Malaysian Journal of Mathematical Sciences 11(3): 303–315 (2017)

MALAYSIAN JOURNAL OF MATHEMATICAL SCIENCES

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

Approximation of Survival Function by Taylor Series for General Partly Interval Censored Data

Aljawadi, B. A.

Department of Mathematics, Hebron University, Palestine

E-mail: aljawadi@yahoo.com

Received: 2 February 2017
Accepted: 31 August 2017

ABSTRACT

In follow up studies including demographical, epidemiological, financial, medical and sociological studies, partly interval censored data is a common practice where observed data include both exact and interval-censored observation on the survival time of interest. This paper discusses approximating survival function under general partly interval censoring model when the covariates are considered in the analysis. The Taylor series is used to approximate the baseline hazard function in the Cox proportional hazard model. The method is evaluated using simulation studies. The results indicate that the approach performs well for practical situations and comparable to the existing methods.

Keywords: General partly interval censoring, Hazard function, Survival function and Taylor series.

1. Introduction

The survival function is a basic quantity employed to describe the probability that an individual survives beyond a specified time. In other words, this is the amount of time until the event of interest occurs. Estimating survival functions has interested statisticians for numerous years. The estimation of survival function has increased greatly over the last several decades because of its large usage in areas related to biostatistics and the pharmaceutical industry. Estimation of survival function can be done via parametric or nonparametric approaches, where in the case of parametric estimation a reasonable parametric model can be assumed and the estimation problem is relatively easy. However, in real life scenarios the exact distribution of the data is usually unknown. In such case, using a nonparametric method is a common alternative, since the nonparametric estimator does not assume that the data come from a specified distribution.

Under various censoring models estimation of survival function is a common practice. Where, in censored data it is not possible to obtain complete information for the entire group of units in the study. Different censoring types arise depending on the way of data collection from the experiment. One of the most common censoring type is general partly interval censoring which studied extensively in Elfaki et al. (2013), Huang (1999), Kim (2003) and Zhao et al. (2008) and so many others. By partly interval censored failure time data we mean, for some subjects, the exact failure times are exactly observed, but for the remaining subjects, the survival time of interest is not observed and the only available information is that the failure times belong to an interval $[t_L, t_R]$; Chen et al. (2012), Chen et al. (2013) and Sun et al. (2005). General partly interval censored data arise often in follow up studies. An example of such data is provided by the Framingham Heart Disease Study; see Elfaki et al. (2013) for a description. In this study, times of the first occurrence of anti D infection through to contaminated blood factor disease patients are of interest. For some patients, time of the first occurrence of infection is recorded exactly. But for others, time is recorded only between two clinical examinations.

2. Model Formulation

Assume that the failure times, $t_1, t_2, ..., t_n$ are independent and identically distributed. If all the random variables are observable, then it is well known that the nonparametric maximum likelihood estimator ($NPMLE$) of $S(t)$ is the empirical survival function. Where estimation the survival function in case of general partly interval censored data is the main goal of this article. How-

ever, for the general partly interval censoring assume that there are m potential examination times $E_1 < E_2 < ... < E_m$ for n participants enrolled in a study then, the exact failure times for n_1 participants will be observed, while the failure times of $n_2 = (n - n_1)$ participants are not observed and that the failure time T is known to be bracketed between two adjacent examination points. Exact failure times mean that any patient has the event of interest during the inspection times or the patient condition necessitates hospital examination where the event of interest is recorded exactly. Interval-censored failure time measurements mean that the event of interest occurs between two examination times, (L_i, R_i) where L_i and $R_i \in (E_1, E_2, ..., E_m)$ and $L_i < R_i$ has a probability of one. If the failure time of the patient occurs before the first examination, then left-censored will be observed, i.e. $t_i \in (0, L_i)$. If the patient did not have the failure time before or during the final examination, then right-censored will be observed, $t_i \in (R_i, \infty)$. Additionally, censoring is presumed independent of the examination time. For the general partly interval-censored i^{th} patient, then, the likelihood function is:

$$L(\theta) = \prod_{i=1}^{n_1} [f(t_i; \theta)] \prod_{i=n_1+1}^{n} [S(L_i; \theta) - S(R_i; \theta)] \tag{1}$$

Where $f(t) = S(t-) - S(t)$ is the mass that S puts at t.

To detect the effect of the covariates on the behaviour of the survival function, then Cox proportional hazard model is necessity due to the vital relation between the hazard and survival functions. However, the survival function is the function that gives the probability of being alive just before duration t such that

$$S(t) = Pr(T \geq t) = 1 - Pr(T < t) = 1 - F(t) = \int_t^{\infty} f(x)dx \tag{2}$$

and as an alternative characterization of the distribution of T is given by the hazard function defined as

$$\lambda(t) = \lim_{dt \to 0} \frac{Pr(t \leq T < t + dt / T \geq t)}{dt} \tag{3}$$

The conditional probability in the numerator may be written as the ratio of the joint probability that T is in the interval $[t, t + dt)$ and $T \geq t$ (which is, of

course, the same as the probability that t is in the interval), to the probability of the condition $T \geq t$. The former may be written as $f(t)dt$ for small dt, while the latter is $S(t)$ by definition. Dividing by dt and passing to the limit gives the useful result

$$\lambda(t) = \frac{f(t)}{S(t)} \tag{4}$$

Which is the hazard function that describes the rate of occurrence of the event at duration t.

When a set of covariates are available in the data set say $Z = (z_1, z_2, ..., z_p)$ then the Cox proportional hazard model can be expressed as follows Ping et al. (1998):

$$\lambda(t/Z, \beta, \theta) = \lambda_\circ(t/\theta)exp(\beta^T Z) \tag{5}$$

Where $\lambda_\circ(t/\theta)$ is the baseline hazard function, $\beta = (\beta_1, \beta_2, ..., \beta_p)^T$ is the parameters vector of the corresponding covariates and θ is the parameters vector of the baseline hazard function.

Note that from equation (2) $-f(t)$ is the derivative of $S(t)$. This suggests rewriting the hazard function given in equation (4) as

$$\lambda(t) = -\frac{d}{dt} log S(t) \tag{6}$$

This expression can be solved to obtain the probability of surviving to duration t as a function of the hazard at all durations up to t:

$$S(t) = exp\Big(- \int_0^t \lambda(x)dx \Big) \tag{7}$$

The integral in brackets is called the *cumulative hazard function* denoted by $\gamma(t)$, which can be defined as follows:

$$\begin{aligned} \gamma(t/Z, \beta, \theta) &= \int_0^t \lambda(x/Z, \beta, \theta)\, dx \\ &= exp(\beta^T Z) \int_0^t \lambda_\circ(x/\theta)dx \\ &= \gamma_\circ(t/\theta)exp(\beta^T Z) \end{aligned} \tag{8}$$

where $\gamma_\circ(t/\theta)$ is the cumulative baseline hazard function.

Based on the above notifications the survival function can be defined as:

$$\begin{aligned} S(t/Z,\beta,\theta) &= exp\Big[-\gamma_\circ(t/\theta)exp(\beta^T Z)\Big] \\ &= S_\circ(t/\theta)^{exp(\beta^T Z)} \end{aligned} \tag{9}$$

where $S_\circ(t/\theta)$ is the baseline cumulative survival function that has no obvious relation with the baseline cumulative hazard function as it is shown in the following expression:

$$S_\circ(t/\theta) = exp\Big(-\int_0^t \lambda_\circ(x/\theta)dx\Big) = exp[-\gamma_\circ(t/\theta)] \tag{10}$$

These results show that the survival and hazard functions provide alternative but equivalent characterizations of the distribution of T. Given the survival function, we can always differentiate to obtain the density and then calculate the hazard using Equation (4). Given the hazard, we can always integrate to obtain the cumulative hazard and then exponentiate to obtain the survival function using Equation (9).

However, based on the given materials then, the log-likelihood function can be written as follows:

$$\begin{aligned} l(\theta) &= \sum_{i=1}^{n_1} -dS(t_i;\theta) \sum_{i=n_1+1}^{n} \Big[S_\circ(L_i/\theta)^{exp(\beta^T Z)} - S_\circ(R_i/\theta)^{exp(\beta^T Z)}\Big] \\ &= \sum_{i=1}^{n_1} [S(t_{i-}) - S(t_i)] \sum_{i=n_1+1}^{n} \Big[S_\circ(L_i/\theta)^{exp(\beta^T Z)} - S_\circ(R_i/\theta)^{exp(\beta^T Z)}\Big] \\ &= \sum_{i=1}^{n_1} \Big[S_\circ(t_{i-}/\theta)^{exp(\beta^T Z)} - S_\circ(t_i/\theta)^{exp(\beta^T Z)}\Big] \\ &\quad \sum_{i=n_1+1}^{n} \Big[S_\circ(L_i/\theta)^{exp(\beta^T Z)} - S_\circ(R_i/\theta)^{exp(\beta^T Z)}\Big] \end{aligned} \tag{11}$$

Maximization of the log-likelihood function can be obtained numerically since no explicit form of the maximum likelihood estimators can be found.

Where, in case of large number of covariates, numerical techniques may produce high levels of bias. Thus, the proposed procedures might be used under some warnings unless an advanced maximization procedure can be adopted.

Maximization of the log-likelihood function depends on the estimation of the baseline hazard function which can be replaced by any adequate distribution such as Weibull, log-normal distributions, or assuming that the baseline survival function is to be piecewise constant which leads to the semi-parametric approach and this technique is available in some statistical software's such as *R*. But under this assumption the continuity of the baseline survival function is violated and it is only a step function. Recently, some researchers employed Taylor series to approximate the base line survival function in attempt to get a smoother function and this approach will be employed for such purpose.

3. Taylor Approximation

In mathematics, functions can be represented by Taylor series which is an infinite sum of terms that are evaluated from the derivatives of the function at a single point. However in this article, Taylor series is used to approximate the baseline survival function by using a finite number of terms of its Taylor series, where the optimal order of this function can be obtained based on the likelihood ratio test.

However, estimation of the baseline survival function can be easily obtained via the estimation of the baseline hazard function as a result of the direct combination between the two functions. Thus, the Taylor series will employ to estimate the baseline hazard function and for simplicity the logarithm of the baseline hazard function will be approximated as follows:

$$\Gamma_{\circ}(t/\theta) = log(\lambda_{\circ}(t/\theta)) = a_{\circ} + a_1 t + \frac{a_2}{2!}t^2 + ... + \frac{a_q}{q!}t^q$$

Where $\Gamma_{\circ}(t/\theta)$ denotes the Taylor series of order q and $\theta = (a_{\circ}, a_1, ..., a_q)$ represents the baseline parameters vector to be estimated. Thus, the baseline cumulative survival function can be obtained as

$$\begin{aligned} S_\circ(t/\theta) &= exp\Big(- \int_0^t \lambda_\circ(y/\theta)dy\Big) \\ &= exp\Big(- \int_0^t exp[\Gamma_\circ(y/\theta)]dy\Big) \\ &= exp\Big(- \int_0^t exp[a_\circ + a_1 y + \frac{a_2}{2!}y^2 + ... + \frac{a_q}{q!}y^q]dy\Big) \end{aligned} \tag{12}$$

The baseline survival function $S_\circ(t/\theta)$ can be involved in the log-likelihood function defined in equation (11) and maximizing the log-likelihood function using maximum likelihood theory to make statistical inference for the desired parameters. The optimal number of terms of baseline hazard function Taylor series can be assigned following the procedure below Chen et al. (2013):

1. Fitting the likelihood function when ($i.e \quad q = 0$) and maximize the likelihood function with respect to the parameters $\hat{\beta}$ and $\hat{\theta} = \hat{a_\circ}$, and denote the fitted value of the likelihood function as $h_\circ = max[l(\hat{\beta}, \hat{\theta})]$.
2. Fitting the likelihood function with one more order of Taylor series ($i.e \quad q = 1$) where the parameters in such case are $\hat{\beta}$ and $\hat{\theta} = (\hat{a_\circ}, \hat{a_1})$ and similar to step (1), the fitted value of the likelihood function represented by $h_1 = max[l(\hat{\beta}, \hat{\theta})]$.
3. Given a preassigned significant level $\alpha = 5\%$ and for degrees of freedom (df) of the Chi square distribution equals to 1, then if $-2(h_\circ - h_1) < \chi^2_{1,(1-\alpha)}$ the selected order of Taylor series is $q = 0$ and hence the maximum likelihood estimates of the parameters are $\hat{\beta}$ and $\hat{\theta} = \hat{a_\circ}$. Otherwise, new estimates of the parameters at $q = 2$ can be obtained, and denote the new fitted value of the likelihood function as $h_2 = max[l(\hat{\beta}, \hat{\theta})]$ and then follow to step (3) again using h_1 and h_2 values.

Repeat this process until a stopping condition such as $-2(h_{q^*-1} - h_{q^*}) < \chi^2_{1,(1-\alpha)}$, where the desired order of Taylor series is $(q = q^* - 1)$ and hence the desired parameters are $\hat{\beta}$ and $\hat{\theta} = (\hat{a_\circ}, \hat{a_1}, ..., \hat{a}_{q^*-1})$.

4. Simulation

The simulation studies based on general partly interval-censored data involve further steps in comparison with the other types of censoring. How-

ever, the data generation in this simulation conducted based on the simulated biomedical clinical trials. For the sake of flexibility, various censoring rates are considered to detect how the pattern of the survival function approximation would progress under fluctuated censoring rates. Each data set comprised 100 observation and to control the generation process it is assumed that the true survival time t, followed an exponential distribution. The following assumptions are made before moving on to the simulation algorithm:

- There are m potential inspection times which are known by design.
- Some subjects may not observed in the first inspection time (*i.e.* possibility of left censoring)
- Survival times are generated from a known $S(t)$.

However, for each data set, the simulation is performed based on the following procedure:

1. A random sample of one hundred observation are randomly generated from binomial distribution with $p = 0.5$ and classified into two different groups; placebo and drug treatment groups, where an indicator variable z is used such that $z_1 = 0$ for placebo group and $z_1 = 1$ for drug treatment group. Another two continuous covariates z_2 and z_3 are generated based on normal distribution with randomly selected parameters (Mean= 2, Variance = 1).
2. The failure time (t) for each patient is generated from exponential distribution with scale parameter $\lambda = exp(b_\circ + b_1 z_1 + b_2 z_2 + b_3 z_3)$, where the initial values of the parameters vector is set to $(b_\circ = 0, b_1 = 0.5, b_2 = 0.5, b_3 = 0.5)$.
3. A set of potential inspection times $E_1 < E_2 < ... < E_{20}$ are generated assuming that the subjects are inspected or examined at these times. The first inspection time, E_1, was generated from a uniform distribution; $U(0, 0.1)$. The next examination time $E_2 \sim U(E_1, E_{1+0.1})$. Afterwards, the consequent examination times were generated in the same manner such that $E_m \sim U(E_{(m-1)}, E_{(m-1)+0.1})$. The patients would assist to each of these scheduled examination times with probability p.
4. Creating the left and right endpoints for all n intervals of a 100×2 empty matrix named I. Such that $I[\ ,1]$ = Left boundaries and $I[\ ,2]$ = Right boundaries. For $i = 1, ..., 100$ and $j = 1, ..., 20$

$$I[i,1] = \begin{cases} 0 & : if\ t[i] < E_1 \\ E_j & : if\ E_j < t[i] < E_{j+1} \\ E_{20} & : if\ t[i] > E_{20} \end{cases}$$

$$I[i,2] = \begin{cases} E_1 & : if\ t[i] < E_1 \\ E_{j+1} & : if\ E_j < t[i] < E_{j+1} \\ \infty & : if\ t[i] > E_{20} \end{cases}$$

5. The censoring indicator ϵ is generated in the most common manner based on the intervals generated in the previous step. However, creating the vector of censoring indicator (ϵ) for all members based on the I matrix such that:

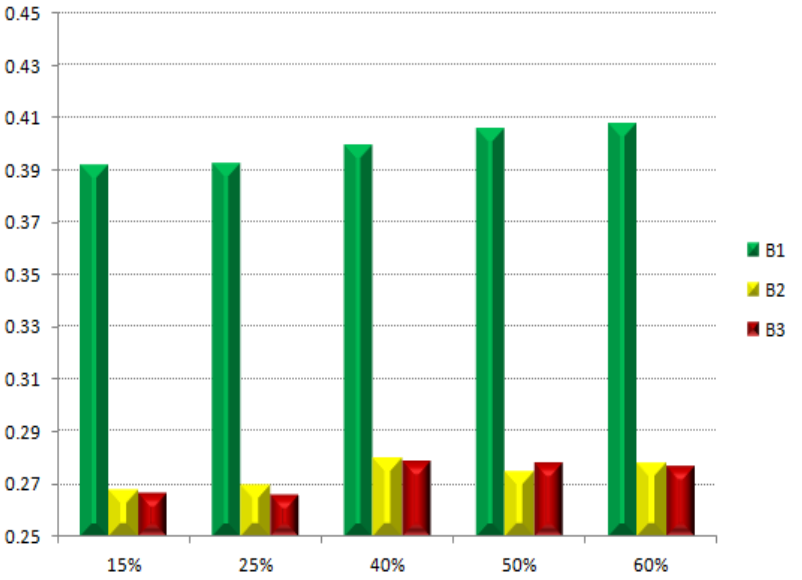
$$\epsilon[i] = \begin{cases} 0 & : if\ I[i,2] = \infty \\ 1 & : otherwise \end{cases}$$

If the failure time is less than E_1 then t is left censoring, while if the failure time is greater than E_{20} then t is right censoring, otherwise t is exact failure time. The simulation carried out 1000 times, where the mean square error and the coverage probability for the estimated parameters are also computed from 1000 bootstrapping samples for some random censoring rates. The simulation is set-up in the R software and the results are shown in the following table and graph.

Table 1: The average of estimated parameters, mean square error and empirical coverage probability.

Censoring Rate (P)		**Average of Estimated Parameters**	MSE	**Empirical Coverage Probability** (CP)
$P = 15\%$	β_1	0.261	0.342	0.961
	β_2	0.317	0.365	0.950
	β_3	0.376	0.373	0.947
$P = 25\%$	β_1	0.282	0.419	0.971
	β_2	0.562	0.606	0.922
	β_3	0.795	0.632	0.940
$P = 40\%$	β_1	0.321	0.843	0.951
	β_2	0.779	0.853	0.911
	β_3	0.898	0.982	0.903
$P = 50\%$	β_1	0.425	1.280	0.882
	β_2	1.795	2.484	0.841
	β_3	1.907	3.741	0.836
$P = 60\%$	β_1	0.447	4.176	0.841
	β_2	1.428	7.432	0.812
	β_3	2.357	10.43	0.801

Figure 1: The Estimated parameters versus censoring rate



In the above table and graph it is very clear that the type of treatment (β_1) has the highest impact on the estimation of the survival function as it can be easily explored from the estimated parameters values. Furthermore, the results reveal the dramatic increasing of the mean square error as a result of the increment in the censoring rate in the data set. The increment pattern of the mean square error for the parameters is more distinctive for high censoring rates (i.e more than 40%) which indicates that the results may distorted under the proposed estimation approach once a high rate of censored observations is found in the considered data set, especially when continuous covariates are considered as it is shown in the table above where the mean square error for β_2 and β_3 belongs to the two continuous covariates have the highest errors compared to β_1. Consequently, a concordance of the conclusions can be found when the coverage probability is reviewed, where the coverage probability of the parameters based on the pre-described maximum likelihood theory given at equation (7) are not much satisfactory when the censoring rate exceeds 40% especially for the parameters from the normal approximation, however, this drawback might be avoided once we resort to an intensive bootstrapping approach, but this is not guaranteed in the existence of heavy censoring in the data sets.

5. Conclusion

In this article, a non-parametric approach for survival function approximation is discussed when the data set consists of general partly interval censored observations. The analysis handled in the presence of some covariates where Taylor series is proposed to approximate the baseline hazard function in Cox proportional hazards regression to mitigate the bias arising from analyzing the imputed time-to-event data. With this formulation, the likelihood ratio test can be used to select an appropriate order for this Taylor series approximation and maximum likelihood techniques used to estimate model parameters and provide statistical inference. The application of this novel method is demonstrated by a simulation study, where the obtained results showed that the proposed method works well for groups of data with low censoring rates and some attention may be paid once a higher censoring rate is available.

References

Chen, D. G., Lili, Y., Paece, K. E., Lio, Y. L., and Wang, Y. (2013). Approximating the baseline hazard function by taylor series for interval censored time to event data. *Journal of Biopharmaceutical Statistics*, 23(3):695–708. DOI:10.1080/10543406.2012.756497.

Chen, D. G., Sun, J., and Peace, K. E. (2012). *Interval-Censored Time-to-Event Data: Methods and Applications*. Chapman and Hall/CRC, USA.

Elfaki, F. A. M., Abobakar, A., Azram, 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. url: http://waset.org/Publications?p=77.

Huang, J. (1999). Asymptotic properties of nonparametric estimation based on partly interval- censored data. *Statistica Sinica*, 9(2):501–519.

Kim, J. S. (2003). Maximim likelihood estimation for the proportional hazards model with partly interval-censored data. *Journal of the Royal Statistical Society: Series B*, 65(2):489–502.

Ping, H., Tsiatis, A., and Davidian, M. (1998). Estimating the parameters in the cox model when covariate variables are measured with error. *Biometrics*, 54(4):1407–1419.

Sun, J., Zhao, Q., and Zhao, X. (2005). Generalized log-rank tests for interval-censored failure time data. *Scandinavian Journal of Statistics*, 32(1):49–57.

Zhao, X., Zhao, Q., Sun, J., and Kim, S. J. (2008). Generralized log-rank test for partly interval-censored failure time data. *Biometrical Journal*, 50(3):375–385.