Journal of Quality Measurement and Analysis *JQMA* 20(1) 2024, 171-186
e-ISSN: 2600-8602 *http://www.ukm.my/jqma*
https://doi.org/10.17576/jqma.2001.2024.13

COMPARISON OF SEVERAL IMPUTATION TECHNIQUES FOR LOG LOGISTIC MODEL WITH COVARIATE AND INTERVAL CENSORED DATA

(Perbandingan antara Beberapa Teknik Imputasi untuk Model Logistik Log dengan Kovariat dan Data Tertapis Selang)

TEEA YUAN XIN* & JAYANTHI ARASAN

ABSTRACT

The main purpose of this study is to compare the performance of midpoint, right, and left imputation techniques for log logistic model with covariate and censored data. The maximum likelihood estimation method (MLE) is used to check the efficiency of imputation techniques by estimating the parameters. The performance of the estimates is evaluated based on their bias, standard error (SE), and root mean square error (RMSE) at different sample sizes, censoring proportions, and interval widths via a simulation study. Based on the results of the simulation study, the right imputation had the best overall performance. Finally, the proposed model is fitted to the real breast cancer data. The findings suggest that the log logistic model fits the breast cancer data well and the covariate of treatment significantly affects the time to cosmetic deterioration of the breast cancer patients.

Keywords: log logistic; imputation techniques; covariate; right censored; interval censored

ABSTRAK

Tujuan utama kajian ini adalah untuk membandingkan prestasi teknik imputasi titik tengah, kanan dan kiri bagi model logistik log dengan kovariat dan data tertapis. Kaedah penganggaran kebolehjadian maksimum (MLE) digunakan untuk memeriksa efisiensi teknik imputasi dengan menganggar parameter. Prestasi anggaran parameter dengan teknik imputasi titik tengah, kanan dan kiri dinilai dan dibandingkan pada saiz sampel, kadar penapisan dan lebar selang yang berbeza. Berdasarkan hasil kajian simulasi, teknik imputasi kanan mempunyai prestasi keseluruhan yang terbaik. Akhirnya, model yang dicadangkan telah dipadan pada data kanser payudara sebenar. Hasilnya mencadangkan bahawa model logistik log sesuai dengan data kanser payudara dan kovariat rawatan mempengaruhi masa kemerosotan kosmetik pesakit kanser payudara secara signifikan.

Kata kunci: logistic log; teknik imputasi; kovariat; tertapis kanan; tertapis selang

1. Introduction

According to Smith and Smith (2001), survival analysis refers to a statistical approach that considers the amount of time of an experimental unit in a study. It is a time study between the entry of a subject and the observation of the event of interest. It is extensively applied in numerous fields such as medical and biological sciences, social and economic sciences, as well as in engineering.

The time to the occurrence of the event of interest is known as survival time, failure time, or time-to-event. Survival time, $T \geq 0$ measures the duration of an event from a particular time origin until the occurrence of the event interest. It can be measured in years, months, weeks, days, minutes, or seconds. The survival data is said to be censored if the exact survival times of

the subjects are unknown while the observations are classified as uncensored if the set of survival times is complete. There are three types of censoring in survival analysis, namely left censoring, right censoring, and interval censoring.

According to Muse *et al.* (2021), log logistic distribution is a continuous probability distribution for a non-negative random variable where its logarithm has a logistic distribution. This distribution has a similarity with log normal distribution in terms of their shape but it exhibits heavier tails. This characteristic makes the log logistic distribution to be more suitable for lifetime data analysis compared with the log normal distribution. Besides, it is very appealing because its cumulative distribution function (CDF) can be expressed in closed form which makes it particularly useful for analysis of survival data with incomplete information such as censoring and truncation. It also has non-monotonic hazard rate which its rate increases at first and eventually decreases. Therefore, this property of its hazard function makes it appropriate for modelling some sets of survival data in medical studies whose rate increases initially and then declines. For instance, the medical studies can be those involving lung cancer, breast cancer, and kidney or heart transplant patients, as described by Arasan and Adam (2014).

Imputation technique is a technique of replacing the missing data with substituted values. In survival analysis, it is used to deal with missing event times of censored observations. In this study, the imputation techniques are employed to approximate the interval censored data which consists of left and right endpoints of the censoring intervals. The three major types of simple imputation techniques are midpoint imputation, right imputation, and left imputation.

Previously, there are some researchers who had done research on the log logistic model with covariate, uncensored, right, and interval censored data. Loh *et al.* (2017) studied the estimation procedure and Wald method for the parameters of the log logistic model with doubly interval, interval, right censored, and uncensored data. In a recent work conducted by Lai and Arasan (2020), the adequacy of the log logistic model with covariate, right, and interval censored data was investigated by applying different types of imputation techniques. So, there are only a few studies done on the log logistic model and it is essential to further explore this model with uncensored, right and interval censored data. In general, we still could not identify the best imputation technique for the log logistic model with covariate, uncensored, right, and interval censored data. Thus, assessing several imputation techniques to deal with the problem of uncensored, right, and interval censored data will be the interest of our study.

This study aims to incorporate covariate into the log logistic model with uncensored, right, and interval censored data and apply maximum likelihood estimation method to obtain its parameter estimates. A simulation study was conducted to assess the performance of the parameter estimates with the midpoint, right, and left imputation techniques to deal with the problem of uncensored, right, and interval censored data at various sample sizes, censoring proportions, and interval widths, and identify the best technique based on the values of bias, standard error (SE), and root mean square error (RMSE) of parameter estimates. Finally, the log logistic regression model was fitted to real data with right and interval censored observations with covariate to evaluate the overall performance of the proposed model in real life situations.

2. Methodology

2.1. *Log logistic regression model*

Let $T \geq 0$ be a non-negative random variable denoting the survival time with the probability density function (PDF) of log logistic model,

$$f(t,\alpha,\beta) = \frac{\left(\frac{\beta}{\alpha}\right)\left(\frac{t}{\alpha}\right)^{\beta-1}}{\left[1+\left(\frac{t}{\alpha}\right)^{\beta}\right]^2} \; ; \; t>0, \alpha>0, \beta>0. \tag{1}$$

In this study, the model is extended to incorporate covariate by letting $\alpha = exp(-\mu)$ and $\beta = \frac{1}{\sigma}$ where $\mu = \boldsymbol{\beta}' x$, t denotes the failure time, $\boldsymbol{X}' = \left(x_0, x_1, \ldots, x_p\right)$ is the vector of covariate values, $x_0 = 1$, $\beta^{'} = \left(\beta_0, \beta_1, \ldots, \beta_p\right)$ is a vector for unknown parameters. Thus, the density function now becomes,

$$f(t) = \frac{\left(\frac{\frac{1}{\sigma}}{exp(-\mu)}\right)\left(\frac{t}{exp(-\mu)}\right)^{\frac{1}{\sigma}-1}}{\left[1+\left(\frac{t}{exp(-\mu)}\right)^{\frac{1}{\sigma}}\right]^2}. \tag{2}$$

Assuming we have n independent random variable, if $i = 1,2,\ldots,n$ and $\mu_i = \beta_0 + \beta_1 x_i$, where x_i is single covariate. Let $z_i = \frac{y_i - \beta_0 - \beta_1 x_i}{\sigma}$ where $y_i = ln(t_i)$ is the log lifetime for i^{th}observations, then the PDF and survival function of log logistic regression model are as follows,

$$f(z_i,\beta,\sigma,x_i) = \frac{exp(z_i)}{\sigma[1+exp(z_i)]^2} \; , -\infty < z_i < \infty. \tag{3}$$

$$S(z_i,\beta,\sigma,x_i) = \frac{1}{[1+exp(z_i)]}, -\infty < z_i < \infty. \tag{4}$$

The failure time, t_i can be simulated using the inverse transform method which is shown as follows,

$$t_i = \left(\frac{u_i}{1-u_i}\right)^{\sigma} exp(\beta_0 + \beta_1 x_i). \tag{5}$$

where u_i is the random number generated from standard uniform distribution.

2.2. *Maximum likelihood estimation*

A censoring indicator variable needs to be defined to determine if an observation is censored or uncensored. Let:

$$c_i = \begin{cases} 1, \text{ if } t_i \text{ uncensored or interval censored;} \\ 0, \text{ if } t_i \text{ right censored.} \end{cases}$$

and let

$$\widetilde{t_i} = \begin{cases} \frac{t_{L_i} + t_{R_i}}{2} & , \text{ for midpoint imputation;} \\ t_{R_i} & , \text{ for right imputation;} \\ t_{L_i} & , \text{ for left imputation;} \\ t_i & , \text{ otherwise.} \end{cases}$$

In this study, midpoint, right, and left imputation techniques will be employed to deal with the interval censored data. For the midpoint imputation, its lifetime is approximated by taking the midpoint of the censoring interval $[t_{L_i}, t_{R_i}]$ as $\frac{(t_{L_i}, t_{R_i})}{2}$. While for the right and left imputation methods, the lifetime is imputed by the right limit of censoring interval, t_{R_i}, and left limit of the censoring interval, t_{L_i}, respectively.

The likelihood function for the log logistic regression model with the presence of uncensored, right, and interval censored lifetime data is given by,

$$L(\beta, \sigma) = \prod_{i=1}^{n} [f(\tilde{t_i})]^{c_i} [S(t_i)]^{1-c_i}$$
$$= \prod_{i=1}^{n} \left[\frac{exp(\tilde{z_i})}{\sigma(1+exp(\tilde{z_i}))^2}\right]^{c_i} ([1 + exp(z_i)]^{-1})^{1-c_i}. \quad (6)$$

The loglikelihood function is,

$$l(\beta, \sigma) = \sum_{i=1}^{n} \{ c_i [\tilde{z_i} - ln(\sigma) - 2\, ln(1 + exp(z_i))]$$
$$-(1 - c_i)\, ln(1 + exp(z_i))\} \quad (7)$$

where $z_i = \frac{ln(t_i) - \beta_0 - \beta_1 x_i}{\sigma}$ and $\tilde{z_i} = \frac{\ln(\tilde{t_i}) - \beta_0 - \beta_1 x_i}{\sigma}$.

Let $A_i = \exp(z_i)\,[1 + \exp(z_i)]^{-1}$ and $B_i = \exp(\tilde{z_i})\,[1 + \exp(\tilde{z_i})]^{-1}$ where $i = 1,2, \ldots, n$, $z_i = \frac{ln(t_i) - \beta_0 - \beta_1 x_i}{\sigma}$ and $\tilde{z_i} = \frac{ln(\tilde{t_i}) - \beta_0 - \beta_1 x_i}{\sigma}$. The first derivatives of loglikelihood function with respect to the parameters are,

$$\frac{\partial(\beta,\sigma)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{x_{ij}}{\sigma} \{c_i(2B_i - 1) + (1 - c_i)A_i\}. \quad (8)$$

where $j = 0,1$ and $x_{i0} = 1$.

$$\frac{\partial(\beta,\sigma)}{\partial \sigma} = \sum_{i=1}^{n} \frac{1}{\sigma} \{c_i \tilde{z_i}(2B_i - 1) - c_i + A_i z_i (1 - c_i)\}. \quad (9)$$

where $x_{i0} = 1$.

Next, Newton-Raphson method will be used to estimate the parameters of the log logistic regression model.

2.3. *Fisher information*

Fisher information matrix of the log logistic regression model which is approximated by the observed information matrix can be expressed as follows,

$$i(\beta_0,\beta_1,\sigma)=\begin{bmatrix} -\frac{\partial^2 l}{\partial\beta_0^2} & -\frac{\partial^2 l}{\partial\beta_0\partial\beta_1} & -\frac{\partial^2 l}{\partial\beta_0\partial\sigma} \\ -\frac{\partial^2 l}{\partial\beta_1\partial\beta_0} & -\frac{\partial^2 l}{\partial{\beta_1}^2} & -\frac{\partial^2 l}{\partial\beta_1\partial\sigma} \\ -\frac{\partial^2 l}{\partial\sigma\partial\beta_0} & -\frac{\partial^2 l}{\partial\sigma\partial\beta_1} & -\frac{\partial^2 l}{\partial\sigma^2} \end{bmatrix}, \tag{10}$$

evaluated at $\hat{\beta}_0, \hat{\beta}_1$, and $\hat{\sigma}$.

2.4. *Log rank test*

Log rank test is a commonly used statistical test to assess whether there is a difference in the survival experience or survival function between two or more groups. The log rank test statistics is given by,

$$\chi^2_{LR}=\frac{\left[\sum_{j=1}^{k}(d_{1j}-e_{1j})\right]^2}{\sum_{j=1}^{k}v_{1j}} \sim \chi^2(1) \tag{11}$$

where d_{1j} is the number of failures at t_j for group 1, $e_{1j}=\frac{n_{1j}d_j}{n_j}$, is the expected number of failures and $v_{1j}=\frac{n_{1j}n_{2j}d_j(n_j-d_j)}{{n_j}^2(n_j-1)}$, is variance of the observed number of failures.

2.5. *Wald confidence interval*

Cox and Hinkley (1979) mentioned if $\hat{\theta}$ is the maximum likelihood estimator for parameter θ then $\hat{\theta}$ is asymptotically normally distributed with mean θ and covariance matrix $I^{-1}(\theta)$ under mild regularity conditions where $I(\theta)$ is the Fisher information matrix,

$$\hat{\theta} \sim N\left(\theta, I^{-1}(\theta)\right). \tag{12}$$

The $100(1-\alpha)\%$ confidence interval for a single parameter θ_j is given by,

$$\hat{\theta}_j \pm z_{1-\frac{\alpha}{2}}\left[\widehat{s.e}\left(\hat{\theta}_j\right)\right]. \tag{13}$$

where $z_{1-\frac{\alpha}{2}}$ is the $100\left(1-\frac{\alpha}{2}\right)^{th}$ percentile of standard normal distribution and $\widehat{s.e}$ is the estimated standard error of $\hat{\theta}_j$. Thus, the $100(1-\alpha)\%$ confidence interval for β is given as follows,

$$\hat{\beta} \pm z_{1-\frac{\alpha}{2}}\left[\widehat{s.e}\left(\hat{\beta}\right)\right]. \tag{14}$$

3. Simulation Study

3.1. *Simulation of log logistic regression model*

After a covariate, right and interval censored data were included in the log logistic model, a simulation study for the log logistic regression model was conducted using R software with 1000 replications at sample sizes of 20, 50, 80, 100 and the censoring proportions, 0%, 20%, 25%, 30%, 35%, 40% to identify the best imputation technique. The value set for the three

parameters were $\beta_0 = 2.4, \beta_1 = 1,$ and $\sigma = 0.6$ to mimic real data with some adjustments. Approximate censoring proportion is denoted by CP, which refers to the combination of right and interval censoring proportions for the simulated data.

Firstly, we generated a sequence of random number $\mathrm{u_i}$ from uniform distribution, $u_i \sim U(0,1)$. Second, the covariate, $\mathrm{x_i}$ was simulated from standard normal distribution. Both u_i and $\mathrm{x_i}$ were used to get survival times, $\mathrm{t_i}$ of log logistic regression model with inverse transform method. Then, censoring times, $\mathrm{s_i}$ were randomly generated from exponential distribution to obtain right censored data. The survival times, t_i will be uncensored if $\mathrm{t_i} \leq \mathrm{s_i}$ and right censored if $\mathrm{t_i} > \mathrm{s_i}$. Next, a small percentage of the right censored data will be randomly selected by using Bernoulli distribution with parameter p and converted into interval censored data. Also, the time intervals were set as 4 and 6 months and they were compared.

The parameter estimates of β_0, β_1, and σ were computed using maximum likelihood estimation with the midpoint, right, and left imputation techniques and Newton-Raphson procedure. In order to examine the performance of the parameter estimates, their values of bias, SE, and RMSE were calculated. The bias of a parameter estimate can be described as the difference between the expected value and the true value of the parameter estimate. It can be used to measure the accuracy of the estimate.

$$bias(\hat{\theta}) = E(\hat{\theta}) - \theta. \tag{15}$$

SE is defined as an important measure of variability between estimates. It is useful in measuring the efficiency of the estimate.

$$SE(\hat{\theta}) = \sqrt{\frac{\sum \hat{\theta}_i^2 - \frac{(\sum \hat{\theta}_i)^2}{N}}{N-1}}. \tag{16}$$

RMSE gives the summary of an estimator's average error. Since RMSE is the combination of both bias and standard error which measure the accuracy and efficiency of estimates, the best imputation technique will be chosen based on it.

$$RMSE = \sqrt{bias(\hat{\theta})^2 + s.e(\hat{\theta})^2}. \tag{17}$$

3.2. *Results and discussion*

The values of bias, SE, and RMSE of $\hat{\beta}_0, \hat{\beta}_1$and $\hat{\sigma}$ will be compared at various sample sizes, censoring proportions, and imputation methods. The tables and figures of the bias, SE, and RMSE for one of the estimates, $\hat{\beta}_0$ with the interval width of 4 and 6 are included in this paper, namely Table 1-6 and Figure 1-6. For the interval width of 4, the bias values of $\hat{\beta}_0$ fluctuate with the left imputation method. The bias of $\hat{\beta}_0$ increases slightly as the censoring proportions increase for the midpoint imputation technique while they gradually increase with the rising censoring proportions when the right imputation technique was employed. Next, it can be observed that the bias values of $\hat{\beta}_1$ and $\hat{\sigma}$ using the left imputation method always increase when the censoring proportions increase. When the midpoint and right imputation methods are used, the bias values of $\hat{\beta}_1$ and $\hat{\sigma}$ decline gradually with the increasing censoring proportions.

The SE values of $\hat{\beta}_0, \hat{\beta}_1$, and $\hat{\sigma}$ for the left imputation method always increase as the censoring proportions increase when the sample size is fixed whereas they are erratic as the

sample size increases. In contrast, the SE values of $\hat{\beta}_0, \hat{\beta}_1$, and $\hat{\sigma}$ with both the midpoint and right imputation methods decrease slightly as the censoring proportions increase when the sample size is fixed. While as the sample size increases, the SE of $\hat{\beta}_0, \hat{\beta}_1$, and $\hat{\sigma}$ are smaller for the same CP.

Table 1: Bias of $\hat{\beta}_0$ with midpoint, right and left imputation techniques (interval width of 4)

n	Right CP	Interval CP	CP	Midpoint	Right	Left
20	0.000	0.000	0.000	0.00749	0.00749	0.00749
	0.200	0.000	0.200	0.07624	0.08405	0.09820
	0.200	0.050	0.250	0.07661	0.09783	0.12806
	0.200	0.100	0.300	0.07835	0.10627	0.13689
	0.200	0.150	0.350	0.07895	0.12647	0.13308
	0.200	0.200	0.400	0.07918	0.13188	0.11939
50	0.000	0.000	0.000	0.00091	0.00091	0.00091
	0.200	0.000	0.200	0.11857	0.16550	0.11360
	0.200	0.040	0.250	0.12007	0.17676	0.09122
	0.200	0.100	0.300	0.12078	0.18716	0.05112
	0.200	0.140	0.350	0.12151	0.19287	0.02493
	0.200	0.200	0.400	0.12333	0.20945	-0.07054
80	0.000	0.000	0.000	-0.00318	-0.00318	-0.00318
	0.175	0.025	0.200	0.11385	0.13011	0.13273
	0.175	0.075	0.250	0.11557	0.14467	0.13990
	0.175	0.125	0.300	0.11819	0.16490	0.12562
	0.175	0.175	0.350	0.11918	0.17593	0.10363
	0.175	0.225	0.400	0.12003	0.18427	0.08065
100	0.000	0.000	0.000	0.00331	0.00331	0.00331
	0.190	0.010	0.200	0.11725	0.13021	0.13499
	0.190	0.060	0.250	0.11910	0.14487	0.14690
	0.190	0.110	0.300	0.12069	0.15929	0.14774
	0.190	0.160	0.350	0.12228	0.17350	0.13498
	0.190	0.210	0.400	0.12490	0.19863	0.06518

Table 2: SE of $\hat{\beta}_0$ with midpoint, right and left imputation techniques (interval width of 4)

n	Right CP	Interval CP	CP	Midpoint	Right	Left
20	0.000	0.000	0.000	0.23818	0.23818	0.23818
	0.200	0.000	0.200	0.29554	0.29463	0.34264
	0.200	0.050	0.250	0.29527	0.29286	0.40720
	0.200	0.100	0.300	0.29387	0.29117	0.44829
	0.200	0.150	0.350	0.29308	0.28678	0.57866
	0.200	0.200	0.400	0.29304	0.28565	0.60956
50	0.000	0.000	0.000	0.14396	0.14396	0.14396
	0.200	0.000	0.200	0.18724	0.18323	0.27655
	0.200	0.040	0.250	0.18719	0.18228	0.30871
	0.200	0.100	0.300	0.18732	0.18184	0.35352
	0.200	0.140	0.350	0.18726	0.18121	0.37756
	0.200	0.200	0.400	0.18673	0.17964	0.46308
80	0.000	0.000	0.000	0.11557	0.11557	0.11557
	0.175	0.025	0.200	0.14765	0.14613	0.16374
	0.175	0.075	0.250	0.14759	0.14514	0.17841
	0.175	0.125	0.300	0.14724	0.14373	0.21180
	0.175	0.175	0.350	0.14727	0.14283	0.24134
	0.175	0.225	0.400	0.14737	0.14256	0.26321
100	0.000	0.000	0.000	0.10525	0.10525	0.10525
	0.190	0.010	0.200	0.12682	0.12623	0.13732
	0.190	0.060	0.250	0.12690	0.12565	0.14910
	0.190	0.110	0.300	0.12669	0.12458	0.16433
	0.190	0.160	0.350	0.12644	0.12339	0.18583
	0.190	0.210	0.400	0.12592	0.12142	0.24344

Table 3: RMSE of $\hat{\beta}_0$ with midpoint, right and left imputation techniques (interval width of 4)

n	Right CP	Interval CP	CP	Midpoint	Right	Left
20	0.000	0.000	0.000	0.23829	0.23829	0.23829
	0.200	0.000	0.200	0.30522	0.30639	0.35644
	0.200	0.050	0.250	0.30504	0.30876	0.42687
	0.200	0.100	0.300	0.30413	0.30996	0.46872
	0.200	0.150	0.350	0.30353	0.31343	0.59377
	0.200	0.200	0.400	0.30355	0.31463	0.62114
50	0.000	0.000	0.000	0.14397	0.14397	0.14397
	0.200	0.000	0.200	0.22163	0.24690	0.29897
	0.200	0.040	0.250	0.22239	0.25391	0.32191
	0.200	0.100	0.300	0.22289	0.26095	0.35720
	0.200	0.140	0.350	0.22323	0.26464	0.37838
	0.200	0.200	0.400	0.22378	0.27594	0.46842
80	0.000	0.000	0.000	0.11561	0.11561	0.11561
	0.175	0.025	0.200	0.18645	0.19566	0.21077
	0.175	0.075	0.250	0.18745	0.20492	0.22672
	0.175	0.125	0.300	0.18881	0.21875	0.24626
	0.175	0.175	0.350	0.18945	0.22661	0.26265
	0.175	0.225	0.400	0.19007	0.23298	0.27529
100	0.000	0.000	0.000	0.10530	0.10530	0.10530
	0.190	0.010	0.200	0.17272	0.18135	0.19255
	0.190	0.060	0.250	0.17404	0.19177	0.20931
	0.190	0.110	0.300	0.17497	0.20222	0.22098
	0.190	0.160	0.350	0.17590	0.21290	0.22968
	0.190	0.210	0.400	0.17736	0.23280	0.25202

Table 4: Bias of $\hat{\beta}_0$ with midpoint, right and left imputation techniques (interval width of 6)

n	Right CP	Interval CP	CP	Midpoint	Right	Left
20	0.000	0.000	0.000	0.00749	0.00749	0.00749
	0.200	0.000	0.200	0.09438	0.10505	0.13283
	0.200	0.050	0.250	0.09705	0.12596	0.17141
	0.200	0.100	0.300	0.09805	0.13389	0.18415
	0.200	0.150	0.350	0.10476	0.17209	0.12165
	0.200	0.200	0.400	0.10557	0.17594	0.11210
50	0.000	0.000	0.000	0.00091	0.00091	0.00091
	0.220	0.000	0.200	0.14755	0.20573	0.13624
	0.220	0.020	0.250	0.14973	0.21810	0.09629
	0.220	0.080	0.300	0.15253	0.23670	0.00350
	0.220	0.120	0.350	0.15403	0.24473	-0.05166
	0.220	0.180	0.400	0.15887	0.27217	-0.28327
80	0.000	0.000	0.000	-0.00318	-0.00318	-0.00318
	0.200	0.0125	0.200	0.13614	0.14671	0.15645
	0.200	0.050	0.250	0.13975	0.16809	0.17927
	0.200	0.100	0.300	0.14385	0.19230	0.17343
	0.200	0.138	0.350	0.14790	0.21601	0.11935
	0.200	0.188	0.400	0.14978	0.22808	0.07169
100	0.000	0.000	0.000	0.00331	0.00331	0.00331
	0.200	0.010	0.200	0.14278	0.15977	0.17554
	0.200	0.050	0.250	0.14576	0.17648	0.19381
	0.200	0.110	0.300	0.14989	0.20063	0.18961
	0.200	0.140	0.350	0.15200	0.21251	0.17158
	0.200	0.200	0.400	0.15724	0.24463	0.04702

Table 5: SE of $\hat{\beta}_0$ with midpoint, right and left imputation techniques (interval width of 6)

n	Right CP	Interval CP	CP	Midpoint	Right	Left
20	0.000	0.000	0.000	0.23818	0.23818	0.23818
	0.200	0.000	0.200	0.30620	0.30514	0.37813
	0.200	0.050	0.250	0.30553	0.30338	0.46905
	0.200	0.100	0.300	0.30564	0.30266	0.50808
	0.200	0.150	0.350	0.30266	0.29634	0.78103
	0.200	0.200	0.400	0.30262	0.29557	0.81457
50	0.000	0.000	0.000	0.14396	0.14396	0.14396
	0.220	0.000	0.200	0.19282	0.18904	0.33090
	0.220	0.020	0.250	0.19323	0.18846	0.37896
	0.220	0.080	0.300	0.19339	0.18746	0.46956
	0.220	0.120	0.350	0.19279	0.18651	0.52298
	0.220	0.180	0.400	0.19219	0.18444	0.69824
80	0.000	0.000	0.000	0.11557	0.11557	0.11557
	0.200	0.0125	0.200	0.15432	0.15352	0.16644
	0.200	0.050	0.250	0.15366	0.15163	0.19017
	0.200	0.100	0.300	0.15354	0.15055	0.23578
	0.200	0.138	0.350	0.15271	0.14877	0.30249
	0.200	0.188	0.400	0.15247	0.14797	0.34853
100	0.000	0.000	0.000	0.10525	0.10525	0.10525
	0.200	0.010	0.200	0.13266	0.13226	0.15189
	0.200	0.050	0.250	0.13232	0.13147	0.16822
	0.200	0.110	0.300	0.13209	0.13015	0.19892
	0.200	0.140	0.350	0.13192	0.12943	0.22331
	0.200	0.200	0.400	0.13120	0.12725	0.31581

Table 6: RMSE of $\hat{\beta}_0$ with midpoint, right and left imputation techniques (interval width of 6)

n	Right CP	Interval CP	CP	Midpoint	Right	Left
20	0.000	0.000	0.000	0.23829	0.23829	0.23829
	0.200	0.000	0.200	0.32041	0.32271	0.40078
	0.200	0.050	0.250	0.32057	0.32849	0.49939
	0.200	0.100	0.300	0.32099	0.33095	0.54043
	0.200	0.150	0.350	0.32028	0.34269	0.79045
	0.200	0.200	0.400	0.32051	0.34397	0.82225
50	0.000	0.000	0.000	0.14397	0.14397	0.14397
	0.220	0.000	0.200	0.24279	0.27939	0.35785
	0.220	0.020	0.250	0.24445	0.28824	0.39100
	0.220	0.080	0.300	0.24630	0.30194	0.46957
	0.220	0.120	0.350	0.24676	0.30770	0.52553
	0.220	0.180	0.400	0.24935	0.32878	0.75351
80	0.000	0.000	0.000	0.11561	0.11561	0.11561
	0.200	0.0125	0.200	0.20579	0.21235	0.22843
	0.200	0.050	0.250	0.20771	0.22637	0.26135
	0.200	0.100	0.300	0.21040	0.24422	0.29270
	0.200	0.138	0.350	0.21259	0.26228	0.32518
	0.200	0.188	0.400	0.21373	0.27187	0.35583
100	0.000	0.000	0.000	0.10530	0.10530	0.10530
	0.200	0.010	0.200	0.19490	0.20740	0.23214
	0.200	0.050	0.250	0.19686	0.22007	0.25663
	0.200	0.110	0.300	0.19978	0.23914	0.27481
	0.200	0.140	0.350	0.20126	0.24882	0.28161
	0.200	0.200	0.400	0.20479	0.27575	0.31930

From Table 3 and Figure 3, it can be summarized that the RMSE values of $\hat{\beta}_0$ with all imputation techniques increase slightly as the censoring proportions increase when the sample size is fixed, while they decline with the increasing sample sizes for the similar CP. Whereas the RMSE values of $\hat{\beta}_1$ and $\hat{\sigma}$ with the left imputation technique increase with the rising censoring proportions when the sample size is fixed, whereas they fluctuate between different sample sizes. On the contrary, the RMSE values of $\hat{\beta}_1$ and $\hat{\sigma}$ using the midpoint and right imputation methods decline slightly as the censoring proportions increase when the sample size is fixed. In addition, the greater the sample sizes, the smaller the RMSE of $\hat{\beta}_1$ and $\hat{\sigma}$ with the midpoint and right imputation for the same overall censoring proportions.

Theoretically, the SE and RMSE of a parameter estimate should be increasing when censoring proportions are increasing. This is because more censored data cause the estimated parameters to deviate more from the actual value which result in higher SE and RMSE. However, the simulation study produced rather inconsistent results due to the mixed case interval censored data, especially when dealing with the log logistic model which is known to produce slightly erratic results.

According to the findings presented in Table 4-6 and Figure 4-6, the interval width of 6 yields greater bias, SE, and RMSE values for $\hat{\beta}_0$ compared to the interval width of 4. This pattern is consistent for both $\hat{\beta}_1$ and $\hat{\sigma}$ as well. Thus, it can be concluded that the smaller interval width works better than the larger interval width with smaller values of bias, SE, and RMSE.

For the simulation study of the log logistic model with covariate, uncensored, right, and interval censored data for various sample sizes, censoring proportions, and interval widths, the performance of the estimate $\hat{\beta}_0$ is the best with the midpoint imputation technique while the estimates $\hat{\beta}_1$ and $\hat{\sigma}$ perform the best when the right imputation technique is employed. Thus, the right imputation is identified as the overall preferred method since it works better for most parameter estimates and it will be employed in the real data analysis later.

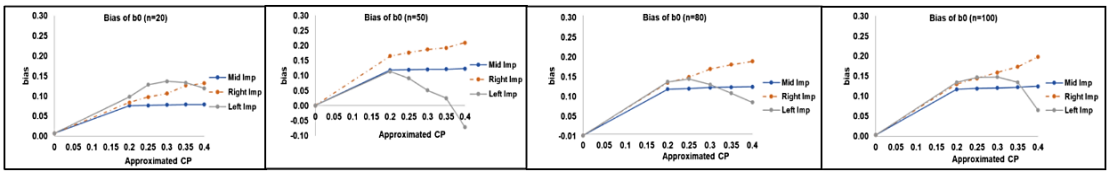

Figure 1: Line plot of bias of $\hat{\beta}_0$ for midpoint, right and left imputation techniques (interval width of 4)

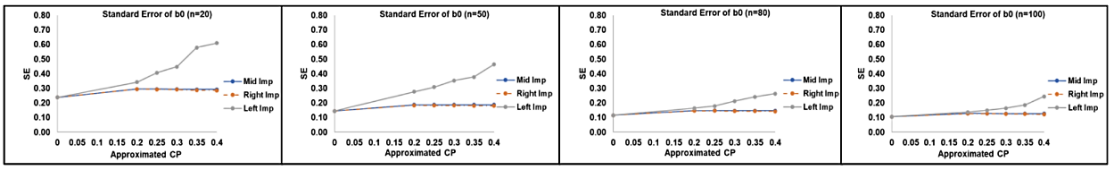

Figure 2: Line plot of SE of $\hat{\beta}_0$ for midpoint, right and left imputation techniques (interval width of 4)

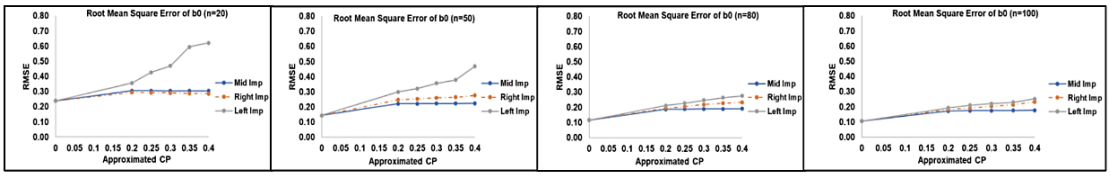

Figure 3: Line plot of RMSE of $\hat{\beta}_0$ for midpoint, right and left imputation techniques (interval width of 4)

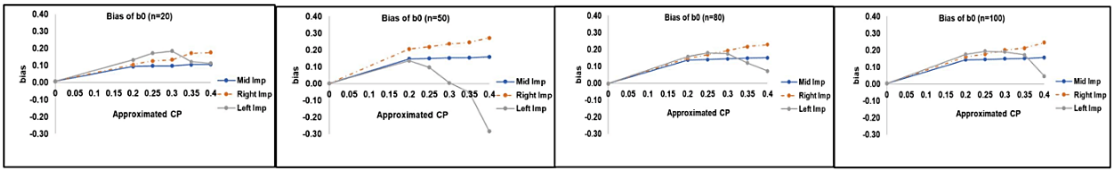

Figure 4: Line plot of bias of $\hat{\beta}_0$ for midpoint, right and left imputation techniques (interval width of 6)

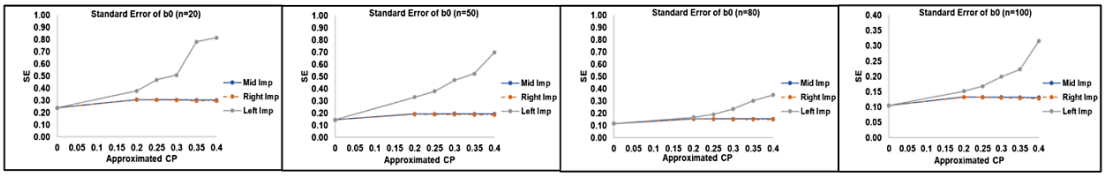

Figure 5: Line plot of SE of $\hat{\beta}_0$ for midpoint, right and left imputation techniques (interval width of 6)

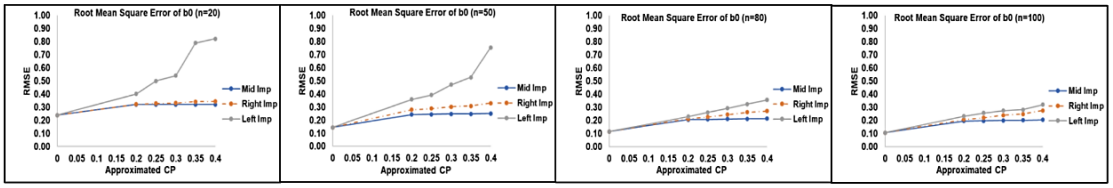

Figure 6: Line plot of RMSE of $\hat{\beta}_0$ for midpoint, right and left imputation techniques (interval width of 6)

4. Real Data Analysis

4.1. *Introduction*

In this research, the proposed log logistic regression model was fitted to breast cancer data. This is a real dataset from Finkelstein and Wolfe (1985). 94 individuals from a retrospective study who were early breast cancer patients receiving treatment at the Joint Center for Radiation Therapy in Boston between 1976 and 1980 are included in this dataset. According to Singh and Totawattage (2013), the purpose of having the retrospective study was to compare the time to cosmetic deterioration between two groups of breast cancer patients. The first group consisted of patients who underwent radiotherapy as their sole treatment (coded as 0), while the second group included patients who received primary radiation therapy followed by chemotherapy (coded as 1). The type of treatment is the covariate in this dataset.

Several parameters were employed to assess the deterioration of the cosmetic appearance. These parameters included breast edema, breast retraction, telangiectasia, arm edema, and overall cosmetic appearance (Beadle *et al.*1984). The development of moderate or severe breast retraction was highly correlated with a fair or poor cosmetic outcome. Thus, the time to cosmetic deterioration can also be determined by the time until the presence of breast retraction.

Out of 94 observations, there are 46 patients treated with radiotherapy only and 48 patients had both radiotherapy and chemotherapy. The patients were monitored every 4 to 6 months, with each of them having various clinic visit times from one another. The time interval between visits is also different from patient to patient because several of the patients missed the scheduled visitation during the study period.

The time to the occurrence of the event of interest in this dataset is the time to cosmetic deterioration. There are 38 right censored observations in total where the patients did not experience breast retraction at the end of the study period. Another 56 patients experienced

breast retraction between the left endpoints, which corresponded to their last clinic visit prior to the appearance of breast retraction, and the right endpoints, which indicated the first clinic visit where the breast retraction was identified. These observations are classified as interval censored data since their exact event time is unknown.

4.2. *Preliminary analysis*

Table 7 illustrates the descriptive statistics of time to cosmetic deterioration based on treatment 0 and treatment 1. It is clearly shown that the median time of cosmetic deterioration for the treatment of radiotherapy alone is higher than the median time for radiotherapy followed by chemotherapy. Therefore, breast cancer patients who received radiotherapy and chemotherapy experienced earlier breast deterioration compared to those who received radiotherapy only.

Table 7: Descriptive statistics of time to cosmetic deterioration by treatment

Treatment	*n*	Events	Median	0.95LCL	0.95UCL
0	46	21	44	35	NA
1	48	35	26	23	35

4.3. *Non-parametric techniques*

Before the parametric log logistic model is fitted to the real dataset, a non-parametric log rank test will be carried out first to verify whether the treatment has a significant effect on survival time as a covariate in this dataset.

Table 8: Log rank test statistics of treatment

$H_0: S_1(t) = S_2(t); H_0: S_1(t) \neq S_2(t)$

	n_{1j}	d_{1j}	e_{1j}	$\frac{(d_{1j} - e_{1j})^2}{e_{1j}}$	$\frac{(d_{1j} - e_{1j})^2}{v_{1j}}$
treatment = 0	46	21	30.8	3.14	7.78
treatment = 1	48	35	25.2	3.85	7.78
Total	94	56	56		

$$\chi^2_{LR} = 7.78 \sim \chi^2_{(1,0.05)}, \qquad p = 0.005$$

Since chi-square value is 7.78, which is greater than the critical value, $\chi^2_{(1,0.05)} = 3.841$, and *p*-value is 0.005, which is smaller than the default level of significance, $\alpha = 0.05$. Therefore, we reject the null hypothesis. There is sufficient evidence to conclude that there is significant difference between the survival function of the two treatment groups. This non-parametric test was conducted because it is not based on any parametric distribution and it is good to see the comparison of the results between the non-parametric approach and the parametric approach. Therefore, we proceed to use the parametric log logistic model to fit the breast cancer data.

4.4. *Parametric techniques*

Figure 7 shows the probability plot for survival time where the right imputation technique was used in approximating the interval censored observations in the breast cancer data. The

visualisation of a probability plot can be used to check if the log logistic model can fit the data well before proceeding to real data analysis. The correlation coefficient for the log logistic model is 0.988, which is very high. All the points fall approximately on the straight line or close to the line on the log logistic probability plot. This indicated that the log logistic model is appropriate to be used in fitting the breast cancer data and it would be a good choice when running parametric distribution analysis.

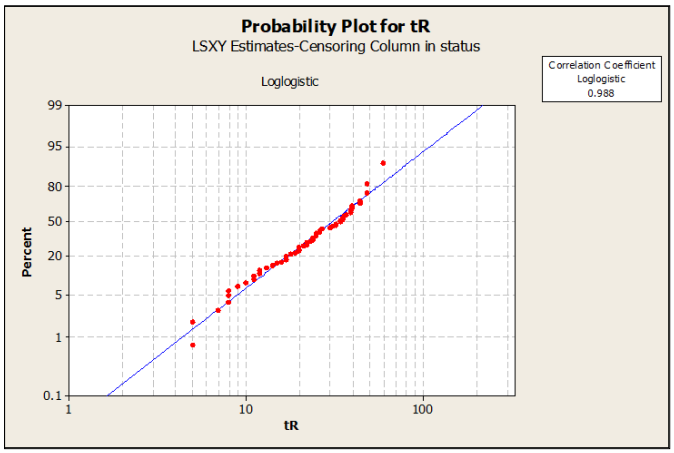

Figure 7: Probability plot of log logistic model

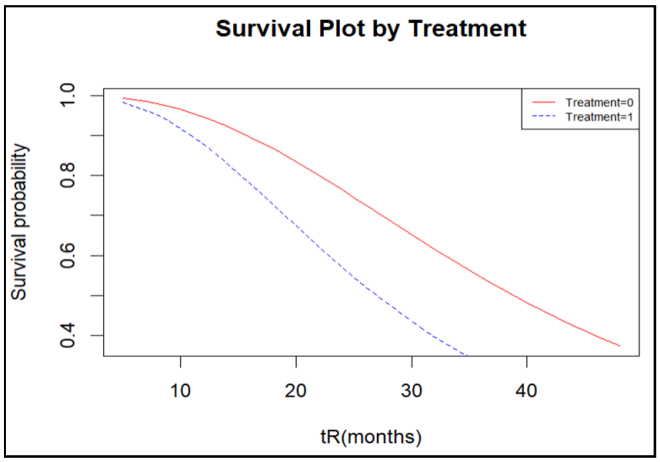

Figure 8: Survival plot by treatment

Figure 8 illustrates that the survival probabilities for breast cancer patients with treatment of radiotherapy only and with the combination of radiotherapy and chemotherapy decrease as the time increases. Particularly, patients who received radiotherapy alone have greater survival rates than those who received radiotherapy and chemotherapy together.

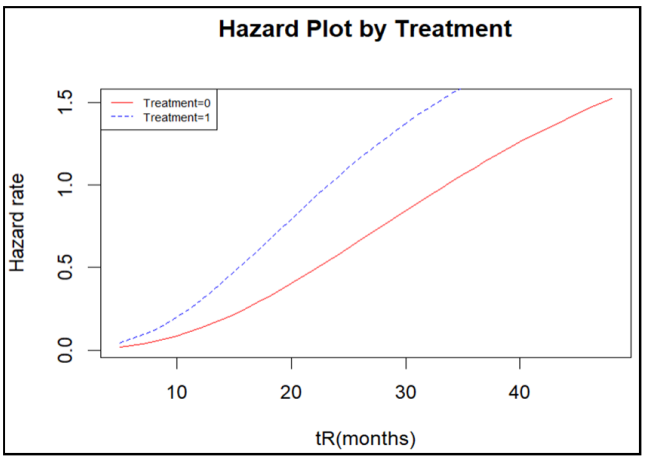

Figure 9: Hazard plot by treatment

Figure 9 shows that the hazard rate for breast cancer patients receiving both radiotherapy and chemotherapy is higher than the hazard rate for those receiving treatment radiotherapy all the time. Thus, it is not recommended that radiotherapy and chemotherapy be given together as a treatment for breast cancer patients to minimise their cosmetic deterioration.

Table 9: Parameter estimation of the log logistic regression model

	Coefficient	SE	z	p	95% CI Lower	95% CI Upper
Intercept($\hat{\beta}_0$)	3.659	0.124	29.42	$< 2e^{-16}$	3.416	3.902
Treatment ($\hat{\beta}_1$)	-0.366	0.161	-2.28	0.023	-0.682	-0.050
Log (scale)	-0.891	0.112	-7.94	$2.1e^{-15}$		
Scale($\hat{\sigma}$)	0.410					

Wald hypothesis test is then conducted to check if the treatment has a significant effect on survival time. β_1 represents the covariate of treatment in breast cancer data. The null hypothesis and alternative hypothesis are given as follows,

$$H_0\colon\ \beta_1 = 0; H_0\colon\ \beta_1 \neq 0$$

As given by Table 9, Wald test statistic, z is -2.28 and p -value is 0.0223. At the default level of significance, $\alpha = 0.05$, the critical region will be $Z \geq 1.96$ and $Z \leq -1.96$. Since p -value is smaller than $\alpha = 0.05$, and Wald test statistic, z is smaller than -1.96, H_0 is rejected. Therefore, there is sufficient evidence to conclude that β_1 is significant and this indicates that treatment has significant effect on the time to cosmetic deterioration. The 95% confidence interval for β_1 is given by,

$$\hat{\beta}_1 \pm z_{1-\frac{\alpha}{2}}\ \widehat{s.e}(\hat{\beta}_1) = (-0.68156, -0.05044\ ) \tag{18}$$

$\beta_1 = 0$ is not included in the confidence interval. This indicates that the β_1 is significant and the covariate of treatment has significant effect on the time to cosmetic deterioration.

Moreover, the median time ratio for treatment can also be obtained. First, the median lifetime, t_m need to be calculated.

$$t_m = \frac{1}{e^{-\beta_0-\beta_1 x}} = \left(e^{-\beta_0-\beta_1 x}\right)^{-1} \tag{19}$$

$$\widehat{TR}(\text{Treatment} = 1, \text{Treatment} = 0) = e^{\widehat{\beta}_1} = 0.6935 \tag{20}$$

The estimated median lifetime for patients receiving both radiotherapy and chemotherapy is 0.6935 times of the median lifetime for patients receiving radiotherapy only, indicating that patients with the treatment of radiotherapy and chemotherapy have shorter lifetime compared to patients with the treatment of radiotherapy alone.

The 95%confidence interval for Median Time Ratio (TR) is given by,

$$0.5058 \leq TR \leq 0.9508 \tag{21}$$

Therefore, we are 95% confident that the median time ratio of patients with the treatment of radiotheraphy and chemotheraphy to patients with treatment of radiotheraphy only is in between 0.5058 to 0.9508 months.

In addition, odds ratio also can be computed for log logistic regression model. Firstly, odds of survival at time, t is given by,

$$\frac{S(z)}{1-S(z)} = e^{-z} \tag{22}$$

The odds ratio is shown as follows (Hosmer & Lemeshow 1999),

$$\widehat{OR}(\text{Treament} = 1, \text{Treatment} = 0) = e^{\frac{\widehat{\beta}_1}{\hat{\sigma}}} = 0.4096 \tag{23}$$

The odds of survival at time t for patients with the treatment of radiotherapy and chemotherapy is 0.4096 to that of patients with the treatment of radiotherapy only. In other words, patients with the treatment of radiotherapy and chemotherapy have smaller odds of survival compared to patients with the treatment of radiotherapy only.

5. Conclusion

In this research, the log logistic model was extended to incorporate a covariate with uncensored, right, and interval censored data. The simulation study was conducted at various sample sizes, censoring proportions, and interval widths to compare the performance of the parameter estimates based on their bias, SE, and RMSE values. The smaller the three values, the better the performance of the parameter estimates in terms of accuracy and efficiency.

Based on the results of the simulation study consisting of a covariate, uncensored, right, and interval censored observations for various sample sizes, censoring proportions, and interval widths, the right imputation is the best overall method since it performs better for most parameter estimates. Moreover, the results of the simulation study also showed that the narrower interval width works better than the wider interval width with lower bias, SE and RMSE in the procedure of parameter estimation. The simulation overall conclusion is that the SE and RMSE values of the parameter estimates decrease as sample sizes increase while they increase as censoring proportions increase, which indicates that bigger sample sizes and smaller censoring proportions yield better parameter estimates.

In the real data analysis, it indicated that the log logistic model fits the breast cancer data well. The covariate of treatment has a significant effect on the time to cosmetic deterioration

since the parameter β_1 is significant. Breast cancer patients who had the treatment of radiotherapy only could survive longer than those who received the combination of radiotherapy and chemotherapy.

References

Arasan J. & Adam M.B. 2014. Double bootstrap confidence interval estimates with censored and truncated data. *Journal of Modern Applied Statistical Methods* **13**(2): 399-419.

Beadle G.F., Silver B., Botnick L., Hellman S. & Harris J.R. (1984). Cosmetic results following primary radiation therapy for early breast cancer. *Cancer* **54**(12): 2911–2918.

Cox D.R. & Hinkley D.V. 1979. *Theoretical statistics*. 1st Ed. New York: CRC Press.

Finkelstein D.M. & Wolfe R.A. 1985. A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics* **41**(4): 933–945.

Hosmer D.W. & Lemeshow S. 1999. *Applied Survival Analysis, Regression Modeling of Time to Event Data*. New York: John Wiley and Sons.

Lai M.C. & Arasan J. 2020. Single covariate log-logistic model adequacy with right and interval censored data. *Journal of Quality Measurement and Analysis* **16**(2): 131–140.

Loh Y.F., Arasan J., Midi H. & Bakar M.R.A. 2017. A parametric model for doubly interval censored lifetime data. *Journal of Science and Technology* **9**(2): 10-16.

Muse A.H., Mwalili S.M. & Ngesa O. 2021. On the log-logistic distribution and its generalizations: A survey. *International Journal of Statistics and Probability* **10**(3): 93-125.

Singh R.S. & Totawattage D.P. 2013. The statistical analysis of interval-censored failure time data with applications. *Open Journal of Statistics* **3**(2): 155-166.

Smith T. & Smith B. 2001. Survival analysis and the application of cox's proportional hazards modeling using SAS. In *Proceedings of the Twenty Sixth Annual SAS Users Group International Conference*.

Department of Mathematics and Statistics
Faculty of Science
Universiti Putra Malaysia
43400 UPM Serdang
Selangor DE, MALAYSIA
E-mail: yuanxinteea@gmail.com[*], *jayanthi@upm.edu.my*

Received: 5 May 2023
Accepted: 29 August 2023

[*]Corresponding author