Applied Mathematics and Computational Intelligence
Volume 13, No. 3, 2024 [115 – 142]

Left, Right, Midpoint and Random Point Imputation Techniques for Weibull Regression Model with Right and Interval-Censored Data

Ahmad Kabeer Bin Naushad Ali[1*] and Jayanthi Arasan[2]

[1,2]Department of Mathematics and Statistics, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia.

* Corresponding author: kabeernaushad2711@gmail.com

Received: 27 October 2023
Revised: 26 February 2024
Accepted: 12 August 2024

ABSTRACT

The research explores several imputation techniques, namely left, right, midpoint and random imputations for the MLE of the Weibull regression model with covariate for uncensored, right, and interval-censored data. A simulation study is conducted to obtain the parameter estimates of the model with different imputation techniques, sample sizes, and censoring proportions and its performance are evaluated using bias, standard error (SE), and root mean square error (RMSE). The simulation result indicates that midpoint imputation technique outperformed other techniques based on the lowest RMSE values. Finally, the model was fitted to diabetic nephropathy data using selected imputation techniques. The result concluded that the Weibull regression model may provide a good fit to the data and that the covariate, gender has a significant effect on the survival time of patient kidneys.

Keywords: covariate, imputation techniques, right-censored, interval-censored, Weibull.

1 INTRODUCTION

Survival analysis, as Smith et al. [1] describe, is all about studying the time between starting an observation and an event happening. Lee and Wang [2] point out that in survival analysis, we measure how long it takes for a particular event to occur, which is a key part of this field. Initially, survival analysis focused on predicting how long things or individuals would last and comparing survival experiences between different groups. Researchers typically use two statistical methods, parametric and non-parametric, to analyze this kind of data. Dealing with missing data is a common challenge in research, and researchers employ techniques, such as imputation, to estimate values that are missing from their datasets, as emphasized by Schafer and Graham [3]. Lai et al. [4] stated that the Weibull distribution is the best lifetime distribution model and describes observed failures of many phenomena and components .

The Weibull distribution plays a central role in survival analysis. When it comes to estimating its parameters, the most common and straightforward method is Maximum Likelihood Estimation (MLE). Cohen [5] provided valuable insights into using MLE for the Weibull Distribution, covering

scenarios involving both censored and complete data. Researchers like Stone and Van [6] have underlined MLE as the preferred method for estimating Weibull Distribution parameters. In a study by Odell et al. [7], MLE's superiority became evident, particularly in the context of the Weibull-based accelerated failure time regression model, especially with larger sample sizes, and when dealing with left or interval-censored data. Guure et al. [8] reinforced the reliability of MLE, especially when estimating scale parameters. Strapasson [9] conducted a comprehensive study comparing different techniques for estimating Weibull Distribution and Weibull Regression Model parameters, including midpoint, lower, and upper limit imputations. Salahaddin [10] found that using the rank regression method is a highly effective way to estimate Weibull Distribution parameters, especially when working with time series wind data. Zhang [11] provided guidance on fitting the Weibull Regression Model using R software. Zyoud et al. [12] conducted an analysis of the parametric Cox model with partly interval-censored data, employing various imputation techniques to handle the challenges posed by missing data. Lai and Arasan [13] focused on estimating Log-Logistic Model parameters using MLE, emphasizing the importance of achieving the lowest Standard Error (SE) and Root Mean Square Error (RMSE) values for optimal parameter estimates. Khairunnisa et al. [14] applied the Weibull regression model with MLE to analyze factors affecting the survival and recovery rates of Covid-19 patients, underscoring the significance of comorbidities. Kiani and Arasan [15] analyzed the Gompertz model with fixed and time-dependent covariates, revealing a decrease in bias, SE, and RMSE values with increasing study period, attendance probability, and sample size.

The Weibull regression model hasn't received as much attention when it comes to dealing with right-censoring and interval-censoring. Most previous research has primarily focused on partial interval-censoring. Additionally, it's unclear which methods work best for estimating the model's parameters in these scenarios. While some studies have used imputation techniques for partly censored data, they haven't specifically examined how effective these techniques are for interval-censored data. For example, Alharpy and Ibrahim [16] used multiple imputation techniques for data that's partly interval-censored and follows the Weibull distribution. Similarly, Saeed and Elfaky [17] conducted research on the Parametric Weibull Model using imputation techniques for partly censored data. To fill this research gap, our study aims to explore the Weibull Regression Model's applicability to situations involving both right-censoring and interval-censoring. We will investigate the effectiveness of imputation techniques in handling such data.

This research aims to achieve three main objectives. Firstly, incorporation of a covariate into the Weibull regression model with uncensored, right and interval censored data, and obtain its maximum likelihood estimation (MLE) via simulation study at various sample sizes, n and censoring proportions, cp using bias, standard error (SE) and root mean square error (RMSE). Secondly, the performance of four imputation techniques (midpoint, random, right, and left) will be compared for dealing with uncensored, right and interval censored data at different sample sizes, censoring proportions, and right censoring proportions. The best imputation technique will be identified based on the RMSE. Lastly, the Weibull regression model will be fitted to real-life right and interval censored data with covariate, and the model's performance will be evaluated using the chosen imputation technique. The results of this research will contribute to a better understanding of the Weibull regression model with censored data and its applications in real-world situations.

2 METHODOLOGY

2.1 The Weibull Regression Model

The Probability Density Function (PDF), $f(t)$, survivorship function, $S(t)$ and the Cumulative Distribution Function (CDF), $F(T)$ of Weibull regression model can be expressed as Equation (1) to (5) stated below :

$$f(t) = \frac{1}{\sigma} e^{\frac{y-\mu}{\sigma} - exp(\frac{y-\mu}{\sigma})} \tag{1}$$

where $\mu = -ln\lambda$ and $\sigma = \frac{1}{\gamma}$; λ is scale parameter and γ is shape parameter;$y = \ln t$; $\mu = \beta_0 + \beta_1 x_1$ and $x_{i0} = 1$, in which it incorporates the effect of single covariate on survival time, where

If $z = \frac{y-\beta' x}{\sigma}$,

$$f(z) = \frac{1}{\sigma} e^{z - exp(z)} \tag{2}$$

$$S(z) = e^{-exp(z)} \tag{3}$$

where $-\infty < z < \infty$

$$F(z) = 1 - S(z) = 1 - e^{-exp(z)}. \tag{4}$$

The lifetime, t_i can be simulated via the inverse transform technique as shown below,

$$t_i = (-\ln(1 - u_i))^{\sigma} \times e^{\beta' x} \tag{5}$$

where random variable U is uniformly distributed in (0,1)

2.2 Maximum Likelihood Estimation

Suppose that the data is categorized into censored and uncensored for $i = 1, 2, \ldots, n$ observations. The variables s_i and $\tilde{t}_i$ are defined as follows,

$$s_i = \begin{cases} 1, & \text{for } t_i \text{ uncensored ,} \\ 0, & \text{for } t_i \text{ censored.} \end{cases}$$

and,

$$\tilde{t_i} = \begin{cases} \frac{tL_i+tR_i}{2}, & \text{for midpoint imputation ,} \\ U_i(tL, tR), & \text{for random imputation} \\ tR_i, & \text{for right imputation,} \\ tL_i, & \text{for left imputation,} \\ t_i, & \text{otherwise.} \end{cases}$$

The likelihood function is given by,

$$l(\beta) = \prod_{i=1}^{n}([f(\tilde{t_i}, \beta, x_i)]^{s_i} \times [S(t_i, \beta, x_i)]^{1-s_i}),$$

where the $f(t, \beta, x_i)$ and $S(t, \beta, x_i)$ represents the probability density function and survivorship function of the Weibull regression model. The log-likelihood function is as Equation (6),

$$\begin{aligned} L(\beta, \sigma) &= \ln(\prod_{i=1}^{n}([f(\tilde{t_i}, \beta, x_i)]^{s_i}, \times [S(t_i, \beta, x_i)]^{1-s_i})), \\ &= \sum_{i=1}^{n}(s_i(-\ln\sigma + \tilde{z_i} - exp(\tilde{z_i}) + exp(z_i)) - exp(z_i)) \end{aligned} \tag{6}$$

where $z_i = \frac{y-\beta' x_i}{\sigma}$.

The first derivative with respect to β_j is given as Equation (7),

$$\frac{\partial(\beta_j, \sigma)}{\partial\beta_j} = \sum_{i=1}^{n}(\frac{x_{ij}}{\sigma}(s_i(1 + e^{\tilde{z_i}} - e^{z_i}) + e^{z_i})) \tag{7}$$

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

The first derivative with respect to σ is given as Equation (8),

$$\frac{\partial(\beta_j, \sigma)}{\partial\sigma} = \sum_{i=1}^{n}(\frac{s_i}{\sigma}(1 - \tilde{z_i} + \tilde{z_i}e^{\tilde{z_i}} - z_i e^{z_i}) + \frac{z_i e^{z_i}}{\sigma}). \tag{8}$$

Newton-Raphson iterative metohd procedures are used to solve the non-linear equations.The covariance matrix of maximum likelihood estimators are approximated using the Fisher's information matrix (Sccott, [18]).

2.3 Log-Rank Test

The log-rank test is done to test the difference in survival between two or more independent groups. The survival curves are first estimated for each groups first and then compared statistically using the log-rank test. The log-rank test can be approximated using the chi-square test statistic, which is given in Equation (9).

$$\chi^2(LR) = \frac{(\sum_{j=1}^{r} d_{1j} - e_{1j}^2}{\sum_{j=1}^{r} v_{1j}} \sim \chi^2(1) \tag{9}$$

where
$e_{1j} = \frac{n_{1j}d_j}{n_j}$
$v_{1j} = \frac{n_{1j}n_{2j}d_j(n_j-d_j)}{n_j^2(n_j-1)}$

2.4 Wald Confidence Interval

Point estimate is used to construct an interval with a certain confidence interval or probability the interval will contain the true parameter value. The confidence interval contains a range of values restricted by an upper and lower limit for the population parameter.

The maximum likelihood estimator for vector of parameters and the loglikelihood function of the θ is denoted by $\hat{\theta}$ and $l((\theta))$. $\hat{\theta}$ is asymptotically normally distributed with mean θ and covariance matrix $I^{-1}(\theta)$, where $I^{-1}(\theta)$ is the Fisher information matrix evaluated at the true value of the θ (Cox & Hinkley, 1974). The $(\hat{var(\hat{\theta})})$ is the $(j,j)^{th}$ element of $I^{-1}(\theta))$. We will approximate I(θ) with observe information matrix i(θ). The $(j,j)^{th}$ element of $i^{-1}(\theta)$ is the estimate of $var(\hat{\theta_j})$. Thus, the formula above is given in Equation (10).

$$\theta_j = \hat{\theta_j} \pm z_{1-\frac{\alpha}{2}} \sqrt{\hat{i^{-1}(\theta_{jj})}}. \tag{10}$$

3 SIMULATION STUDY

Simulation study is known as computer experiments that use pseudo-random sampling to create data (Maria, [19]) . In this research, a simulation study was carried out to assess the performance of parameter estimates of the Weibull regression model. The simulation was conducted with R software in different combinations of sample sizes and censoring proportions with 1000 replications. The sample sizes were $n = 20, 40, 60, 80$ and 100. Right censored were fixed at approximately 10% and 20%, while the interval-censored is approximated to 0%, 5%, 10%, 15%, 20%, and 25%. There is also a simulation done for 0% censoring proportion which consists of 0% right censoring and 0% interval censoring. Initial values of 1.088, -0.006, and 0.090 were chosen as the parameters for β_0, β_1, and σ respectively. These parameters were actually chosen from a study that identifies the parameter estimates for the Weibull regression model using the HIV data.

A sequence of random numbers from the uniform distribution, $U(0,1)$ is simulated. The covariate, x is generated using a standard normal distribution. These u_i and x_i were used to generate

failure times, t_2 using the inverse transform method. Censoring time, c_2 was generated from the exponential distribution to obtain censored lifetimes. t_2 and c_2 were compared, and the minimum was chosen to be the t_i which is the final failure time. The censoring indicator, c_i is obtained when t_2 is less than c_2 which will return 1 and 0 if otherwise. This indicates that, if t_2 is less than c_2, it is uncensored, and if it is otherwise, then it is censored.

The interval chosen for the simulation study was 0 to 4 months, 4 to 8 months, 8 to 12 months, 12 to 16 months, 16 to 20 months, and 20 to 24 months. We also generate p from the Bernoulli distribution to obtain interval-censored observations. To obtain data that consists of uncensored, right-censored, and interval-censored data, we use the following indicator. The observation will be uncensored if $c = 1$ and $p = 0$, interval-censored if $c = 1$ and $p = 1$, and finally right-censored if $c = 0$ and $p = 0$ or $c = 0$ and $p = 1$. If the observation failure time, t is greater than 24 months, then the observation will automatically become right censored. The R code for the simulation study was shown in the Appendix.

If the observation is interval-censored, the imputed time will be used instead of the observed time. Left imputation time, tL, and right imputation time, tR can be obtained from the interval where we will choose the left endpoint for tL and right endpoint for tR. For example, if the observation fall between 4 to 8 months, it's tL will be 4 and tR will be 8. For midpoint imputation, $tMid$, we will take the average of tL and tR. The value for random imputation, $tRandom$ is obtained by generating random value using Uniform Distribution, $u(tL, tR)$.

The parameter estimates for β_0, β_1, and σ were obtained using the maximum likelihood estimation with imputation, which consists of left, right, midpoint, and random imputation. The values of root mean square error (RMSE) were calculated to compare the performance of parameter estimates at different sample sizes, for different censoring proportions, and for different imputation techniques as well.

Root Mean Square Error (RMSE) is the expected value of the square of the difference between the estimator and the true parameter. Better performance of parameter estimates can be considered when the RMSE is small. The formula of RMSE can be calculated as per the formula given in Equation (11).

$$RMSE = \sqrt{Bias(\hat{\theta})^2 + SE(\hat{\theta})^2} \tag{11}$$

where $Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$ and $SE(\hat{\theta}) = \sqrt{\frac{\sum \hat{\theta_i}^2 - \frac{\sum \hat{\theta_i}^2}{N}}{N-1}}$.

3.1 Results and Discussion

The root mean square error (RMSE) values of parameter estimates will be compared across different sample sizes, censoring proportions, and right-censored proportions (rcp). Tables 1 and 2 display the RMSE results for $\hat{\beta_0}$ estimates using various imputation techniques (midpoint, random, right, and left) at 10% and 20% rcp, and these results are further visualized in Figures 1 and 2.

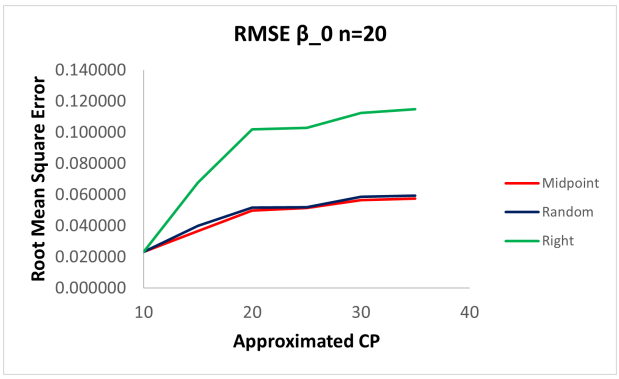
The RMSE for left imputation is notably larger than other methods. To ensure clear comparisons, the midpoint, random, and right imputation techniques are graphed together for different sample sizes while the left imputation technique is separately graphed with varying sample sizes. This separation is due to the left imputation's erratic behavior, resulting in significantly larger values compared to other techniques. It's important to note that all figures in this chapter follow the same plotting style as Figure 1.

Tables 3 and 4 display root mean square error values for $\hat{\beta}_1$ estimates at varying sample sizes and censoring proportions. These values are calculated using different imputation techniques, including midpoint, random, right, and left, with rcp at 10% and 20%. The corresponding analyses are presented in Figures 3 and 4.

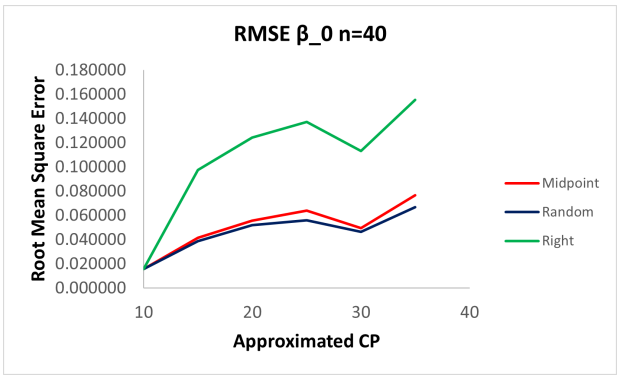
Tables 5 and 6 show root mean square error values for $\hat{\sigma}$ estimates under different sample sizes, censoring proportions, and imputation techniques (midpoint, random, right, and left) with rcp set at 10% and 20%. Figures 5 and 6 provide a visual representation of these results

Table 1 : Root Mean Square Error of $\hat{\beta}_0$ for different imputation techniques when $rcp = 10\%$

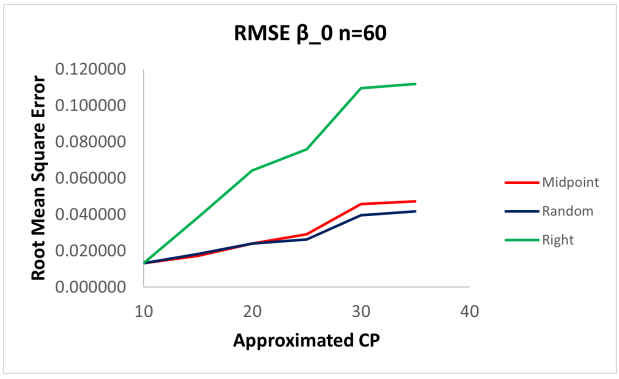
n	CP(%)	Imputation Techniques			
		Midpoint	**Random**	**Right**	**Left**
20	10	0.023265	0.023265	0.023265	0.023265
	15	0.036666	0.039840	0.067687	0.467256
	20	0.049688	0.051762	0.101997	0.860986
	25	0.051304	0.051839	0.102874	0.898541
	30	0.056434	0.058530	0.112329	1.039923
	35	0.057481	0.059276	0.114913	1.089045
40	10	0.015540	0.015540	0.015540	0.015540
	15	0.041465	0.038641	0.097548	0.792383
	20	0.055461	0.051927	0.124129	1.214502
	25	0.064028	0.055874	0.137141	1.584922
	30	0.049440	0.046242	0.113198	0.943159
	35	0.076626	0.066682	0.155292	2.138894
60	10	0.013278	0.013278	0.013278	0.013278
	15	0.017829	0.018346	0.038355	0.115052
	20	0.024088	0.024072	0.064299	0.238725
	25	0.029158	0.026285	0.075864	0.264653
	30	0.045763	0.039578	0.109621	0.926966
	35	0.047310	0.041670	0.111933	1.009733
80	10	0.011458	0.011458	0.011458	0.011458
	15	0.022639	0.021689	0.061306	0.183470
	20	0.024770	0.023113	0.067392	0.147127
	25	0.036529	0.031744	0.093817	0.506495
	30	0.042349	0.034791	0.108342	0.490999
	35	0.049516	0.039311	0.123337	0.688577
100	10	0.010001	0.010001	0.010001	0.010001
	15	0.022411	0.020671	0.062900	0.118171
	20	0.030793	0.026759	0.085531	0.272705
	25	0.032192	0.027792	0.088336	0.305988
	30	0.045466	0.036851	0.113285	0.667625
	35	0.055169	0.042662	0.131587	1.035322



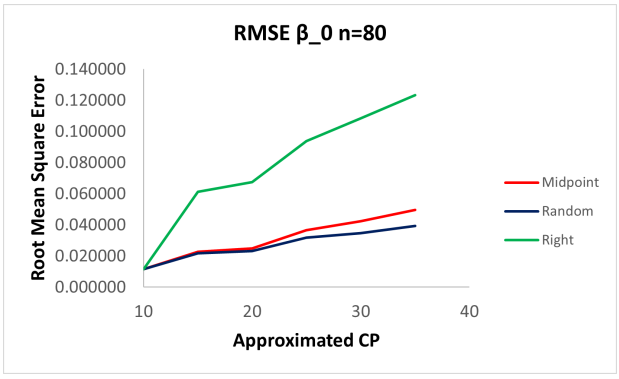
(a) n=20



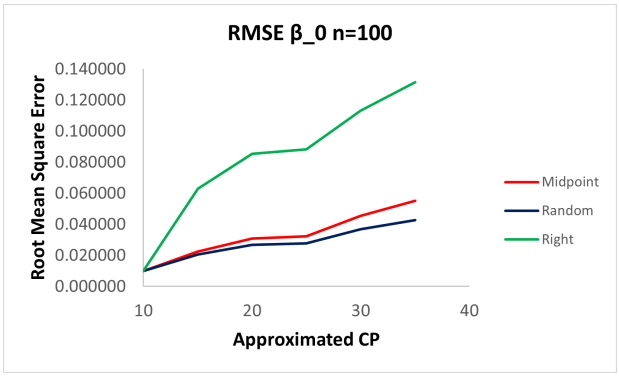
(b) n=40



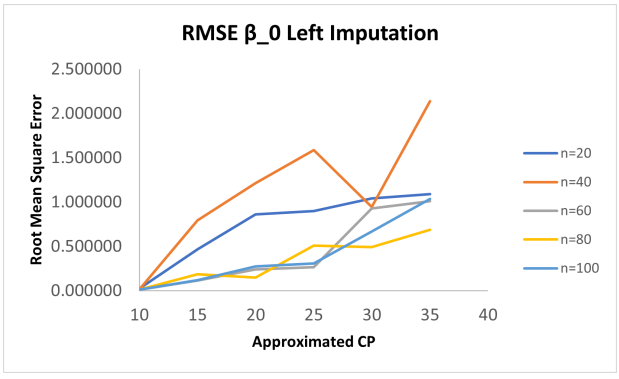
(c) n=60



(d) n=80



(e) n=100

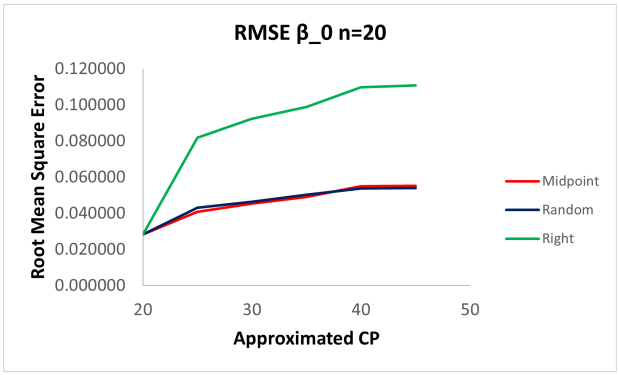


(f) Left Imputation for n=20, 40, 60, 80, 100

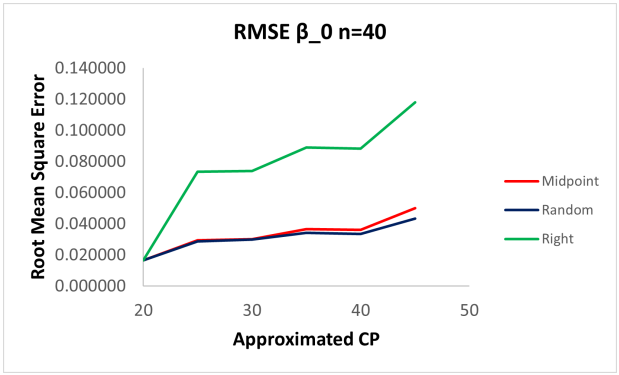
Figure 1 : RMSE of $\hat{\beta}_0$ at $n = 20, 40, 60, 80, 100$ when $rcp = 10\%$.

Table 2 : Root Mean Square Error of $\hat{\beta}_0$ for different imputation techniques when $rcp = 20\%$

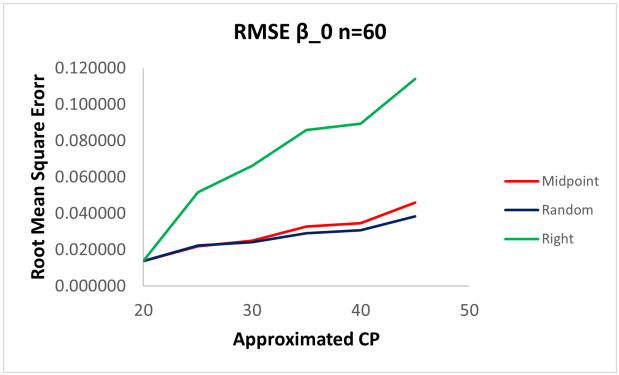
n	CP(%)	Imputation Techniques			
		Midpoint	**Random**	**Right**	**Left**
20	20	0.028220	0.028220	0.028220	0.028220
	25	0.040827	0.043033	0.081865	0.706017
	30	0.045304	0.046258	0.092289	0.716387
	35	0.048934	0.050131	0.098851	0.790845
	40	0.054842	0.053762	0.109649	0.939213
	45	0.055127	0.053989	0.110665	0.946278
40	20	0.016488	0.016488	0.016488	0.016488
	25	0.029421	0.028700	0.073442	0.306330
	30	0.030140	0.029849	0.073971	0.315526
	35	0.036538	0.034201	0.088974	0.420535
	40	0.036065	0.033442	0.088203	0.419782
	45	0.050099	0.043308	0.117905	0.709074
60	20	0.013637	0.013637	0.013637	0.013637
	25	0.021893	0.022268	0.051626	0.170377
	30	0.024886	0.024161	0.066090	0.217069
	35	0.032819	0.029051	0.085800	0.302740
	40	0.034664	0.030688	0.089302	0.393485
	45	0.045949	0.038317	0.113995	0.493728
80	20	0.011878	0.011878	0.011878	0.011878
	25	0.020865	0.020350	0.054834	0.199266
	30	0.028580	0.024681	0.079056	0.195099
	35	0.033485	0.029160	0.088283	0.260439
	40	0.044414	0.034450	0.113489	0.411963
	45	0.058523	0.045592	0.137947	0.958640
100	20	0.010357	0.010357	0.010357	0.010357
	25	0.013733	0.013994	0.032876	0.078030
	30	0.020382	0.019474	0.058644	0.162151
	35	0.028855	0.024515	0.081129	0.231814
	40	0.047098	0.037178	0.116566	0.737179
	45	0.047265	0.036251	0.120512	0.414955



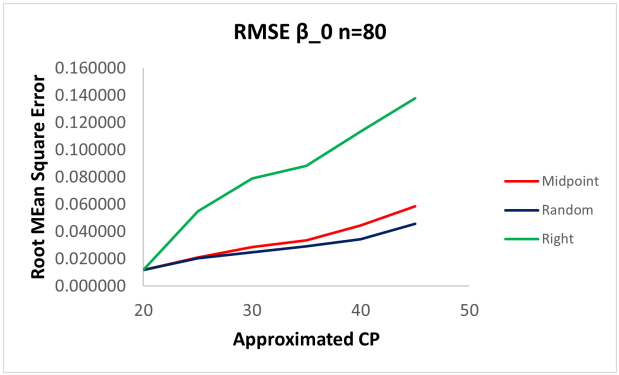
(a) n=20



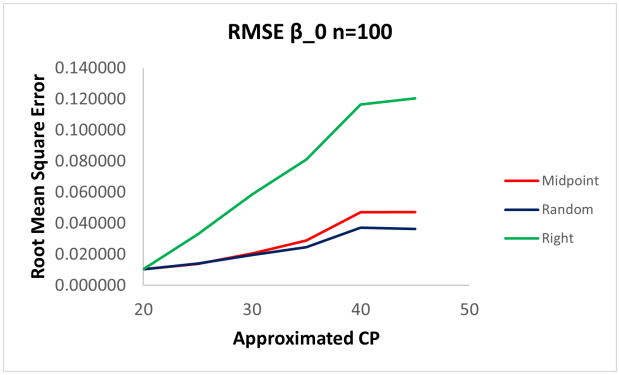
(b) n=40



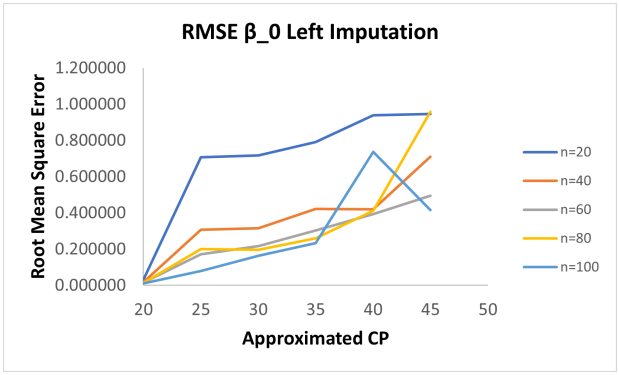
(c) n=60



(d) n=80



(e) n=100

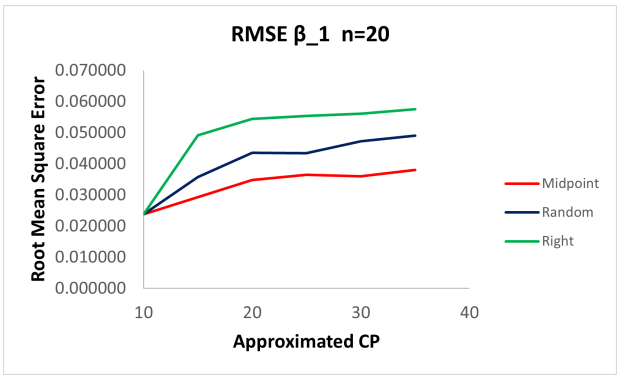


(f) Left Imputation for n=20, 40, 60, 80, 100

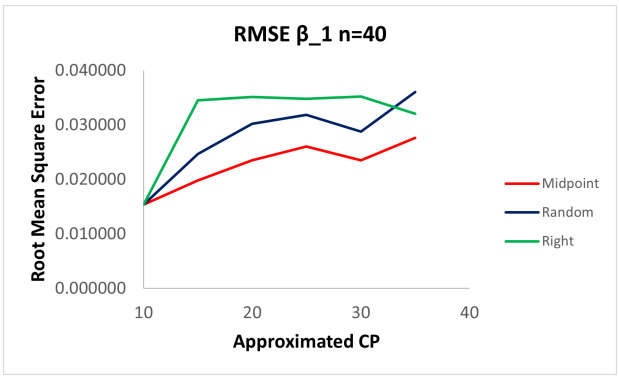
Figure 2 : RMSE of $\hat{\beta}_0$ at $n = 20, 40, 60, 80, 100$ when $rcp = 20\%$.

Table 3 : Root Mean Square Error of $\hat{\beta}_1$ for different imputation techniques when $rcp = 10\%$

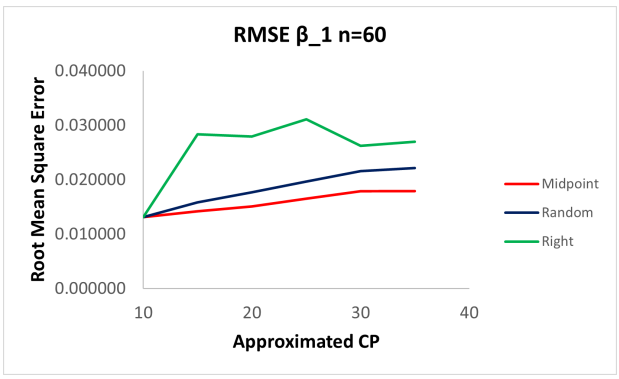
n	CP(%)	Imputation Techniques			
		Midpoint	**Random**	**Right**	**Left**
20	10	0.023771	0.023771	0.023771	0.023771
	15	0.029306	0.035762	0.049155	0.479243
	20	0.034800	0.043477	0.054414	1.037234
	25	0.036455	0.043435	0.055397	1.114783
	30	0.035954	0.047266	0.056075	1.219540
	35	0.038042	0.049052	0.057489	1.185579
40	10	0.015336	0.015336	0.015336	0.015336
	15	0.019769	0.024618	0.034491	0.509276
	20	0.023513	0.030159	0.035080	0.801801
	25	0.026003	0.031801	0.034787	0.923060
	30	0.023481	0.028728	0.035164	0.643641
	35	0.027571	0.035979	0.032048	1.135075
60	10	0.013108	0.013108	0.013108	0.013108
	15	0.014184	0.015802	0.028372	0.085742
	20	0.015049	0.017678	0.027934	0.186075
	25	0.016528	0.019633	0.031127	0.283278
	30	0.017868	0.021556	0.026205	0.443372
	35	0.017901	0.022162	0.026957	0.477932
80	10	0.011382	0.011382	0.011382	0.011382
	15	0.012539	0.014255	0.023551	0.144676
	20	0.013284	0.015658	0.024238	0.181290
	25	0.014433	0.017685	0.024510	0.312459
	30	0.016226	0.019647	0.024597	0.427711
	35	0.016705	0.021308	0.024022	0.508635
100	10	0.009834	0.009834	0.009834	0.009834
	15	0.011461	0.013580	0.022755	0.141416
	20	0.012291	0.014701	0.021801	0.239265
	25	0.012627	0.015027	0.021612	0.250376
	30	0.014314	0.017698	0.020343	0.370032
	35	0.015164	0.018072	0.020350	0.478557



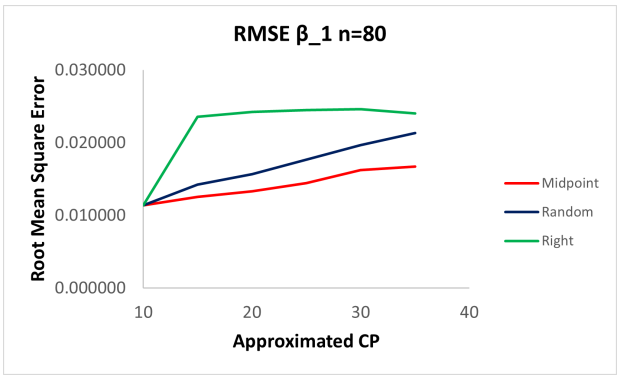
(a) n=20



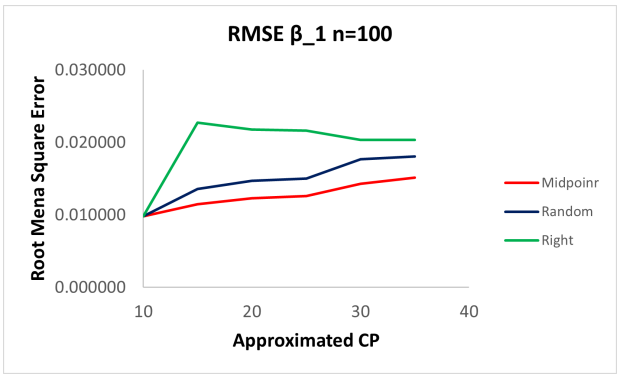
(b) n=40



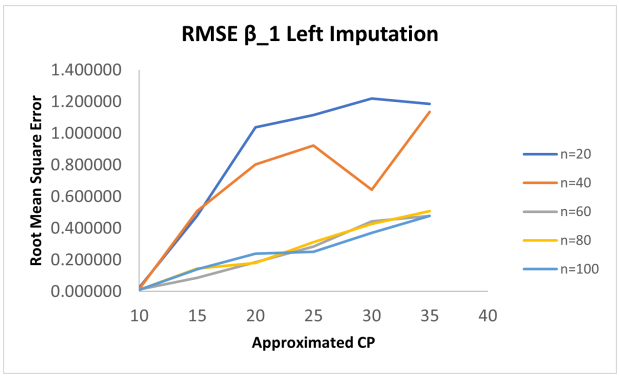
(c) n=60



(d) n=80



(e) n=100

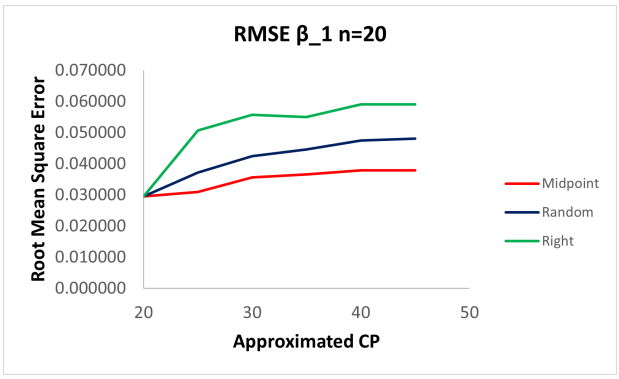


(f) Left Imputation for n=20, 40, 60, 80, 100

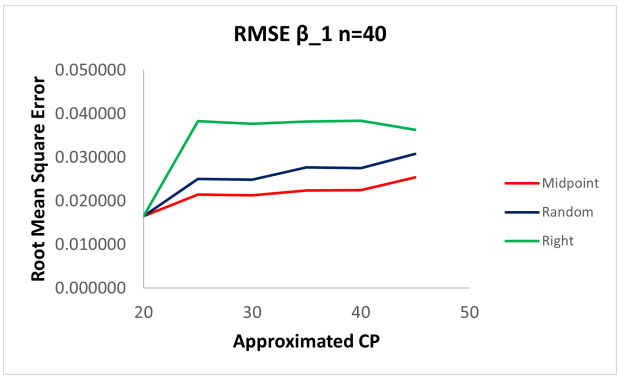
Figure 3 : RMSE of $\hat{\beta}_1$ at $n = 20, 40, 60, 80, 100$ when $rcp = 10\%$.

Table 4 : Root Mean Square Error of $\hat{\beta}_1$ for different imputation techniques when $rcp = 20\%$

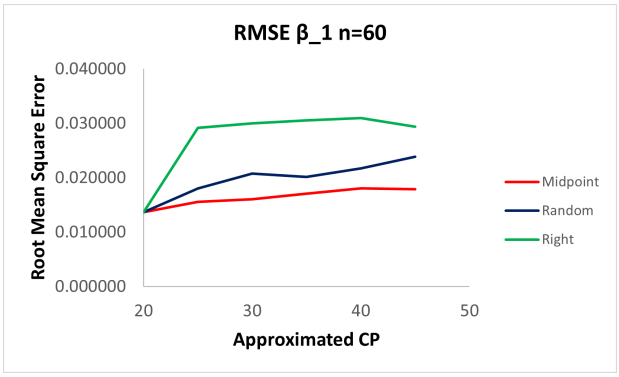
n	CP(%)	Imputation Techniques			
		Midpoint	**Random**	**Right**	**Left**
20	20	0.029535	0.029535	0.029535	0.029535
	25	0.030869	0.037102	0.050623	0.694886
	30	0.035643	0.042466	0.055670	0.948134
	35	0.036506	0.044577	0.054927	1.034325
	40	0.037825	0.047476	0.058991	1.160445
	45	0.037888	0.048006	0.059065	1.177668
40	20	0.016503	0.016503	0.016503	0.016503
	25	0.021485	0.025039	0.038258	0.36940
	30	0.021292	0.024868	0.037654	0.367451
	35	0.022429	0.027726	0.038180	0.560624
	40	0.022499	0.027535	0.038361	0.549789
	35	0.025397	0.030817	0.036295	0.831870
60	20	0.013629	0.013629	0.013629	0.013629
	25	0.015536	0.017997	0.029182	0.150926
	30	0.016003	0.020749	0.029993	0.238335
	35	0.017084	0.020129	0.030510	0.332034
	40	0.018052	0.021705	0.030927	0.388232
	45	0.017869	0.023836	0.029399	0.561630
80	20	0.011951	0.011951	0.011951	0.011951
	25	0.013147	0.015397	0.025753	0.153061
	30	0.013660	0.016381	0.024606	0.232286
	35	0.014050	0.016979	0.025599	0.284361
	40	0.016179	0.020093	0.024262	0.488222
	45	0.016814	0.021716	0.022603	0.629927
100	20	0.010356	0.010356	0.010356	0.010356
	25	0.011193	0.012414	0.021729	0.054464
	30	0.011457	0.013387	0.023060	0.138863
	35	0.012537	0.015578	0.023601	0.247609
	40	0.014522	0.017610	0.021063	0.394549
	45	0.014879	0.017990	0.021932	0.474685



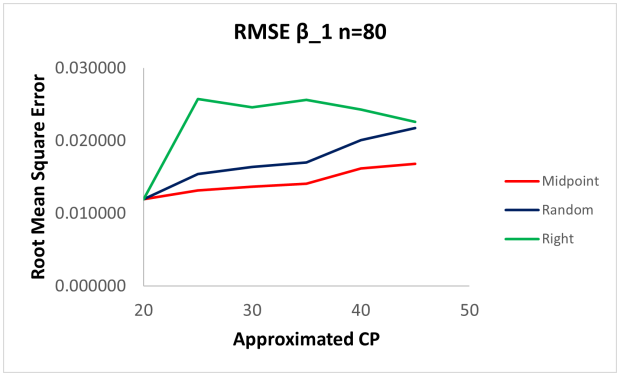
(a) n=20



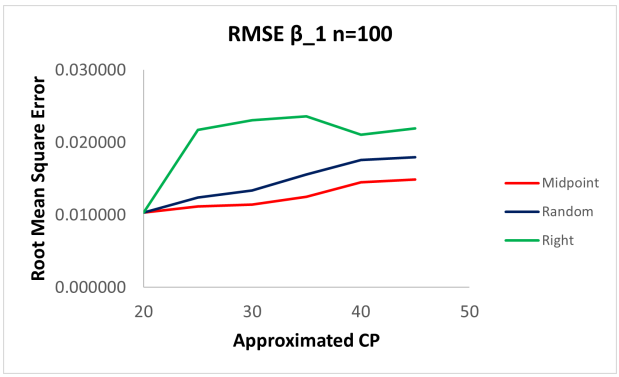
(b) n=40



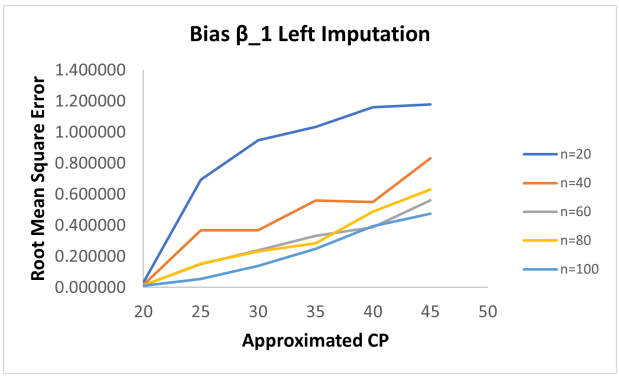
(c) n=60



(d) n=80



(e) n=100

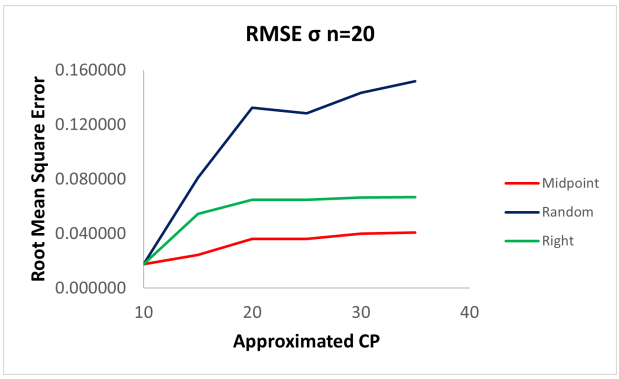


(f) Left Imputation for n=20, 40, 60, 80, 100

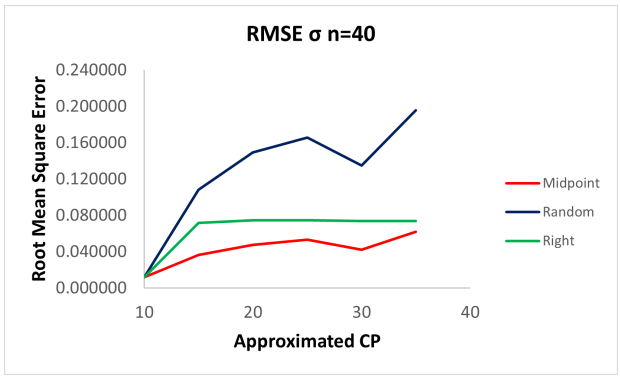
Figure 4 : RMSE of $\hat{\beta}_1$ at $n = 20, 40, 60, 80, 100$ when $rcp = 20\%$.

Table 5 : Root Mean Square Error of $\hat{\sigma}$ for different imputation techniques when $rcp = 10\%$

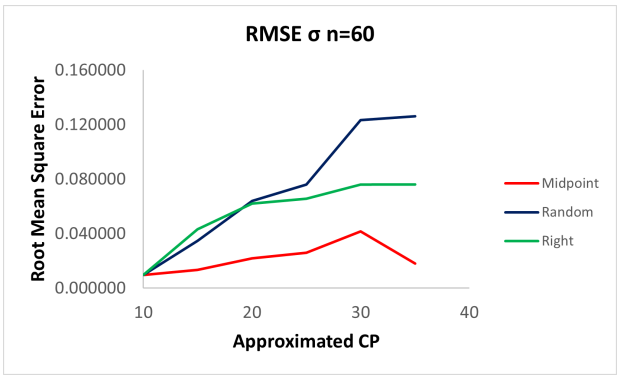
n	CP(%)	Imputation Techniques			
		Midpoint	**Random**	**Right**	**Left**
20	10	0.017586	0.017586	0.017586	0.017586
	15	0.024425	0.080929	0.054360	2.982548
	20	0.036139	0.132376	0.064816	4.830288
	25	0.036072	0.128277	0.064745	4.890446
	30	0.040018	0.143183	0.066557	5.431131
	35	0.040697	0.151801	0.066754	5.580538
40	10	0.011767	0.011767	0.011767	0.011767
	15	0.036483	0.107961	0.071724	4.612888
	20	0.047358	0.149283	0.074422	6.128125
	25	0.053173	0.165766	0.074697	6.881151
	30	0.042100	0.134817	0.073561	5.463443
	35	0.061744	0.195741	0.073824	7.920065
60	10	0.009654	0.009654	0.009654	0.009654
	15	0.013359	0.034595	0.043186	1.285930
	20	0.021974	0.063976	0.062114	2.707797
	25	0.026040	0.076136	0.065503	3.237399
	30	0.041570	0.123256	0.075969	5.270100
	35	0.017901	0.126075	0.076176	5.417171
80	10	0.008430	0.008430	0.008430	0.008430
	15	0.021280	0.061474	0.062994	2.550239
	20	0.022981	0.067628	0.064800	2.800593
	25	0.034303	0.100708	0.073692	4.330900
	30	0.040233	0.119835	0.075304	5.075757
	35	0.046781	0.138901	0.076345	5.932943
100	10	0.007684	0.007684	0.007684	0.007684
	15	0.021322	0.060726	0.063390	2.557964
	20	0.030365	0.086445	0.071572	3.790339
	25	0.031601	0.090860	0.072430	3.939698
	30	0.043287	0.123463	0.077369	5.417524
	35	0.051540	0.148417	0.077556	6.478165



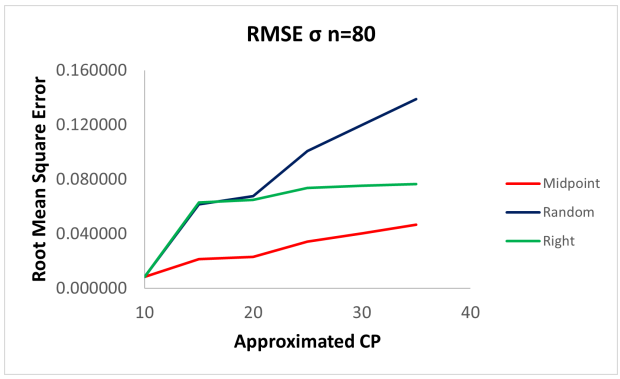
(a) n=20



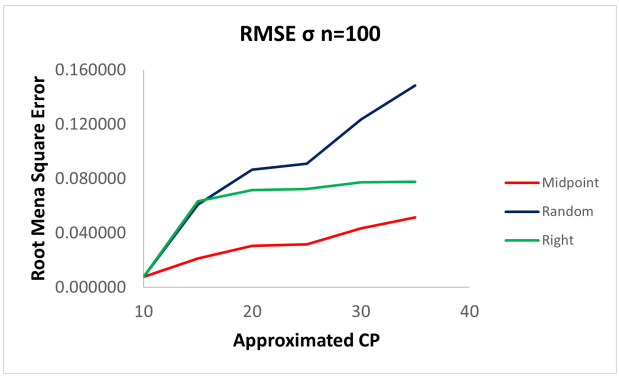
(b) n=40



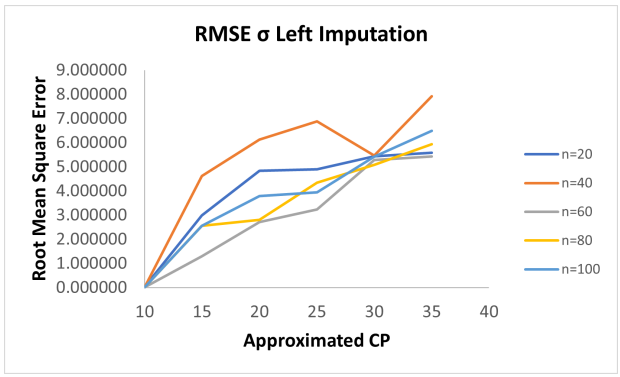
(c) n=60



(d) n=80



(e) n=100

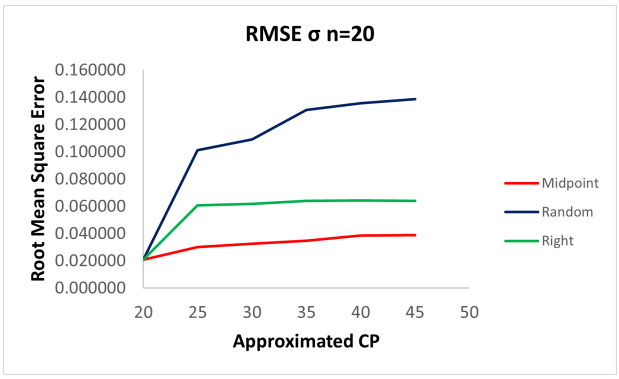


(f) Left Imputation for n=20, 40, 60, 80, 100

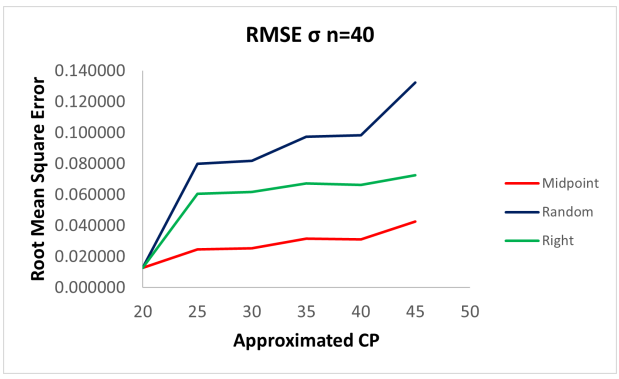
Figure 5 : RMSE of $\hat{\sigma}$ at $n = 20, 40, 60, 80, 100$ when $rcp = 10\%$.

Table 6 : Root Mean Square Error of $\hat{\sigma}$ for different imputation techniques when $rcp = 20\%$

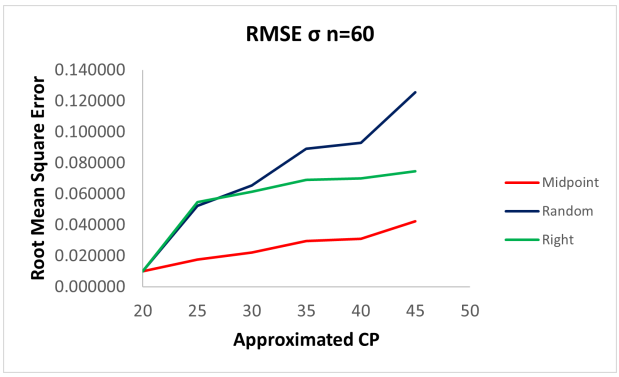
n	CP(%)	Imputation Techniques			
		Midpoint	**Random**	**Right**	**Left**
20	20	0.020688	0.020688	0.020688	0.020688
	25	0.030126	0.101284	0.060564	3.783700
	30	0.032484	0.109077	0.061753	4.272115
	35	0.034653	0.130648	0.063888	4.687632
	40	0.038664	0.135751	0.064256	5.233033
	45	0.038932	0.138718	0.064101	5.279228
40	20	0.012670	0.012670	0.012670	0.012670
	25	0.024630	0.079784	0.060612	3.073102
	30	0.025449	0.081795	0.061611	3.121097
	35	0.031606	0.097412	0.067235	3.962946
	40	0.030985	0.098357	0.066320	3.914184
	45	0.042667	0.132331	0.072544	5.604269
60	20	0.009942	0.009942	0.009942	0.009942
	25	0.017608	0.052135	0.054750	1.982828
	30	0.022010	0.065476	0.061404	2.679788
	35	0.029587	0.089215	0.068929	3.754236
	40	0.030908	0.092862	0.070006	3.937432
	45	0.042366	0.125458	0.074529	5.380565
80	20	0.008625	0.008625	0.008625	0.008625
	25	0.017958	0.051568	0.057111	2.095049
	30	0.027435	0.079708	0.068631	3.399555
	35	0.031531	0.093271	0.071775	3.940929
	40	0.041840	0.122384	0.075678	5.318627
	45	0.053552	0.157789	0.076417	6.804488
100	20	0.008028	0.008028	0.008028	0.008028
	25	0.011015	0.028557	0.042638	1.069301
	30	0.019259	0.054868	0.060525	2.285490
	35	0.027874	0.078908	0.069453	3.461618
	40	0.044863	0.126961	0.077451	5.601950
	45	0.045425	0.130883	0.076802	5.731358



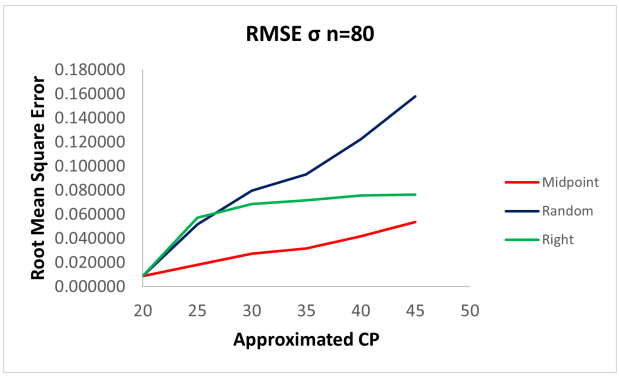
(a) n=20



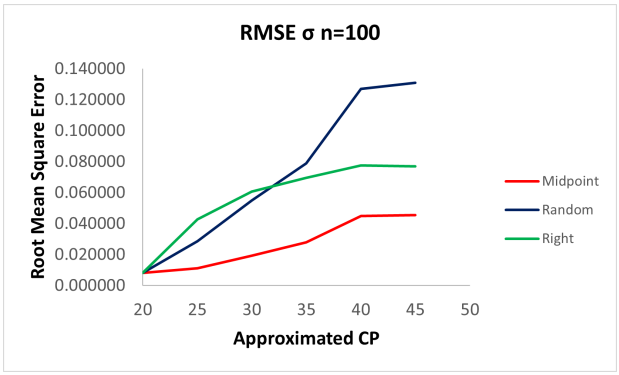
(b) n=40



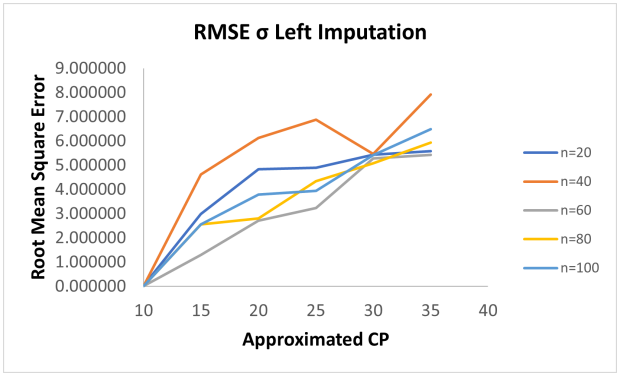
(c) n=60



(d) n=80



(e) n=100



(f) Left Imputation for n=20, 40, 60, 80, 100

Figure 6 : RMSE of $\hat{\sigma}$ at $n = 20, 40, 60, 80, 100$ when $rcp = 20\%$.

It can be observed that the values of root mean square error of $\hat{\beta}_0$, $\hat{\beta}_1$ and for $\hat{\sigma}$ increases when the censoring proportion increases almost all the times except at some censoring proportions where there is a slight decrease in the values.

When the $rcp = 10\%$, for of $\hat{\beta}_0$, the random imputation records the lowest root mean square error values followed by the midpoint and right imputation techniques. The midpoint imputation technique records the lowest root mean square error value for $\hat{\beta}_1$ and $\hat{\sigma}$. For $\hat{\beta}_1$, the random

imputation technique records the second lowest root mean square error value, followed by the right imputation technique, meanwhile right imputation technique records the second lowest root mean square error values for $\hat{\sigma}$. The left imputation technique gives the largest root mean square error values for all the parameter estimates.

The $rcp = 20\%$ exhibit almost a similar pattern as $rcp = 10\%$ and gives the same conclusions for the imputation techniques. The value of the root mean square error is higher in $rcp = 10\%$ compared to 20% when the censoring proportion is the same. The rcp value is approximated at 10% and 20% respectively. The remaining censoring proportion value is from the interval-censored proportion value, icp. This indicates that when the censoring proportion has the same value for different combinations of rcp and icp, the value of the root mean square is lower when the rcp is approximated at 20% compared to 10%. Thus, it can concluded that a higher rcp value gives a lower root mean square error value.

The conclusion of the best imputation techniques was chosen based on the lowest value of the root mean square error ($RMSE$) for the different imputation techniques. This is because, lower $RMSE$ values yields a more accurate and efficient estimates for $\hat{\beta}_0$, $\hat{\beta}_1$ and for $\hat{\sigma}$. Thus, from the simulation study, there are two imputation techniques that gives lowest RMSE values which is random imputation technique and midpoint imputation technique. Random imputation technique gives the lowest root mean square error value when it comes to $\hat{\beta}_0$ while the midpoint imputation technique gives the lowest root mean square error value for $\hat{\beta}_1$ and $\hat{\sigma}$. Thus, the midpoint imputation technique is chosen as the best imputation technique as it yields the lowest RMSE for most of the parameters. Thus, these technique will be used for real data analysis in the next chapter.

4 REAL DATA ANALYSIS

In this study, real data of diabetic nephropathy were fitted to the Weibull distribution model. This data was obtained from Steno Memorial Hospital in Denmark, which describes the survival time for Type I diabetes patients to develop Diabetic Nephropathy (DN). Diabetic Nephropathy is a sign of kidney failure for the diabetes patient.

According to Mayo Clinic [20], diabetic nephropathy is a serious complication of Type I and Type II diabetes. About 1 in 3 people with diabetes in the United States have diabetic nephropathy. This disease damages blood vessels and cells in the kidneys, leading to the loss of kidney function. High blood pressure caused by poorly controlled diabetes further damages the kidney and may progress to end-stage kidney disease. Treatment options for end-stage diabetic nephropathy include dialysis or a kidney transplant. To delay or prevent the disease, it is important to maintain a healthy lifestyle, manage diabetes and high blood pressure, and seek early treatment.

The data consist of 731 patients on survival times in months and all the patients have developed Diabetic Nephropathy by the end of the study. The data consist of 454 males and 277 females where the gender of the patient indicates 0 if the patient is male and 1 if the patient is female. The data consist of uncensored and interval censored only where the censoring indicator, c_i is 1 if the data is uncensored and 0 if the data is interval-censored. This shows that

the data consists of 138 patients out of 731 patients or 18.87% of the data is interval censored. There are two survival times shown in the data, which are $tleft$, which indicates the left endpoint of the survival time, and $tright$, which indicates the right endpoint of the survival time when the patient is interval-censored. If the patient is uncensored, the $tleft$ and $tright$ values are the same. The data was futher modified to obtain right-censored, interval-censored, and uncensored data.

Both, real data and modified real data are analyzed to determine whether there is a significant effect on the gender of the survival time of the patients.

4.1 Preliminaries

The results shown in the preliminaries were done by using Midpoint Imputation technique.

The non-parametric Kaplan-Meier Survival curve for diabetic nephropathy data was plotted. The Weibull distribution model was fitted to the diabetic nephropathy data and the survival curve was then plotted on the same graph.

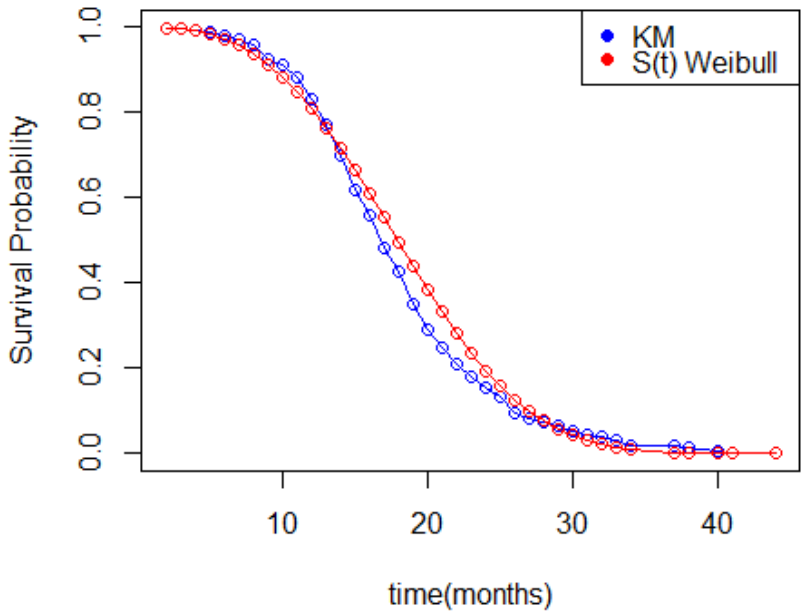


Figure 7 : Survival Curve.

From Figure 7, it can be observed that the $\hat{S(t)}$ Weibull is close to the Kaplan-Meier Survival Curve. This indicates that the Weibull Model may provide a good fit for the Diabetic Nephropathy data.

Descriptive statistics, which are mean, standard error of the mean, standard deviation, and

mean for survival times are analyzed.

Table 7 : Descriptive Statistics of Survival Times

n	Mean	Std.Error	Std.Dev	Median
731	16.6990	0.2311	6.2492	16

Table 7 shows the overall descriptive statistics for the survival time of patients' kidneys. The mean survival time is 16.6990 months, whereas the median survival time is 16 months. The median survival time is lower than the mean survival time. This means that the survival time is positively skewed, and this skewness can be clearly seen in the histogram plotted in Figure 8.

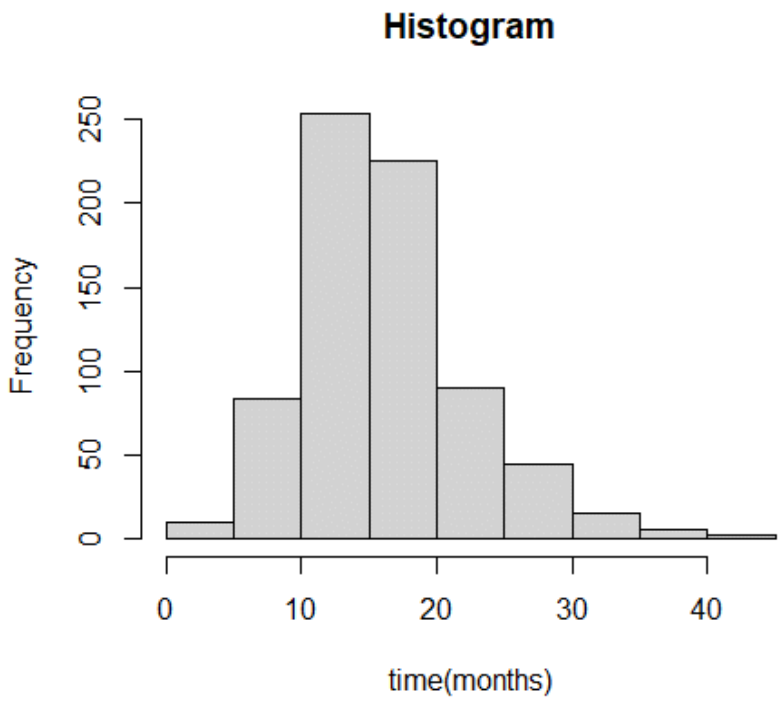


Figure 8 : Histogram of Survival Time.

Table 8 : Descriptive Statistics of Survival Time by Gender

Gender	n	Mean	Std.Error	Std.Dev	Median
Male	454	17.0892	0.2869	6.1135	16
Female	277	16.0596	0.3861	6.4255	15

Table 8 illustrates descriptive statistics based on the Gender of the patients. The majority of patients are Males with 454 out of a total of 731 patients. The mean survival time of the male kidney is 17.0892 months, whereas the mean survival time of the female kidney is 16.0596 months. The mean survival time of the male kidney is higher than the mean survival time of the female kidney. This might be due to post-menopausal females having an increased risk of developing Diabetic Kidney Disease and End-Stage Kidney Disease and glomerular hyperfiltration than women (Shephard, [21]). On the other hand, the median survival time of the male kidney is 16 which is higher than the median survival time of the woman kidney which is 15. This shows that the male kidney survives slightly longer than the female kidney.

Figure 9 illustrates the survival plots based on gender (male and female).

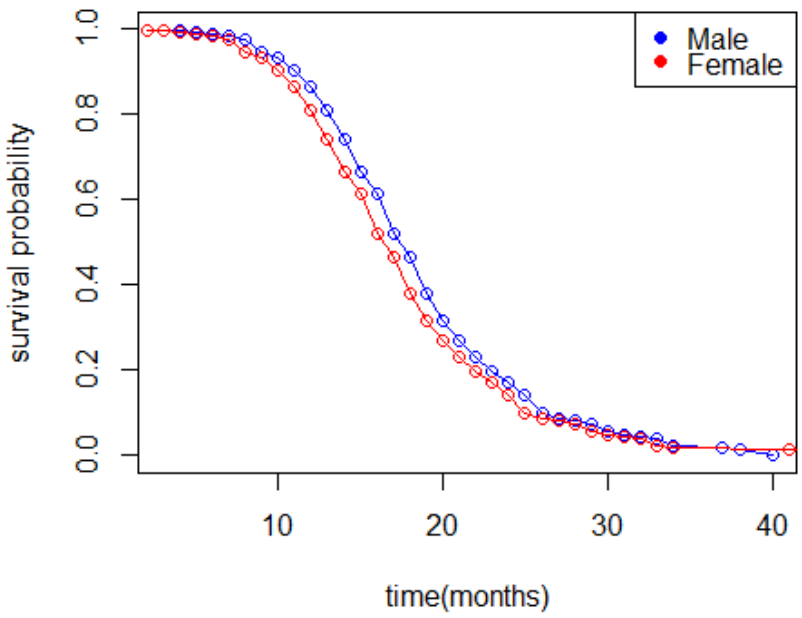


Figure 9 : Survival Plot by Gender.

From Figure 9, it is clearly shown that survival probabilities for both male kidneys and female kidneys decrease over time. However, the survival probability for male kidneys is slightly higher than female kidneys because the survival curve of males is above the survival curve of females almost all the time.

Figure 10 : Hazard Plot by Gender.

Both male and female kidneys have the same hazard rate pattern as illustrated in Figure 10. Both hazard rates rise over time, but the hazard rate of the female kidneys is higher than the male kidneys. This result shows that the failure rate of female kidneys is higher than the male kidneys.

4.2 Non-Parametric Approach

Since the preliminary analysis shows that gender affects survival time, hence, the non-parametric log-rank test for midpoint imputation will be conducted to check whether gender which acts as a covariate has a significant effect on survival time. Table 9 show the log-rank test statistics based on gender (male and female based on midpoint imputation. There is no difference in the survival difference between male and female if $S_1(t) = S_2(t)$.

$$H_0 : S_1(t) = S_2(t)$$
$$H_1 : S_1(t) \neq S_2(t)$$

$$\chi^2(LR) = 6.5 \sim \chi^2(1), p = 0.01$$

Table 9 : Log-Rank test statistics of Gender based on Midpoint Imputation

Gender	n_{1j}	d_{1j}	e_{ij}	$\frac{(d_{1)} - e_{1j})^2}{e_{1j}}$	$\frac{(d_{1j} - e_{1j})^2}{v_{1j}}$
Male	454	356	384	2.02	6.5
Female	277	237	209	3.70	6.5

The Chi-Square test statistic is 6.5 with 1 degree of freedom and the corresponding p-value is 0.01. Since the p-value is less than 0.05, we will reject the null hypothesis. Thus, it can be concluded that there is significant difference in survival times of 2 genders.

4.3 Real Data Analysis with Weibull Regression Model

The real data consist of uncensored and interval-censored observations only.

Based on the diabetic nephropathy data, we conduct the hypothesis testing on the covariate, β_1(gender) to check whether it has a significant effect on the survival times of the patient's kidney.

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Table 10 : Descriptive statistics of Parameter Estimates using Midpoint Imputation

	Est.	Std.Error	Wald	p
Intercept	3.1029	0.0421	73.75	<0.001
Gender	-0.0670	0.0284	-2.36	0.018
Log(Scale)	-1.0836	0.0290	-37.32	<0.001

Table 10 illustrates the descriptive statistics of the parameter estimates using Midpoint Imputation. Wald statistics for Gender is -2.36 which falls in the rejection region, which is -1.96 when $\alpha = 0.05$. Thus, the null hypothesis is rejected and it can be concluded that Gender is significant in the model.

Table 11 : Wald Confidence Interval

	95% Confidence Interval
Intercept	(3.0204 , 3.1854)
Gender	(-0.1227 , -0.0113)

Wald test is further carried out to check the significance of gender on survival time. Based on Table 11, a 95% confidence interval for the parameter of β_1 is estimated to fall in the region between -0.1227 to -0.0113. It can be observed that the zero is not included in the confidence intervals. Thus, the null hypothesis is rejected and it can be concluded that the gender has a significant effect on the survival time of patients' kidneys.

4.4 Modified Real Data Analysis with Weibull Regression Model

The real data is modified, which makes the data consist of uncensored, right-censored and interval-censored observations.

Table 12 : Descriptive statistics of Parameter Estimates using Midpoint Imputation

	Est.	Std.Error	Wald	p
Intercept	3.1216	0.0428	72.96	<0.001
Gender	-0.0750	0.0289	-2.59	0.0095
Log(Scale)	-1.0668	0.0250	-36.22	<0.001

Table 12 illustrates the descriptive statistics of the parameter estimates using Midpoint Imputation. Wald statistics for Gender is -2.59 which falls in the rejection region, which is -1.96 when $\alpha = 0.05$. Thus, the null hypothesis is rejected and it can be concluded that the gender is significant in the model.

Table 13 : Wald Confidence Interval

	95% Confidence Interval
Intercept	(3.0377 , 3.2055)
Gender	(-0.1316 , -0.0184)

Wald test is further carried out to check the significance of gender on the survival time. Based on Table 13, a 95% confidence interval for the parameter of β_1 falls in the region between -0.2845 to -0.0587. It can be observed that zero is not included in the confidence intervals. So, the null hypothesis is rejected, and it can be concluded that gender has a significant effect on the survival time of patients' kidneys.

5 CONCLUSION

In this research, the covariate is incorporated into the Weibull regression model with right and interval-censored data. The maximum likelihood estimation method is approached to obtain the parameter estimates. To solve the nonlinear likelihood equation, Newton Raphson's iterative procedure was performed. We compare the performance of the parameter estimates at different sample sizes and censoring proportions using the values of bias, standard error and root mean square error. Better estimates with lowest root mean square error for all the parameters can be obtained using midpoint imputation technique at all sample sizes.

Furthermore, in general, the results showed that the parameter estimates for the Weibull regression model with covariate perform best with low censoring proportion level and large sample size. On the other hand, high censoring proportion and small sample size give higher values of root mean square error, which indicates that the model is less efficient.

It was discovered that the Weibull regression model provided a good fit for diabetic nephropathy data. From the preliminary analysis, it can be observed that the survival probability for male kidneys is slightly higher than the female kidneys. Additionally, it was concluded that gender has a significant effect on the survival time of patients' kidneys through non-parametric log-rank, the Wald test and Wald confidence interval test.

For the analysis of real data analysis and modified real data analysis, we found out that gender has a significant effect on the survival time of patients' kidneys by obtaining the summary of the maximum likelihood estimation of the Weibull regression model, the Wald test and Wald confident interval test as well.

ACKNOWLEDGEMENT

First of all, I would like to express my deepest gratitude to Assoc. Prof. Dr. Jayanthi a/p Arasan for her invaluable guidance and support as my supervisor throughout this project. Dr. Jayanthi's advice and assistance have been instrumental in helping me gain a deeper understanding of this project and its broader implications in life.

I would also like to extend my thanks to the Department of Mathematics and Statistics, Faculty of Science, University Putra Malaysia, for providing the resources and support that made this project possible.

A heartfelt thank you to my family for always supporting my dreams and standing by me in times of need.

Lastly, I want to express my solidarity with the cause of Free Palestine.

REFERENCES

[1] T. Smith, B. Smith, and M. A. Ryan, "Survival analysis using cox proportional hazards modeling for single and multiple event time data," in *Proceedings of the twenty-eighth annual SAS users group international conference, SAS Institute, Inc, Cary, paper*, 2003, pp. 254–228.

[2] E. T. Lee and J. Wang, *Statistical methods for survival data analysis*. John Wiley & Sons, 2003.

[3] J. L. Schafer and J. W. Graham, "Missing data: our view of the state of the art." *Psychological methods*, vol. 7, no. 2, p. 147, 2002.

[4] C.-D. Lai, D. Murthy, and M. Xie, "Weibull distributions and their applications," in *Springer Handbooks*. Springer, 2006, pp. 63–78.

[5] A. C. Cohen, "Maximum likelihood estimation in the weibull distribution based on complete and on censored samples," *Technometrics*, vol. 7, no. 4, pp. 579–588, 1965.

[6] G. Stone and H. R. Van, "Parameter estimation for the weibull distribution," *IEEE Transactions on Electrical Insulation*, no. 4, pp. 253–261, 1977.

[7] P. M. Odell, K. M. Anderson, and R. B. D'Agostino, "Maximum likelihood estimation for interval-censored data using a weibull-based accelerated failure time model," *Biometrics*, pp. 951–959, 1992.

[8] C. B. Guure, N. A. Ibrahim, and M. B. Adam, "On partly censored data with the weibull distribution," *ARPN Journal of Engineering and Applied Sciences*, vol. 7, no. 10, pp. 1329–1334, 2006.

[9] E. Strapasson, "A simulation study to compare imputation methods to handle g rouped surv iv al data'," *Rev. Bras. Biom*, vol. 27, no. 2, p. 210, 2009.

[10] A. A. Salahaddin, "Comparative study of four methods for estimating weibull parameters for halabja, iraq," *International Journal of Physical Sciences*, vol. 8, no. 5, pp. 186–192, 2013.

[11] Z. Zhang, "Parametric regression model for survival data: Weibull regression model as an example," *Annals of translational medicine*, vol. 4, no. 24, 2016.

[12] A. Zyoud, F. M. Elfaki, and M. Hrairi, "Parametric model based on imputations techniques for partly interval censored data," in *Journal of Physics: Conference Series*, vol. 949, no. 1. IOP Publishing, 2017, p. 012002.

[13] M. C. Lai and J. Arasan, "Single covariate log-logistic model adequacy with right and interval censored data," *Journal of Quality Measurement and Analysis*, vol. 16, no. 2, pp. 131–140, 2020.

[14] S. F. Khairunnisa, S. Suyitno, and S. Mahmuda, "Weibull regression model on hospitalization time data of covid-19 patients at abdul wahab sjahranie hospital samarinda," *Jurnal Matematika, Statistika dan Komputasi*, vol. 19, no. 2, pp. 286–303, 2023.

[15] K. Kiani and J. Arasan, "Gompertz model with time-dependent covariate in the presence of interval-, right-and left-censored data," *Journal of Statistical Computation and Simulation*, vol. 83, no. 8, pp. 1472–1490, 2013.

[16] A. M. Alharpy and N. A. Ibrahim, "Parametric tests for partly interval-censored failure time data under weibull distribution via multiple imputation," *Journal of Applied Sciences*, vol. 13, no. 4, pp. 621–626, 2013.

[17] N. M. Saeed and F. A. M. Elfaki, "Parametric weibull model based on imputations techniques for partly interval censored data," *Austrian Journal of Statistics*, vol. 49, no. 3, pp. 30–37, 2020.

[18] W. A. Scott, "Maximum likelihood estimation using the empirical fisher information matrix," *Journal of Statistical Computation and Simulation*, vol. 72, no. 8, pp. 599–611, 2002.

[19] A. Maria, "Introduction to modeling and simulation," in *Proceedings of the 29th conference on Winter simulation*, 1997, pp. 7–13.

[20] S. Mayo Clinic, "Diabetic nephropathy (kidney disease)," Oct (2021). [Online]. Available: https://www.mayoclinic.org/diseases-conditions/diabetic-nephropathy/symptoms-causes/syc-20354556

[21] B. D. Shepard, "Sex differences in diabetes and kidney disease: mechanisms and consequences," *American Journal of Physiology-Renal Physiology*, vol. 317, no. 2, pp. F456–F462, 2019.