample sizes considered are $n = 30$ and 50, and 100 replications are done. In this study, we will compare four estimators for the median survival time t_M:

(1) $\hat{t}_P$: parametric distribution method using the true function

(2) $\hat{t}_I$: mean imputation method

(3) $\hat{t}_S$: self-consistency method

(4) $\hat{t}_L$: logistic regression method

Also, we will compare four estimators for the survival function $S(t)$:

(1) $\hat{S}_P(t)$: parametric distribution method using the true function

(2) $\hat{S}_I(t)$: mean imputation method

(3) $\hat{S}_S(t)$: self-consistency method

(4) $\hat{S}_L(t)$: logistic regression method

To compare the numerical performance of each estimator we will evaluate mean squared error(MSE) for estimators of the median survival time and mean integrated squared error(MISE) for estimators of the survival function.

The distribution we considered for the survival time is Weibull distribution with parameters λ and γ denoted by Weibull(λ, γ). Here we consider two types of Weibull distribution, Weibull$(1, 1)$ and Weibull$(1/2, 2)$. First, Weibull$(1, 1)$ is the exponential distribution with mean 1, and the true median is $t_M = 0.69$. Second, Weibull$(1/2, 2)$ is the rightly skewed distribution, and the true median is $t_M = 1.67$. Also, we consider $1/2$ and 1 for a and b in Step 2, and we consider $\alpha = 0.1$ and $\alpha = 0.3$ in Step 3.

To decide the number of subintervals k, we proposed to use the cross validation criterion given in (3.4). The see the behavior of this criterion we compute MSE and CV for $k = 2, 3, \ldots, 20$, respectively. Figure 4.1 shows them when the true distribution is Weibull$(1, 1)$, the sample size is 30, $a = b = 1$, and $\alpha = 0.1$. We see that MSE is minimized when $k = 9$ and CV is minimized when $k = 11$. Also, both curves show very similar pattern. For other distributions with different sample size, different a and b, and different α, they show a quite similar pattern. In fact, the minimizer of MSE is between 7 and 15, and the minimizer of CV is between 7 and 14 in all situations considered. Also, the difference between minimizers of MSE and CV in all situations was less than 4. Conclusively, the CV criterion in (3.4) reflects the behavior of MSE very well.

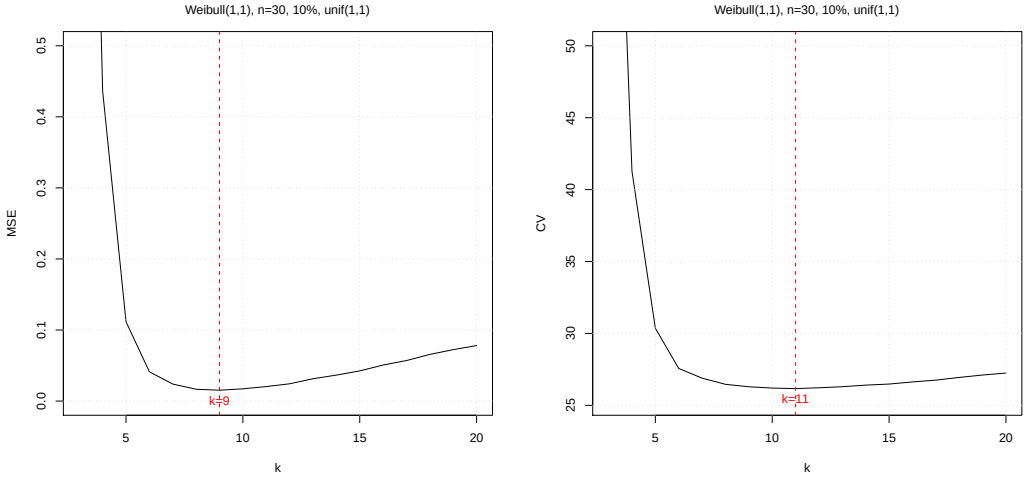

Figure 4.1. MSE and CV from Weibull$(1, 1)$ with $a = b = 1$, $\alpha = 0.1$, and $n = 30$.

Table 4.1. MSE of $\hat{t}_M$

Dist.	n	% of ∞	(a, b)	$\hat{t}_P$	$\hat{t}_I$	$\hat{t}_S$	$\hat{t}_L$
Weibull$(1, 1)$	30	10%	(0.5, 0.5)	2.99	3.74	4.58	2.35
			(1, 1)	3.62	3.60	5.36	1.52
		30%	(0.5, 0.5)	5.65	3.28	3.99	2.44
			(1, 1)	4.16	2.55	4.09	1.62
	50	10%	(0.5, 0.5)	2.08	2.39	3.19	1.64
			(1, 1)	2.27	1.75	3.08	0.82
		30%	(0.5, 0.5)	3.82	3.23	3.53	1.46
			(1, 1)	2.96	1.46	2.16	0.82
Weibull$(2, 2)$	30	10%	(0.5, 0.5)	3.60	4.75	5.89	1.77
			(1, 1)	3.85	4.82	6.80	3.50
		30%	(0.5, 0.5)	6.32	4.66	5.85	1.43
			(1, 1)	5.15	4.16	6.70	3.76
	50	10%	(0.5, 0.5)	2.44	3.48	3.61	1.49
			(1, 1)	2.74	3.36	4.66	3.71
		30%	(0.5, 0.5)	5.28	4.36	4.94	1.59
			(1, 1)	3.32	3.59	4.26	4.09

To see the numerical performance we evaluate MSE of four estimators, $\hat{t}_P$, $\hat{t}_I$, $\hat{t}_S$, and $\hat{t}_L$ of the median survival time with the results listed in Table 4.1. The proposed estimator $\hat{t}_L$ performs very well and in most cases it is best in the sense of MSE. Also, the mean imputation method $\hat{t}_I$ is quite comparable to $\hat{t}_P$ and is even superior in many cases. In general, the MSE decreases as n increases, however, it does not depend a lot on the percentage of infinity and the length of interval.

For the performance of four estimators of the survival function, we compute MISE. As shown in Table 4.2, we see similar results as in estimators of the median survival time. One difference fact is that MISE of the estimators of the survival function depends on the length of interval. This phenomenon stems from the fact that a wide interval can cause poor estimate of survival function at boundaries, while it may not affect the estimation of the median survival time.

Table 4.2. MISE of $\hat{S}(t)$

Dist.	n	% of ∞	(a, b)	$\hat{S}_P$	$\hat{S}_M$	$\hat{S}_S$	$\hat{S}_L$
Weibull(1, 1)	30	10%	(0.5, 0.5)	1.32	2.04	2.89	1.62
			(1, 1)	1.48	2.43	3.67	2.07
		30%	(0.5, 0.5)	5.06	7.82	7.34	1.89
			(1, 1)	4.06	6.83	6.67	2.20
	50	10%	(0.5, 0.5)	0.96	1.51	1.85	1.26
			(1, 1)	0.98	1.83	1.99	1.57
		30%	(0.5, 0.5)	4.41	8.00	7.09	1.30
			(1, 1)	3.58	7.10	6.32	1.61
Weibull(2, 2)	30	10%	(0.5, 0.5)	1.26	1.96	2.71	2.20
			(1, 1)	1.37	1.97	3.27	5.48
		30%	(0.5, 0.5)	3.20	5.08	5.37	2.20
			(1, 1)	2.40	4.07	5.11	5.23
	50	10%	(0.5, 0.5)	0.85	1.35	1.76	1.71
			(1, 1)	0.93	1.28	1.90	4.87
		30%	(0.5, 0.5)	2.85	5.04	4.91	1.74
			(1, 1)	1.77	3.47	3.50	5.37

Table 5.1. Observed intervals in months for times to breast retraction of early breast cancer patients

Group	Observed intervals in months
RT	(45,], (27, 37], (37,], (4, 11], (17, 25], (6, 10], (46,], (0, 5], (33,], (15,] (0, 7], (26, 40], (18,], (46,], (19, 26], (46,], (46,], (24,], (11, 15], (11, 18], (46,], (27, 34], (36,], (37,], (22,], (7, 16], (36, 44], (5, 12], (38,], (34,], (17,], (46,], (19, 35], (46,], (5, 12], (9, 14], (36, 48], (17, 25], (36,], (46,], (37, 44], (37,], (24,], (0, 8], (40,], (33,]
RCT	(8, 12], (0, 5], (30, 34], (16, 20], (13,], (0, 22], (5, 8], (13,], (30, 36], (18, 25], (24, 31], (12, 20], (10, 17], (17, 24], (18, 24], (17, 27], (11,], (8, 21], (17, 26], (35,], (17, 23], (33, 40], (4, 9], (16, 60], (33,], (24, 30], (31,], (11,], (15, 22], (35, 39], (16, 24], (13, 39], (15, 19], (23,], (11, 17], (13,], (19, 32], (4, 8], (22,], (44, 48], (11, 13], (34,], (34,], (22, 32], (11, 20], (14, 17], (10, 35], (48,]

∗ RT: radiation therapy alone
RCT: radiation therapy plus adjuvant chemotherapy

5. An Example

For an illustrative example, we use the data set of the RCT (radiation therapy plus adjuvant chemotherapy) group of early breast cancer (Finkelstein and Wolfe, 1985) given in Table 5.1. Here we focus on the RT group which consists of 46 interval censored observations. The number of finite R_i's is $d = 21$ and the average length of finite intervals is $A = 8.97$. Also, $L_{(n)} = 44$ and $R_{(d)} = 60$. Therefore, we estimate w as $\hat{w} = \max(44, 60) + 8.97 = 68.97$. In addition, we estimate the number of subintervals $k = 26$ based on the cross-validation criterion in (3.4), and the behavior of CV(k) is given in Figure 5.1(a). We compute three estimators of the survival function; $\hat{S}_I(t)$, $\hat{S}_S(t)$, and $\hat{S}_L(t)$. Using these results, we compute three estimators of the median survival time; $\hat{t}_I$, $\hat{t}_S$, and $\hat{t}_L$. The results are given in Figure 5.1(b). As shown in this Figure we see that $\hat{t}_I = 21.5$, $\hat{t}_S = 20.0$, $\hat{t}_L = 17.2$. Note that the mean imputation method is quite similar to the Kaplan-Meier estimator in the right censored data, and it is well known that the estimator of the median survival based on the Kaplan-Meier estimator in the right censored data tends to overestimate the true median survival time. In this sense, it is preferable if the median survival

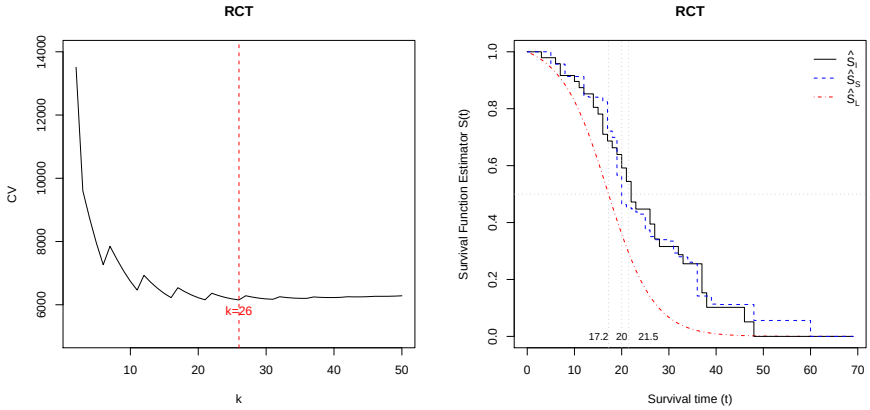

Figure 5.1. (a) Cross-validation curve and (b) Estimators of survival function and median survival times in breast cancer data.

time estimator is less than the estimator based on the mean imputation method. Note that the proposed estimator is less than that of the mean imputation method.

6. Concluding Remarks

Interval censoring is a type of censoring that has become increasingly common in the areas of medicine and clinical trials. Interval-censored data are incomplete data, where the exact time of the event is unavailable; however, the event is known to occur between two defined times. Interval-censored data are more difficult to deal with than the right-censored data because of the complexity and special structure of the interval-censored data. Thus, existing methods in right-censored data cannot be directly applicable to the interval-censored data. For example, the Kaplan-Meier type estimator for the interval censored data is not analytically given. There are four types in interval-censored data; case I interval-censoring, case II interval censoring, double interval censoring, and panel count data.

In this paper, studied estimation of survival function and median survival time in case II (general) interval censoring among four types of interval-censored data. Specifically, we suggested a logistic regression method to estimate the median survival time and survival function in interval-censored data. To do this we partition the support into k subintervals of equal length, and the response variable will be 0 until the subinterval meet the mean of interval-censored time. Since then the response variable is assigned as 1. The covariate is the time representing the subintervals. Also, we proposed a new and novel method of generating random numbers under interval-censoring set-up. Through simulation studies we compare our methods to imputation method and nonparametric maximum likelihood estimator. Under various simulation schemes the proposed estimators showed better numerical performance in the sense of the mean squared error for estimating the median survival time and the mean integrated squared error for estimating the survival function. Also, the proposed estimator can be used when covariates exist while existing estimators cannot.

References

Efron, B. (1967). The two sample problem with censored data, In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley, 831–853.

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

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

Finkelstein, D. M. and Wolfe, R. A. (1985). A semiparametric model for regression analysis of interval-censored failure time data, *Biometrics*, **41**, 933–945.

Groeneboom, P. and Wellner, J. A. (1992). *Information Bounds and Nonparametric Maximum likelihood Estimation*, DMV Seminar, Band 19, Birkhauser, New York.

Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoring, *The Annals of Statistics*, **24**, 540–568.

Hudgens, M. G. (2005). On nonparametric maximum likelihood estimation with interval censoring and left truncation, *Journal of the Royal Statistical Society, Series B*, **67**, 573–587.

Jongbloed, G. (1998). The iterative convex minorant algorithm for nonparametric estimation, *Journal of Computational and Graphical Statistics*, **7**, 310–321.

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

Kim, C., Park, B. U., Kim, W. and Lim, C. (2003). Bézier curve smoothing of the Kaplan-Meier estimator, *The Annals of the Institute of Statistical Mathematics*, **55**, 359–367.

Lawless, J. F. and Babineau, D. (2006). Models for interval censoring and simulation-based inference for lifetime distributions, *Biometrika*, **93**, 671–686.

Lindsey, J. C. and Ryan, L. M. (1998). Tutorial in biostatistics: Methods for interval-censored data, *Statistics in Medicine*, **17**, 219–238.

Peto, R. (1973). Experimental survival curves for interval-censored data, *Applied Statistics*, **22**, 86–91.

Rubin, D. B. (1987). *Multiple Imputation for Noresponse in Surveys*, Wiley, New York.

Sun, J. (2006). *The Statistical Analysis of Interval-Censored Failure Time Data*, Springer.

Tanner, M. A. and Wong, W. H. (1987). The application of Imputation to an estimation problem in grouped lifetime analysis, *Technometrics*, **29**, 23–32.

Tanner, M. A. (1991). *Tools for Statistical Inference: Observed Data and Data Augmentation Methods*, Springer-Verlag, New York.

Turnbull, B. W. (1976). The empirical distribution with arbitrarily grouped censored and truncated data, *Journal of the Royal Statistical Society, Series B*, **38**, 290–295.

Wellner, J. and Zhan, Y. (1997). A hybrid algorithm for computation of the nonparametric maximum likelihood estimator from censored data, *Journal of the American Statistical Association*, **92**, 945–959.