Parametric Weibull Model Based on Imputations Techniques for Partly Interval Censored Data

Noora Mohsin Saeed
Qatar University

Faiz A. M. Elfaki
Qatar University

Abstract

The term survival analysis has been used in a broad sense to describe collection of statistical procedures for data analysis for which the outcome variable of interest is time until an event occurs, the time to failure of an experimental unit might be censored and this can be right, left, interval, and Partly Interval Censored data (PIC). In this paper, the analysis of this model is conducted based on parametric Weibull model via PIC data. Moreover, two imputation techniques are used, which are: left point and right point. The effectiveness of the proposed model is tested through numerical analysis on simulated and secondary data sets.

Keywords: partly interval censored, Weibull model, imputation techniques.

1. Introduction

Statistical method is one of the strategies used by researchers as it provides various kinds of methods in analyzing data. One of the methods used in the data analysis is the survival analysis method. Survival analysis or failure time analysis is described as one of the most significant and advanced methods in statistics during the last quarter of the 20th century (Sam and Krings (2008)). It is one of the significant statistical methods as it is involved with the failures of components (Singh and Totawattage (2012)). Kleinbaum and Klein (2005) described the survival analysis as the procedure to analyze data statistically and the outcome is the time until an event occurs. There are many applications of survival analysis, for example in medical, engineering, education, economic, and other areas. Mostly the survival analysis method has been widely used in the biomedical as well as in the engineering applications. As mentioned by Liu (2012), one of the examples of engineering application that deals with the survival analysis method is the testing of life time or durability of a mechanical or an electrical component. Scientists apply this technique to track the products and material's life span for predicting the product reliability and durability. Lawless (2003) described that the duration could be compared with lifetime of a marriage; a marriage may end due to annulment, divorce, or death. Other example from education scope, as mentioned by Eagle and Barnes (2014), used survival analysis approach for measuring time until an event occurs and account for teacher's attrition. However, this research uses Weibull model for secondary data form medical and simulated data based on education data sets. While dealing with

survival data, censoring should be involved. Censoring occurs when the information of a failure time of some subjects is incomplete. There are different reasons for censoring which lead to different types of censored data. These are right, left, interval, and one of the most important types of interval censored data is partly interval censored data which means that for some of the subjects the event of interest is exactly observed while for others it lies within an interval (Kim (2003)). In this paper, analysis is conducted based on partly interval censored via simulated and secondary medical data.

2. Weibull distribution model

In lifetime data one of the most useful distributions for analyzing and modelling is the Weibull distribution in various fields such as; medical, biology, engineering, and others. It is applicable for various failure situations and was proposed by Weibull (1939). Lee and Wang (2003) proposed that Weibull distribution it used in many mortalities of the human disease studies and reliability studies. It is described by two parameters, which are: shape parameter that determine the distribution curve and the other parameter that determine the scaling (the two parameters are shown in equation 1). Lee and Wang (2003) proposed the probability density function of the two Weibull distribution parameters as,

$$f(t; \alpha, \beta) = \begin{cases} \frac{\beta}{\alpha} \left(\frac{t}{\alpha}\right)^{\beta-1} e^{-\left(\frac{t}{\alpha}\right)^{\beta}}, & t \geq 0 \\ 0, & t < 0, \end{cases} \tag{1}$$

where β and α presents the shape and scale parameters respectively. The cumulative Weibull distribution function is given as,

$$F(t; \alpha, \beta) = \begin{cases} 1 - e^{-\left(\frac{t}{\alpha}\right)^{\beta}}, & t \geq 0 \\ 0, & t < 0, \end{cases} \tag{2}$$

This study presents the model under survivorship and the estimation of the curve of the survival probability function is calculated by using censoring data based on Weibull distribution model. The likelihood function for Weibull distribution with data involve failure, right censored, and interval-censored is given as (Guure, Ibrahim, and Adam (2012)):

$$L(t_i, u_i, v_j, \alpha, \beta) = \prod_{i=1}^{k} f(t_i) \prod_{i=k+1}^{r} (1 - F(t_i)) \prod_{j=r+1}^{n} (F(v_j, \alpha, \beta) - F(u_j, \alpha, \beta)). \tag{3}$$

This implies that,

$$L(t_i, u_i, v_j, \alpha, \beta) = \prod_{i=1}^{k} \frac{\beta}{\alpha} \left(\frac{t_i}{\alpha}\right)^{\beta-1} e^{-\left(\frac{t_i}{\alpha}\right)^{\beta}} \times \prod_{i=k+1}^{r} e^{-\left(\frac{t_i}{\alpha}\right)^{\beta}} \times \prod_{j=r+1}^{n} \left(e^{-\left(\frac{u_j}{\alpha}\right)^{\beta}} - e^{-\left(\frac{v_j}{\alpha}\right)^{\beta}}\right) \tag{4}$$

Then the log-likelihood is taken from equation (4) and differentiated with respect to α and β. Then numerical method, such as Newton Raphson method is applied to obtain the values of β and α.

3. Results and data analysis

This paper illustrates the implementation of the methods, discussed in the earlier sections, by using two data sets. The first one is breast cancer data, and the second one is auto generated data. All calculations were computed by using R software. In the simulation study in 5 we used simple imputation methods to impute the missing data or the exact data. The two imputation methods are right-point where the event time is imputed by the right limit of the interval, and left-point where the event time is imputed by the left limit of the interval.

4. Breast cancer data

Several researchers used this data in their study such as; Zyoud, Elfaki, and Hrairi (2016) modified the partly interval censored data and compared with the Turnbull method. There were two failure times that are Radiation (R) and Radiation + Chemotherapy (R+C). In the first failure time (R) there were 66 patients and 68 patients for the second failure time (R+C). The objective is to compare the cosmetic effects of the first failure time against the second failure time on women with early breast cancer and the event of interest was represented by the time to first occurrence of breast retraction. The actual dates are recorded based on the availability of the patient while visiting the clinic for every 4 to 6 months. To set up the data as the PIC, the same way is followed that was used by Alharphy and Ibrahim (2013) and Zyoud *et al.* (2016). Figure 1 shows the survival curve for Radiotherapy and Radiation + Chemotherapy by using Weibull distribution and Turnbull method. It is clear from the figure that the estimated survival curves obtained by the Weibull model lies close to the one obtained by Turnbull. These results indicated that the proposed model is fitted well compare with Turnbull method. Parameters estimates (shape and scale) of Weibull distribution and standard errors (se) of the two treatments which are presented in Table 1. Moreover, the likelihood ratio test for this model shows the value as 12.86, where P-value is almost zero.

Table 1: Parameters estimated based on Weibull distribution for breast cancer PIC Data

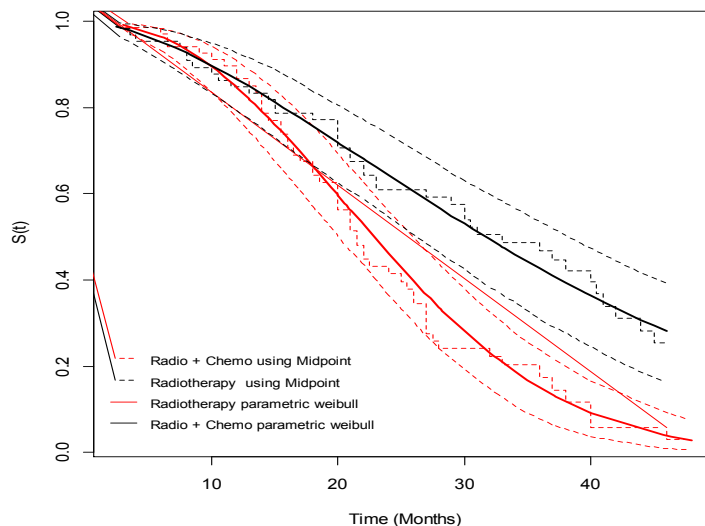|	Shape			Scale		
	Estimation	CI	SE	Estimation	CI	SE
All	1.858	(1.574, 2.193)	0.157	32.402	(29.095, 36.085)	1.780
Radiotherapy	1.620	(1.240, 2.110)	0.220	39.690	(32.720, 48.160)	3.920
Radio+Chemo	2.195	(1.785, 2.699)	0.232	26.786	(33.176, 30.254)	1.664

Figure 1: Survival function obtained by Weibull model and Turnbull methods based on PIC data

5. Simulation data

To evaluate and study the behavior of statistical procedures in statistic, simulation procedure is mostly used especially for the situation when a problem cannot be solved analytically

(Elfaki, Bin Daud, Ibrahim, Abdullah, and Usman (2007)). The technique requires setup of many samples. The samples are then individually reckoned in terms of statistics of interest, and the overall statistics of interest is used to study distribution properties. The objective of this simulation study is to compare the survival function for local and international students based on partly interval censored data. The simulated data were generated based on the education data set with two failure times that is are local and international students (This education data set is not provided in this paper, readers are referred to Saeed (2018)). The Weibull distribution is used to generate the data (the Weibull was found to be fit the original data very well as in Saeed (2018)). To generate the data we used the mean and standard deviation as 19.7538, 1.3354 for local students and 11.749, 0.0548 for international student, respectively. The generated data were 500 in count for local and international students as sample size.

6. Results from partly interval censored data

There are two scenarios; in the first scenario the study has taken 50% exact observed and 50% observed as interval. In the second scenario, the study has taken the exact observation of 70% and 30% for the observed as interval. The result of this partly interval censored data shows in Figure 2, 3, 4 and Figure 5. These figures show that the survival curves obtaining by Weibull for the local students and international students compared with Turnbull method. The local student showed a longer survival compared to the international students for both scenarios, which is indicate that the local students are more stable compare with international students. Table 2 and Table 3 show the shape and scale parameter obtained by our model based on right point and the results look similar for both students (local and international) in the two scenarios. Likewise, the estimated parameters obtained by left point imputation for two scenarios are almost similar with respect to the parameters and standard error (Tables 4 & 5). Moreover, the likelihood ratio test obtained by right point for two scenarios are 38.11 (0) and 35.6 (0), respectively. Similarly, the likelihood ratio test obtained by left point for two scenarios are 38.48 (0) and 40.38 (0), respectively, which implement the significant of the model.

Table 2: The parameters estimated based on Weibull distribution from simulation with PIC Data (50:50) - Right Point

	Shape			Scale		
	Estimation	CI	SE	Estimation	CI	SE
All	32.203	(29.106, 35.629)	1.6614	12.035	(11.996, 12.079)	0.0213
Local	66.156	(55.921, 78.265)	5.6736	12.075	(12.042, 12.109)	0.0171
International	25.113	(22.153, 28.468)	1.6067	11.909	(11.840, 11.978)	0.0351

Table 3: The parameters estimated based on Weibull distribution from simulation with PIC Data (70:30) - Right Point

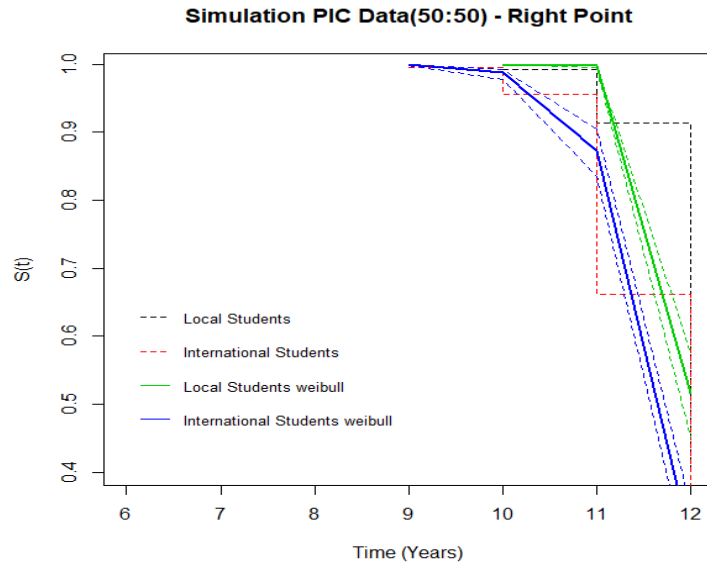|	Shape			Scale		
	Estimation	CI	SE	Estimation	CI	SE
All	51.783	(46.436, 57.745)	2.8793	12.0663	(12.039, 12.093)	0.0138
Local	387.88	(328.82, 457.54)	32.687	12.018	(12.012, 12.024)	0.003
International	32.957	(28.798, 37.716)	2.2681	11.993	(11.939, 12.047)	0.0276

Figure 2: The survival function based on PIC data (50:50) - Right point imputation
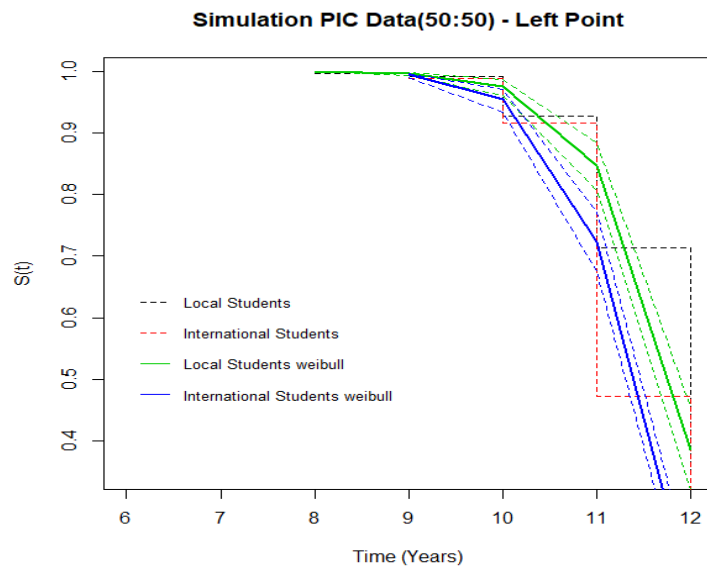

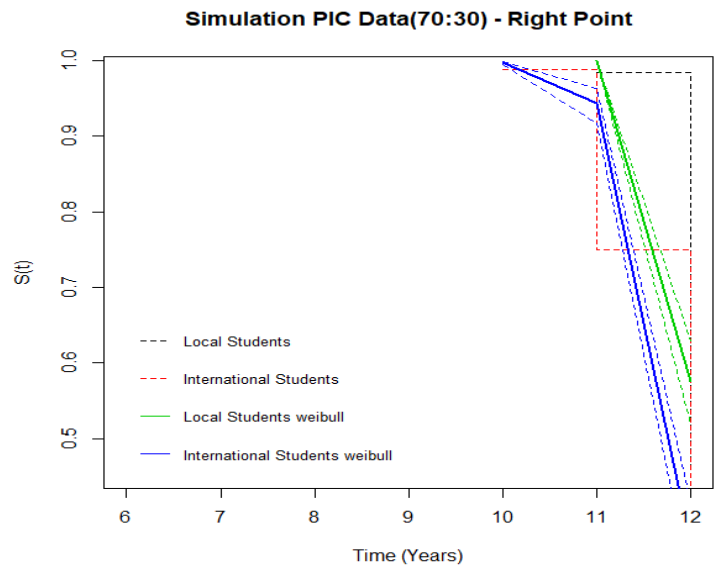Figure 3: The survival function based on PIC (50:50) - Left point imputation

Figure 4: The survival function based on PIC data 70:30 (70% exact) - Right point imputation
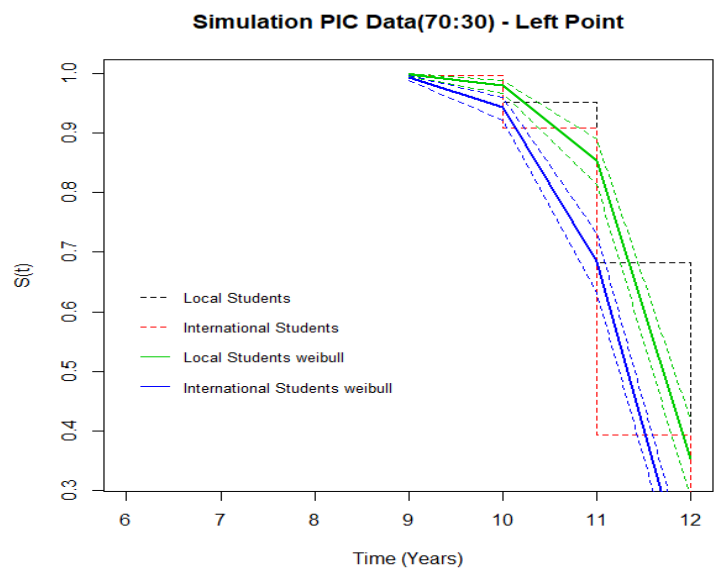

Figure 5: The survival function based on PIC data 70:30 (70% exact) - Left point imputation

Table 4: The parameters estimated based on Weibull distribution from simulation with PIC Data (50:50) - Left Point

	Shape			Scale		
	Estimation	CI	SE	Estimation	CI	SE
All	19.663	(17.967, 21.518)	0.9046	11.827	(11.759, 11.894)	0.0344
Local	20.032	(17.323, 23.165)	1.485	12.029	(11.922, 12.136)	0.0548
International	20.304	(18.143, 22.722)	1.165	11.626	(11.542, 11.711)	0.043

Table 5: The parameters estimated based on Weibull distribution from simulation with PIC Data (70:30) - Left Point

	Shape			Scale		
	Estimation	CI	SE	Estimation	CI	SE
All	19.714	(18.067, 21.512)	0.8779	11.775	(11.709, 11.840)	0.0335
Local	21.627	(18.800, 24.878)	1.545	11.979	(11.884, 12.075)	0.0487
International	19.764	(17.719, 22.046)	1.102	11.5507	(11.466, 11.636)	0.0433

7. Concluding

In this study, Weibull model is used based on simple imputation technique to simplify the procedure for partly interval censored data, which are the right point and left point. The estimated survival function was obtained based on the maximum likelihood estimation and comparisons were made with existing literature. From the breast cancer data, it is confirmed that the proposed model is fit to use well and easy to implement compared with the one obtained by Turnbull method. The simulation data was used based on the education data. The data generated for 500 times from international and local students. It can be concluded that the Weibull model based on simulation results are suitable for partly interval censored data compared with interval data. Finally, the result reflects that when the observed data have more exact in the data, the model is better fitted which is same line with other results obtained by some researchers such as Kim (2003), Zyoud *et al.* (2016), and Alharphy and Ibrahim (2013).

References

Alharphy A, Ibrahim N (2013). "Parametric Test for Partly Interval-censored Failure Time Data under Weibull Distribution via Multiple Imputation." *Journal of Applied Science*, **3**(4), 621–626.

Eagle M, Barnes T (2014). "Survival Analysis on Duration Data in Intelligent Tutors." *Intelligent Tutoring Systems*, **8474**, 178–187.

Elfaki FA, Bin Daud I, Ibrahim NA, Abdullah MY, Usman M (2007). "Competing Risks for Reliability Analysis Using Cox's Model." *Engineering Computations*, **24**(4), 373–383.

Guure CB, Ibrahim NA, Adam MB (2012). "Competing Risks for Reliability Analysis Using Cox's Model." *Science Asia*, **39S**, 75–79.

Kim JS (2003). "Maximum Likelihood Estimation for the Proportional Hazards Model with Partly Interval-censored Data." *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, **65**(2), 489–502. ISSN 1369-7412. doi:10.1111/1467-9868.00398. URL https://doi.org/10.1111/1467-9868.00398.

Kleinbaum DG, Klein M (2005). *Survival Analysis: A Self-learning Text.* Springer, New York.

Lawless JF (2003). *Statistical Models and Methods for Lifetime Data.* John Wiley and Sons, Canada.

Lee ET, Wang JW (2003). *Statistical Methods for Survival Data Analysis.* John Wiley and Sons, Canada.

Liu X (2012). *Survival Analysis Models and Applications.* John Wiley and Sons, United Kingdom.

Saeed NM (2018). *Weibull Distribution Based on Education Partly Interval Censored Data.* Master Project, Qatar University.

Sam Z, Krings AW (2008). "Survival Analysis Approach to Reliability, Survivability and Prognostics and Health Management (PHM)."

Singh R, Totawattage D (2012). "The Statistical Analysis of Interval-Censored Failure Time Data with Applications." *Journal of Statistics*, **3**, 155–166.

Zyoud A, Elfaki FA, Hrairi M (2016). "Non-parametric Estimate Based in Imputations Techniques for Interval and Partly Interval Censored Data." *Science International (Lahore)*, **28**, 879–884.

Affiliation:

Noora Mohsin Saeed
Department of Mathematics, Statistics and Physics
College of Arts and Sciences, Qatar University
P.O. Box 2713, Doha, Qatar
E-mail: 200355083@student.qu.edu.qa

Faiz A.M. Elfaki (Corresponding Author)
Department of Mathematics, Statistics and Physics
College of Arts and Sciences, Qatar University
P.O. Box 2713, Doha, Qatar
E-mail: felfaki@qu.edu.qa