Malaysian Journal of Mathematical Sciences 14(2): 285–293 (2020)

MALAYSIAN JOURNAL OF MATHEMATICAL SCIENCES

Journal homepage: http://einspem.upm.edu.my/journal

Nonparametric Estimation of a Survival Function with Interval Censored Data

Aljawadi, B. A.

Department of Mathematics, Hebron University, Palestine

E-mail: baderj@hebron.edu

Received: 20 March 2019
Accepted: 4 March 2020

ABSTRACT

Censored data represent a general problem in many fields especially in medical survival analysis. Where in some cases the censored data replaced with the exact data, which will distort the real distribution of the data set. The disjoint interval censored data are a special case of the general form of interval censoring which are found in a variety of applications including grouped data and survey responses. However, little attention has been given to their analysis even though they are a recurrent type of data. In contrast of Turnbull's standard non-parametric method for estimation of survival function in case of disjoint interval censored data, an alternative approach for the estimation of survival function developed in this study. This approach investigated by optimizing the non-parametric maximum likelihood function without any iterative numerical algorithms, where a simple closed form solution to the non-parametric maximum likelihood function exist. the advantages of the proposed estimation approach are illustrated using real data set with some other examples.

Keywords: Disjoint interval censoring, survival function, covariates and Hessian matrix.

1. Introduction

Interval censoring is a very common data problem which are often found in longitudinal studies. A special case of interval censoring is disjoint interval censored data, where subjects are assessed only periodically for the response of interest. The time when the interested event occurs (i.e failure time) is not exactly observed but the given information is only that this time place within a set of non-overlapping intervals, and hence the set of data comprised of several discrete class intervals. For example, in a clinical trial individuals may asked to visit a clinic for assessment at predetermined visiting times. Thus, the interested event is not fully observed and it is only known that it belongs to the preassigned intervals.

Disjoint interval censored data are also found in many other applications such as in the compilation and summarization of a continuous random variable on the form of a frequency table which is the typical case with population-wide surveys. (see in Aljawadi et al. (2012))

The analysis of interval censored data are commonly analyses using special statistical techniques using both parametric and non-parametric maximum likelihood approaches (see Lindsey and Ryan (1998), Day (2007),Turnbull (1976)), where the parametric approach relies on a prior assumption about underlying distribution of the data set. But in real life the assumption of data distribution is easy to violate and the analysis may generate inconsistent estimates as a result of misspecification of the distribution function. Therefore, the non-parametric maximum likelihood techniques is the viable alternative since no distribution function is imposed on the data. In the non-parametric approach, the survival function is estimated and then it can be used to estimate the empirical distribution function (see Wu (2008)).

Disjoint interval censored data is a recurrent type of data and little attention has been given to the non-parametric analysis of such data sets in the non-parametric literature, Chen et al. (2013b). In this study, an alternative approach for survival function estimation is developed using the non-parametric maximum likelihood function where no iterative numerical algorithms are required or any advanced statistical software packages. While the solution of the non-parametric problem is given in a simple closed form. The proposed technique is a contrast of the standard non-parametric approach (Turnbull's method).

2. Methodology

Suppose that T_i is the unobserved survival time of interest for subject i, $1 \leq i \leq n$, and let $(M_{j-1}, M_j], j = 1, 2, ..., m$, be the observed interval for which T_i is observed, every respondent i is presented with m disjoint open intervals and the survival curve $S(t) = P(T_i > t)$ is estimated. Consequently, every t_i is observed to fall into one of the intervals $(M_\circ, M_1], (M_1, M_2], ..., (M_{j-1}, M_j]$ and the probability that t_i is in the j^{th} interval with boundary values M_{j-1} and M_j is given by:

$$P(M_{j-1} \leq t_i \leq M_j) = S(M_{j-1}) - S(M_j), i = 1, ..., n, j = 1, ..., m,$$

let α_{ij} be a dummy variable that indicates whether the j^{th} interval contains the i^{th} failure time such that

$$\alpha_{ij} = \begin{cases} 1 & if\ t_i \in (M_{j-1}, M_j). \\ 0 & otherwise \end{cases}$$

In the case in which t_i is not observed at one or several intermediate intervals. Then in such case, the intervals with no observation need to be merged (i.e. pooled) as follows:

- Identify the 1^{st} and m^{th} intervals by the first and last non-empty intervals respectively whatever the number of empty intervals before or after them.
- Identify the intermediate intervals with no observations.
- If no observations found in the $(m+1)^{th}$ interval then the m^{th} and $(m+1)^{th}$ intervals are merged into one interval containing M_k individuals with limit values of m_{k-1} and m_{k+1}. Continue in the same manner until all remaining intervals are pooled sufficiently and have observations, and then applying the proposed procedure for estimating the survival function.

For these assumptions we may consider the likelihood function conditional upon the observed intervals, where the likelihood function can be represented as follows: (see Chen et al. (2013a))

$$\begin{aligned} L(S) &= \prod_{i=1}^{n}\Big(\sum_{j=1}^{m}\alpha_{ij}[S(M_{j-1}) - S(M_j)]\Big) \\ &= \prod_{i=1}^{n}\Big(\sum_{j=1}^{m}\alpha_{ij}[S_{j-1} - S_j]\Big) \end{aligned}$$

and hence the log-likelihood function can be written as:

$$l(S/n) = \sum_{i=1}^{n} log\Big(\sum_{j=1}^{m}\alpha_{ij}[S_{j-1} - S_j]\Big). \tag{1}$$

In the parametric approach it is assumed that the failure time T follows a particular statistical distribution with parameter vector ϕ, such as exponential, Weibull, log-normal and some other distributions. However, the optimization algorithms are used to find the vector ϕ that maximizes the log-likelihood function.

While in the nonparametric maximum likelihood procedure there is no assumption about the probability distribution for the interested variable T and hence survival function is unknown and the nonparametric procedure considers the survival function $S(t)$ is a parameter to be estimated. Moreover, the maximum likelihood estimation needs to be expressed as a constrained maximization problem such that

$$Max\Big[l(S/n) = \sum_{i=1}^{n} log\Big(\sum_{j=1}^{m}\alpha_{ij}[S_{j-1} - S_j]\Big)\Big] \tag{2}$$

subject to the constraint: $1 = S_\circ \geq S_1 \geq S_2 \geq ... \geq S_k = 0$.

The estimates $\hat{S}$ are usually obtained using the self-consistent algorithm proposed by Turnbull (see Elfaki et al. (2013)) with some limitations of this procedure and some alternatives to the non parametric maximum likelihood problem which discussed in Elfaki et al. (2013). However, the maximization problem in equation 2 and in some special cases when a set of disjoint closed intervals are available, there is a closed form solution in the contrast of the claims by some authors that no closed form solutions can be generated to the problem described in equation 2. (see Wu (2008), Elfaki et al. (2013))

The unconstrained version of the maximization problem in equation 2 can be written as:

$$\begin{aligned} l(S/n) &= \sum_{i=1}^{n} log\Big(\sum_{j=1}^{m} \alpha_{ij}[S_{j-1} - S_j]\Big) \\ &= \sum_{j=1}^{m} M_j log(S_{j-1} - S_j), \end{aligned} \tag{3}$$

where $S_\circ = 1, S_m = 0$ and M_j is the number of individuals who chose the M^{th} interval.

The first order conditions for the equation in equation 2 are given by:

$$\begin{aligned} \frac{\partial l}{\partial S_j} &= \frac{M_j}{S_{j-1} - S_j} - \frac{M_{j+1}}{S_j - S_{j+1}} = 0, j = 1, ..., m-1 \\ &= -M_{j+1}S_{j-1} + \Big(M_j + M_{j+1}\Big)S_j - M_jS_{j+1} = 0. \end{aligned} \tag{4}$$

The equations in equation 4 are a system of $m-1$ linear equations that can be expressed in a matrix form as follows:

$$MS = \Pi \tag{5}$$

where

$$M = \begin{bmatrix} (M_1 + M_2) & -M_1 & 0 & \dots & 0 \\ -M_3 & (M_2 + M_3) & -M_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & (M_{j-1}, M_j) \end{bmatrix}$$

$$S = \begin{bmatrix} S_\circ \\ S_1 \\ \vdots \\ S_{j-1} \end{bmatrix} \; and \; \Pi = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ M_{j-1} \end{bmatrix}.$$

However, the solution to system of equations in equation 5 is the vector $\hat{S} = M^{-1}\Pi$ and it is easy to show that the j^{th} element of $\hat{S}$ can be given by:

$$\hat{S_\omega} = \frac{\sum_{j=1}^{\omega} M_j}{n}, \qquad \omega = 1, ..., m-1, \qquad and \qquad n = \sum_{j=1}^{m} M_j. \tag{6}$$

The estimates in equation 6 compared to those obtained by Turnbull procedure can be estimated simply without any numerical technique which is an advantage of the nonparametric maximum likelihood problems under the proposed situations. It is also important to note that the solutions vector in equation 6 ensures that $0 < \hat{S_j} < 1$ and $\hat{S_j} < \hat{S_{j-1}}, \forall j$ which satisfying the constrains involved in equation 2.

The variance of the estimates of survival function $var(\hat{S})$ is given by the inverse of $-E[H(S)]$, where $H(S)$ is the *Hessian* matrix.

The Hessian matrix is the matrix of partial derivatives of the first order condition given in (2.4) with respect to $S_{j's}$ such that:

$$\frac{\partial^2 l}{\partial S_j^2} = \frac{\partial^2}{\partial S_j^2}\Big[\frac{M_j}{S_{j-1} - S_j} - \frac{M_{j+1}}{S_j - S_{j+1}}\Big], j = 1, ..., m-1 \tag{7}$$

and

$$\frac{\partial^2 l}{\partial S_j \partial S_{j+1}} = \frac{\partial^2}{\partial S_j \partial S_{j+1}}\Big[\frac{M_j}{S_{j-1} - S_j} - \frac{M_{j+1}}{S_j - S_{j+1}}\Big], \tag{8}$$

such that

$$Var(\hat{S}) = \Big(-E[H(S)]\Big)^{-1} =$$

$$n^{-2}\begin{bmatrix} \left(\frac{1}{M_1} + \frac{1}{M_2}\right) & \frac{-1}{M_2} & 0 & \cdots & 0 \\ \frac{-1}{M_2} & \left(\frac{1}{M_2} + \frac{1}{M_3}\right) & \frac{-1}{M_3} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \frac{-1}{M_j} & \left(\frac{1}{M_{j-1}} + \frac{1}{M_j}\right) \end{bmatrix}^{-1} \tag{9}$$

3. Simulation

Let T_i denote a non-negative continuous random variable representing survival time of the subject i with a known survival function. The observation on T is interval censored when the exact value of T is unknown and it is only known that it belongs to an observed interval (L, R). L and R may only be certain predetermined points or discrete follow up times, because it is impossible to observe subjects continuously. However, to simulate interval censored data, we define a set of potential inspection times assuming that the subjects are inspected at these times regularly.

The data set consists of 1000 observations using a normal distribution with mean $\mu = 50$ and standard deviation $\sigma = 30$. The generated observations were then allocated into 10 class intervals: $[0, 9.99], [10, 19.99], [20, 29.99], [30, 39.99], [40, 49.99], [50, 59.99], [60, 69.99], [70, 79.99], [80, 89.99], [90, 99.99]$. For empty interval pooling procedure can be used to merge these intervals as follows:

- For $I = 2 \rightarrow I - 1$, identify empty intervals.
- If no observation found in the $(I+1)^{th}$ interval then the I^{th} and $(I+1)^{th}$ intervals need to be merged into one interval containing N_k observations with boundary values $L_{(k-1)}$ and $R_{(k+1)}$.
- Continue in the same manner until intervals pooled sufficiently so that the remaining intervals have observations.

The corresponding non-parametric survival function $S(t)$ estimates, as well as the distribution of the data set and their standard errors are estimated and shown in the following table. The $S(t)$ values can be directly estimated using the raw proportions of observations belonging to each category. Estimated survival values and the corresponding standard errors are evaluated based on equations 6 and 9.

Table 1: Survival Probabilities of the Simulated Data

Interval	Number of Observations	S(t)	Standard Error
0-9.99	19	0.988	0.0060
10-19.99	44	0.957	0.011
20-29.99	62	0.897	0.090
30-39.99	85	0.806	0.0130
40-49.99	320	0.530	0.0230
50-59.99	276	0.210	0.0212
60-69.99	91	0.125	0.0194
70-79.99	60	0.063	0.0106
80-89.99	31	0.019	0.0034
90-100	12		

Furthermore, to demonstrate the efficiency of the proposed technique a comparison between the proposed approach for survival function approximation and the generalized Turnbull estimator is shown in the figure below.

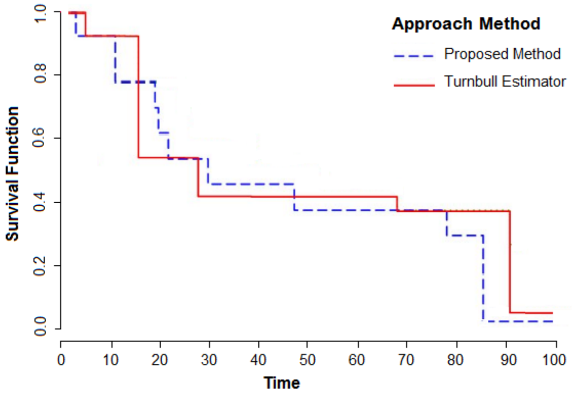

Figure 1: Survival probabilities based on the proposed approach and Turnbull algorithm

4. Summary and Results

Interval censoring usually represents a sampling scheme or an incomplete data structure which are found in a variety of applications, from applications in survey responses to grouped data represented to report population-wide surveys; thus the robust and practical approach to analysis of this type of data is necessity. In this article an alternative approach to estimate the empirical survival estimates of the variables of interest is developed by optimizing their corresponding nonparametric maximum likelihood function. The focus is on a special censoring model where the observed data are a set of disjoint intervals. This approach in contrast to Turnbull standard nonparametric method does not require iterative numerical procedures or the use of advanced statistical software packages. The computations are very intuitive and easy to compute even though this type of analysis tends to be sensitive to the number of intervals used, as well as to the values of the interval boundaries.

References

Aljawadi, B. A., Bakar, M. R. A., and Ibrahim, N. A. (2012). Nonparametric versus parametric estimation of the cure fraction using interval censored data. *Journal Communications in Statistics-Theory and Methods*, 41(23):4251–4275. doi: 10.1080/03610926.2011.569678.

Chen, D. G., Lili, Y., Paece, K. E., Lio, Y. L., and Wang, Y. (2013a). Approximating the baseline hazard function by taylor series for interval censored time to event data. *Journal of Biopharmaceutical Statistics*, 23:695–708.

Chen, D. G., Sun, J., and Peace, K. E. (2013b). Interval-censored time-to-event data: Methods and applications. *Boca Raton, FL: Chapman and Hall.*

Day, B. (2007). Distribution-free estimation with interval-censored contingent valuation data: Troubles with turnbull? *Environmental and Resource Economics*, 37:777–795.

Elfaki, F. A. M., Abobakar, A., Rizam, M., and Usman, M. (2013). Survival model for partly interval-censored data with application to anti d in rhesus d negative studies. *International Journal of Biological, Biomolecular, Agricultural, Food and Biotechnological Engineering*, 7(5):347–350.

Lindsey, J. C. and Ryan, L. M. (1998). Tutorial in biostatistics methods for interval-censored data. *Statistics in Medicine*, 17(2):219–238.

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

Wu, S. F. (2008). Interval estimation for a pareto distribution based on a doubly type-ii censored sample. *Computational Statistics and Data Analysis*, 52:3779–3788.