PAPER • OPEN ACCESS

Parametric Model Based On Imputations Techniques for Partly Interval Censored Data

To cite this article: Abdallah Zyoud *et al* 2017 *J. Phys.: Conf. Ser.* **949** 012002

View the article online for updates and enhancements.

Parametric Model Based On Imputations Techniques for Partly Interval Censored Data

Abdallah Zyoud[1], F. A. M. Elfaki[2,*], and Meftah Hrairi[1]

[1]Department of Mechanical Engineering, Faculty of Engineering, IIUM, P.O. Box 10, 50728 Kuala Lumpur, Malaysia

[2]Department of Mathematics, Statistics and Physics, College of Arts and Science, Qatar University, P.O. Box 2713, Doha, Qatar

[*]E-mail: felfaki@qu.edu.qa

Abstract: The term 'survival analysis' has been used in a broad sense to describe collection of statistical procedures for data analysis. In this case, outcome variable of interest is time until an event occurs where the time to failure of a specific experimental unit might be censored which can be right, left, interval, and Partly Interval Censored data (PIC). In this paper, analysis of this model was conducted based on parametric Cox model via PIC data. Moreover, several imputation techniques were used, which are: midpoint, left & right point, random, mean, and median. Maximum likelihood estimate was considered to obtain the estimated survival function. These estimations were then compared with the existing model, such as: Turnbull and Cox model based on clinical trial data (breast cancer data), for which it showed the validity of the proposed model. Result of data set indicated that the parametric of Cox model proved to be more superior in terms of estimation of survival functions, likelihood ratio tests, and their P-values. Moreover, based on imputation techniques; the midpoint, random, mean, and median showed better results with respect to the estimation of survival function.

1. Introduction

One of the main objectives in the survival studies is the comparison of survival functions, which is due to the fact that it may be needed to compare two different treatments for one disease. In this paper, the comparison problem for several imputation techniques were discussed based on the existence of PIC failure time data. PIC data is common in numerous applications in medicine, engineering, biology and economy. One of its most common occurrences is in medical or health studies that have periodic follow-up hospital visits, where the event, usually a patient displaying a certain disease symptoms or recovery signs, may occur between visits, but it also may occur in the hospital, hence partly interval censoring. For some subjects, PIC data indicates the exact failure times which are observed. For the remaining subjects, survival time of interest is observed only for a certain time interval instead of being exact [1] [2] [3]. [4] provided an example of this kind of data that is; Framingham Heart Disease Study which follows a group of women diagnosed with breast cancer and separates them into two different groups based on the type of treatment that they receive. Another example about Fatigue Failure (Crack size data) which measures the number of cycles it takes before a crack appears in the subjects may be considered as PIC data. Using imputation greatly

simplifies the calculation process especially when dealing with larger sets of data. Applying several imputation techniques to our data was relatively easy to implement. Using R software, we were able to develop a programming code to apply different imputation methods to the data sets and proceed with the parametric analysis. In this study, several imputation techniques were used to estimate survival function and compared with the one that was obtained by Turnbull based on interval censored and PIC failure time data. In the next two sections, parametric model and imputation techniques will be discussed.

2. Parametric Model

While nonparametric methods have been the standard approach for analyzing simple homogeneous survival data without covariate information for making fewer assumptions, parametric survival models are sometimes used for inference. Let's take ($X_1, X_2, ..., X_n$) to be the true survival times, and also are assumed to be independent and identically distributed with survival function $S(t; \beta)$ and hazard function $h(t; \beta)$, but only a right censored sample $(\widetilde{X}_i, D_i)$, $i = 1, ..., n$ is observed. If we assume independent censoring, then the likelihood function for the parameter is presented by Equation (1).

$$L(\beta) = \prod_{i=1}^{n} (h(\widetilde{X}_i, \beta))^{D_i} S(\widetilde{X}_i, \beta) \tag{1}$$

where β is a parameter, that needs to be estimated.

The values of β can be estimated by maximizing the likelihood as shown in Equation (2). (The estimation of β's is not address it here in the paper, reader refer to [6])

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{\exp(\beta' z_i)}{\sum_{j \in R_i} \exp(\beta' z_j)} \right]^{D_i} \tag{2}$$

where $R_i = \{j : \widetilde{X}_j \geq \widetilde{X}_i\}$ is the risk set at time $\widetilde{X}_i$, which is a set of subjects which are still alive and uncensored at the time. We can use the Breslow estimator to estimate the cumulative baseline hazard $A_0(t)$ as presented in Equation (3).

$$\widehat{A}_0(t) = \sum_{\widetilde{X}_i \leq t}^{n} \left[\frac{D_i}{\sum_{j \in R_i} \exp(\widehat{\beta}' z_j)} \right] \tag{3}$$

Nelson Aalen estimator could be used if β is known and is also equal to the maximum likelihood estimate $\widehat{\beta}$. The estimator based on Equations (2) and (3) for which have a parametric maximum likelihood interpretation. Equation (2) can be used to estimate β in a large sample, in which $\widehat{\beta}$ is normally distributed with the proper mean and with a covariance that can be estimated by information matrix.

The function of Equation (1) may be analyzed using standard large sample theory. Thus, standard tests, that are: Wad test, Score test, and likelihood test, are used as inferential tools. Mostly distributions used in estimation parametric survival models that are Weibull distribution and exponential distribution. Both distributions contain very simplest model.

3. Imputation Methods

Several imputation techniques were used in this study and compared with Turnbull's model. The performances of the techniques were also compared with each other. The imputation techniques are based on Kaplan-Meir estimator.

The imputation techniques can be classified into two different categories, which are:

1. ***Simple imputation techniques***
   There are three main types of simple imputation techniques:
   1. Right imputation where the event time is imputed by the right limit of the censoring interval R_i .

   2. Left imputation where the event time is imputed by the left limit of the censoring interval L_i

   3. Mid imputation which refers to imputing the event time by the midpoint of the censoring interval $[L_i, R_i]$ as $(L_i + R_i)/2$.

2. ***Probability-based imputation techniques***
   It is usually require estimating the distribution of the partly interval censored data based on the observed intervals and using our knowledge of the distribution to impute the missing data. The most common probability-based imputation methods are, conditional mean imputation, conditional median imputation, and random Imputation.

4. An Example

The proposed method was applied to the modified breast cancer data which was found in some studies [6] [7]. The data set here consist of 46 patients who were treated by the means of Radiation therapy (R) only and 48 other patients who were treated by the means of adjuvant Chemotherapy next to Radiation (R+C). This study aimed to compare cosmetic effects of the treatment through Radiation therapy alone (R) against the combination of Radiation therapy with Chemotherapy (R+C) on women who suffer from breast cancer in its early stages. The event of interest in this study is the time when breast retraction is first observed. The patients were checked in the clinic once every 4 to 6 months, where the actual dates of the event were recorded exactly if available. If the dates are not available, interval of events was noted. To set up data as partly interval censored data, for instance: for radiation, 25 observations were set up as right censored, 21 as interval censored, and 20 as exact. Likewise, for R+C, the set up was for 13 observations as right censored, 35 as interval censored, and 20 as exact, reader refer to [6]. Results of this data set were analyzed as PIC as presented next.

Based on parametric analysis for cancer PIC data, Figures 1, 2, 3, 4, 5, and 6 show the results obtained by random imputation, median imputation, midpoint, left point, right point, and mean imputation techniques respectively. The figures look like Turnbull method especially the one which was obtained by mean and median imputation. Likelihood ratio test shows better result in the case of midpoint, right point, mean, and random imputations as shown in Table 1.

Table 1: Likelihood Ratio Test and their P-value based on parametric model from cancer PIC data

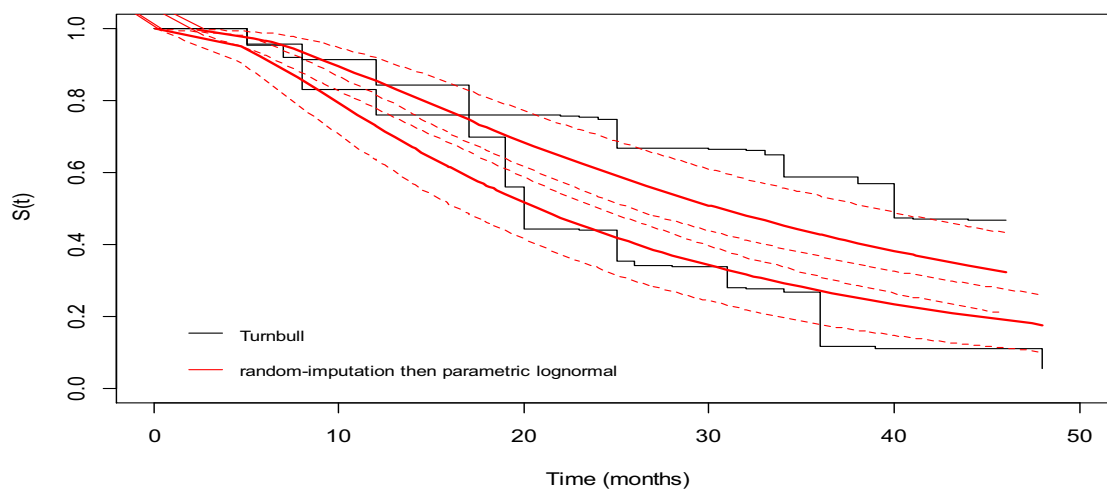| Imputation Techniques	Likelihood Ratio Test	P-value
Midpoint	4.6	0.032
Left-point	3.03	0.082
Right-point	4.74	0.029
Mean	4.49	0.034
Median	4.22	0.04
Random	5.13	0.023

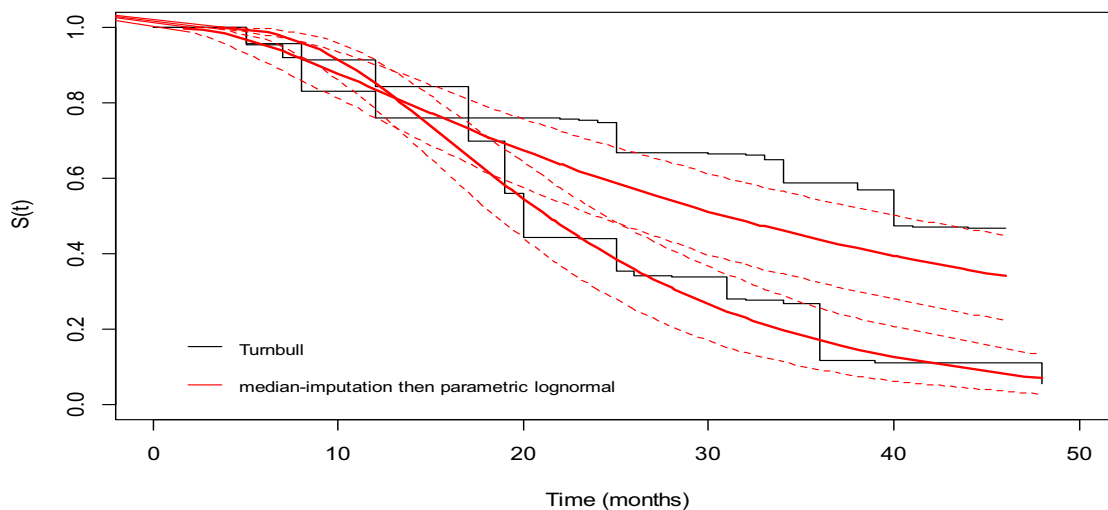**Figure 1:** Estimated survival function obtained by random imputation vs Turnbull based on parametric model from cancer PIC data.

Figure 2: Estimated survival function obtained by median imputation vs Turnbull based on parametric model from cancer PIC data.
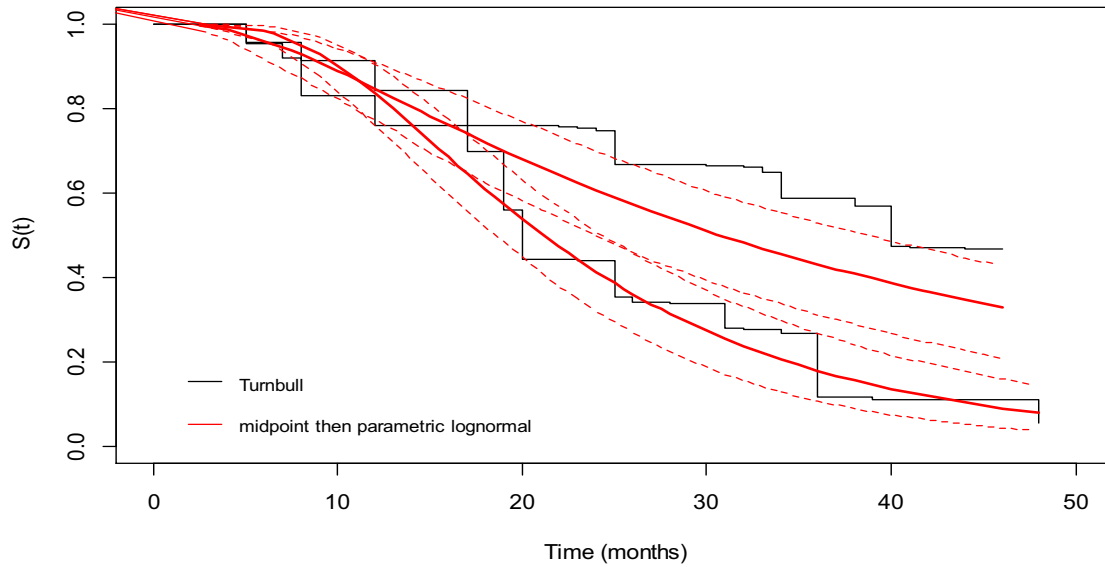
Figure 3: Estimated survival function obtained by midpoint imputation vs Turnbull based on parametric model from cancer PIC data.
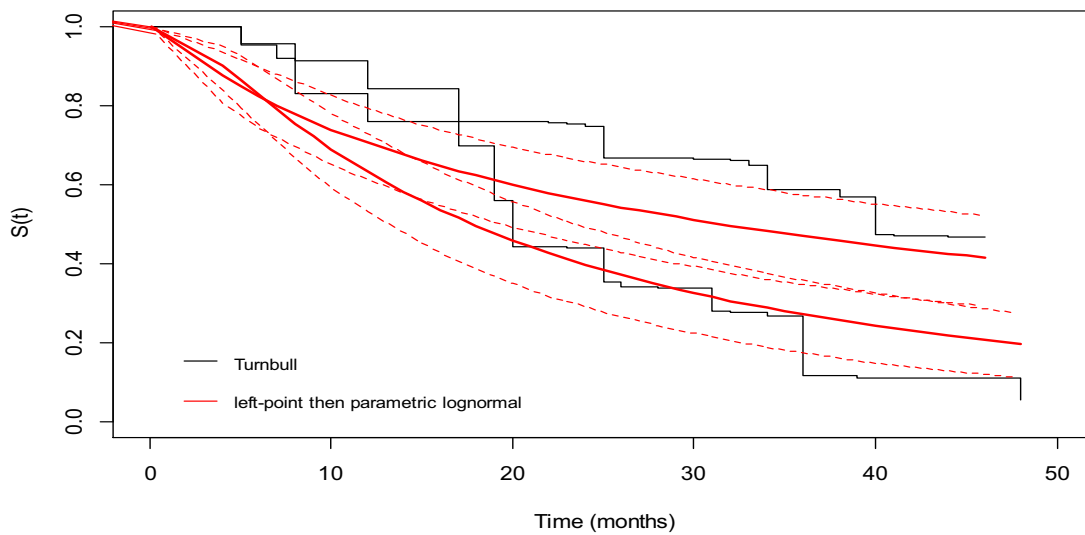

Figure 4: Estimated survival function obtained by left imputation vs Turnbull based on parametric model from cancer PIC data.
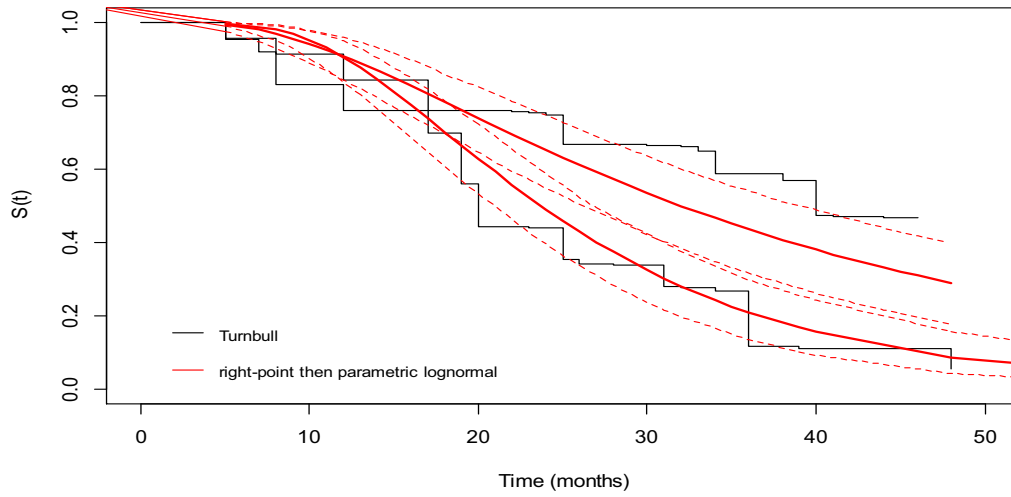
Figure 5: Estimated survival function obtained by right imputation vs Turnbull based on parametric model from cancer PIC data.
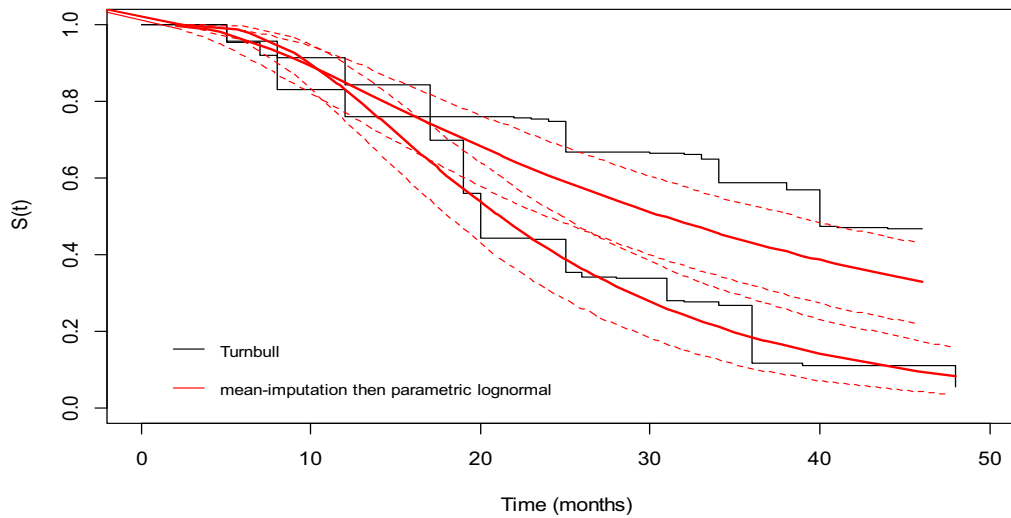

Figure 6: Estimated survival function obtained by mean imputation vs Turnbull based on parametric model from cancer PIC data.

5. Conclusion

In this paper, a simple modification of estimating survival function for PIC data using parametric model based on imputation techniques was proposed. Modification of breast cancer data is used and R software is also used to generate the results. Obtained results are same as the result which is obtained by Turnbull. However, based on partly interval censored data, random imputation and mean & median imputation show better results compared with others imputation techniques as well as Turnbull with respect to the smallest P-value (Table 1).

The parametric Cox model based on imputation techniques gives essentially better estimations of survival function as a Turnbull method, and is also much easier in term of predictions.

References

[1] Huang, J. Asymptotic properties of nonparametric estimation based on partly interval-censored data. *Statistica Sinia*, (1999): 9; 501 – 519.

[2] Kim, J. S. . Maximum likelihood estimation for the proportional hazards model with partly interval-censored data. Journal of the Royal Statistical Society, (2003): 65; 489- 502.

[3] Zhao, X. Zhao, Q.J. Sun. Generalized log-rank test for partly interval-censored failure time data. Biometrical Journal. (2008):3;375-385.

[4] Odell, P. M., Anderson, K. M. and D'Agostino, R. B. Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model. Biometrics. (1992):48;951-959.

[5] Kumar, P. Modeling and analysis of degradation data. Indian Institute of TechnologyBombay,Available:http://www.academia.edu/266427/Modelling_and_Analysis_of_Degradation_Data [Accessed: 5/1/2013]. (2010).

[6] Abdallah M T Zyoud. Imputation Techniques for Reliability Analysis Based On Partly Interval Censored Data. M.Sc., thesis. International Islamic University Malaysia. (March 31, 2017).

[7] Finkelstein, D. M., and Wolfe, R. A. A semi parametric model for regression analysis of interval-censored failure time data. Biometrics. (1985): 41;933-945.