International Conference Mathematical and Computational Biology 2011
International Journal of Modern Physics: Conference Series
Vol. 9 (2012) 112–118
© World Scientific Publishing Company
DOI: 10.1142/S2010194512005168

SIMULATION OF INTERVAL CENSORED DATA IN MEDICAL AND BIOLOGICAL STUDIES

KAVEH KIANI

Institute for Mathematical Research, Universiti Putra Malaysia,
43400 UPM Serdang, Selangor, Malaysia
kaveh@inspem.upm.edu.my

JAYANTHI ARASAN

Institute for Mathematical Research, Universiti Putra Malaysia,
43400 UPM Serdang, Selangor, Malaysia,
Department of Mathematic, Faculty of Science, Universiti Putra Malaysia,
43400 UPM Serdang, Selangor, Malaysia
jayanthi@math.upm.edu.my

This research looks at the simulation of interval censored data when the survivor function of the survival time is known and attendance probability of the subjects for follow-ups can take any number between 0 to 1. Interval censored data often arise in the medical and biological follow-up studies where the event of interest occurs somewhere between two known times. Regardless of the methods used to analyze these types of data, simulation of interval censored data is an important and challenging step toward model building and prediction of survival time. The simulation itself is rather tedious and very computer intensive due to the interval monitoring of subjects at prescheduled times and subject's incomplete attendance to follow-ups. In this paper the simulated data by the proposed method were assessed using the bias, standard error and root mean square error (RMSE) of the parameter estimates where the survival time T is assumed to follow the Gompertz distribution function.

Keywords: Interval censored; follow-up study; survival analysis; Gompertz model.

1. Introduction

Interval censored data arise in studies when the event of interest occurs somewhere between two known times. Similar to most studies in survival analysis the main objective is still estimating actual survival times of subjects.

Interval censored data can be analyzed using parametric, semi-parametric and non-parametric models. In parametric models the survivor function of the survival time or $S(t)$ is known whereas $S(t)$ is known partially in semi-parametric models and completely unknown in non-parametric models. This study focuses on an approach of simulating interval censored data in the parametric setting.

There are some reviews and books which have been written in the area of analyzing interval censored data such as, Huang and Wellner,[1] Lindsey and Ryan,[2] Gomez et al.,[3] Lesaffre et al.[4] and Sun.[5] Odell et al.[6] presented a likelihood function that could contain interval, left and right censored data. Lindsey[7] studied interval censoring in nine parametric regression models and has compared the results of the two methods, using the exact likelihood and density estimation (by midpoint).

Although, there are many literatures about analyzing interval censored data, there are hardly any literatures on the methods of simulating interval censored data. Lawless[8] describe how to estimate inspection process from a real interval censored data set. After estimating the process he would be able to simulate interval censored data sets from the estimated process. Lee [9] has proposed a method which can be used for estimating $S(t)$ where the data are interval censored. Also in his work, $S(t)$ is unknown and has to be estimated by simulating interval censored data from a real data set.

But, sometimes inference about a real data set is not the motivation of that study. Motivation could be extending a new model for analyzing interval censored data and check the performance of this model by simulation.

In this paper it has been assumed $S(t)$ is known and method of simulating interval censored data for this known distribution is the objective of this study. In section 3 in order to assess the quality of the simulated interval censored data an example is presented by assuming $S(t)$ follows the Gompertz distribution. The properties of the Gompertz distribution is presented in Johnson et al.[10] Some authors have done studies on different characteristics of the Gompertz distribution, for instance, Makany[11] and Chen[12].

2. Simulation of Interval Censored Survival Time Data

Let T denote a nonnegative random variable representing survival time of a subject. T is a continuous variable with a known survivor function $S(t)$. An observation on T is interval censored when the exact value of T is unknown and only an interval (L, R) is observed where $T \in (L, \mathrm{R})$ and $L \leq R$ with probability 1.

In the simulation of interval censored data triplets (t_i, l_i, r_i) where $i = 1,2,\dots,n$ are produced. The t_i's can be easily generated via simulation. However, the same is not the case with the simulation of l_i's and r_i's.

In real life, l_i and r_i may only be certain predetermined points on the time axis or discrete follow-up times, because it is impossible to observe subjects continuously. In fact, this is the main reason for the existence of interval censored data in most long term studies. Specially in follow-up studies involving biological subjects several reasons may exists as to why a subject fail to attend or simply miss some of the follow-ups.

In order to simulate these phenomena, we define a set of potential inspection times and assume that the subjects are inspected or examined at these times. Here after the set of potential inspection times is shown by P where $P = \{p_1, p_2, \dots, p_k\}$ and subjects attendance probability to each of the p_j's is shown by q where $0 \leq q \leq 1$ and $\mathrm{j} = 1,2,\dots,k$.

Also, when

- $q = 1$, subjects attend all of the p_j's.
- $q = 0$, subjects miss all of the p_j's.
- $0 < q < 1$, subjects will attend to some of the p_j's and will miss others.

Depending on the value of q each subject will have a set of actual inspection times or A, where for the i^{th} subject, $A_i = \{a_{i1}, a_{i2}, \ldots, a_{im_i}\}$ and $A_i \subset P$.

We should bear in mind that right and left censored data can be regarded as special cases of interval censored data, Kalbfleisch and Prentice.[13] The i^{th} subject is right censored when $t_i \in (l_i, +\infty)$ or, the subject has been event free at the last actual inspection time. A subject is left censored when $t_i \in (0, r_i)$ or, the subject has met the event before the first actual inspection time.

Algorithm of the simulation interval censored data with different q's is described in next section.

2.1. *Algorithm of the simulation*

The following assumptions are made before moving on to the simulation algorithm.

- There are k potential inspection times which are known by design.
- All subjects are observed in the first inspection time or p_1.
- Subjects will attend follow-ups with probability q.
- Survival times are generated from a known $S(t)$.

In order to generate interval censored data for the 1^{st} subject or (l_1, r_1), the following algorithm is used:

(i) Generate $u_2 \sim Uniform(0,1)$.

(ii) Define an indicator function,

$$I = \begin{cases} 1, & \text{if subject attend } \mathrm{p_2} \ (\mathrm{u_2} \leq \mathrm{q}); \\ 0, & \text{if subject miss } \mathrm{p_2} \ (\mathrm{u_2} > q). \end{cases}$$

(iii) Repeat steps (i) and (ii) for p_j where $j = 3,4, \ldots, k$.

(iv) Create vector of attendance for all the k members of P. Consequently this vector will direct us to a set of actual inspection times or A_1. For example vector (1,0,0,1,0,1,0,0,1) shows $A_1 = (a_{11}, a_{12}, a_{13}, a_{14}) = (p_1, p_4, p_6, p_9)$.

(v) Select the largest member of A_1 which is less than t_1 as a l_1 and smallest member of A_1 which is more than t_1 as a r_1. It is clear that if,

$$\begin{cases} t_1 < a_{11}, & \text{observation is left censored } \Rightarrow (\mathrm{l_1}, \mathrm{r_1}) = (0, \mathrm{a_{11}}); \\ t_1 > a_{1m_1}, & \text{observation is right censored } \Rightarrow (\mathrm{l_1}, \mathrm{r_1}) = (\mathrm{a_{1m_1}}, +\infty); \\ t_1 = a_{1j}, & \text{observation is exact survival time } \Rightarrow \mathrm{l_1} = \mathrm{r_1} = \mathrm{a_{1j}}. \end{cases}$$

(vi) Repeat steps (i) to (v) to obtain intervals for the other subjects.

3. Assessing the Simulation Process: An Example

In this section, we would like to apply the simulation in section (2.1) also discuss when applied to the Gompertz survival time distribution. The hazard function of this distribution is,

$$h(t) = \lambda exp(\gamma t), \qquad t \geq 0, \ \ \lambda > 0, \ \ \gamma > 0,$$

where $\theta = (\lambda, \gamma)$ is the vector of parameters. T is the non-negative continuous random variable which denotes the subject's survival time. The scale parameter is λ and the shape parameter is γ. The survivor function of the model is,

$$S(t) = exp\left[\frac{\lambda}{\gamma}(1 - e^{\gamma t})\right],$$

and the probability density function is,

$$f(t) = \lambda exp(\gamma t) \times exp\left[\frac{\lambda}{\gamma}(1 - e^{\gamma t})\right].$$

The parameters of this model can be estimated by the method of maximum likelihood (MLE). If there are no censored observation, then the likelihood function for the full sample is,

$$L(\theta) = \prod_{i=1}^{n} f(t_i) = \prod_{i=1}^{n} \left\{h(t_i) \times exp\left[\frac{\lambda}{\gamma}\left(1 - e^{\gamma t_i}\right)\right]\right\}.$$

In order to incorporate interval and right censored data to the likelihood function we need to define an indicator functions. Here is assumed there is not any left censored or exact data. For the i^{th} observation,

$$\delta_i = \begin{cases} 1, & \text{observation is right censored;} \\ 0, & \text{observation is interval censored}, \end{cases}$$

then the likelihood function will be,

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} [S(l_i) - S(r_i)]^{(1-\delta_i)}[S(l_i)]^{\delta_i} \\ &= \prod_{i=1}^{n} \left\{exp\left[\frac{\lambda}{\gamma}(1 - e^{\gamma l_i})\right] - exp\left[\frac{\lambda}{\gamma}(1 - e^{\gamma r_i})\right]\right\}^{(1-\delta_i)} \\ &\times \left\{exp\left[\frac{\lambda}{\gamma}(1 - e^{\gamma l_i})\right]\right\}^{\delta_i}. \end{aligned}$$

The log-likelihood function will be,

$$l(\theta) = \sum_{i=1}^{n} (1-\delta_i) ln\left\{1 - exp\left[\frac{\lambda}{\gamma}(e^{\gamma l_i} - e^{\gamma r_i})\right]\right\} + \left[\frac{\lambda}{\gamma}(1 - e^{\gamma l_i})\right].$$

The MLEs of the parameters can be obtained by using any iterative procedure such as the Newton-Raphson algorithm.

A simulation study using 1000 samples each with n = 50, 100, 150 and 200 subjects conducted for this model. The values of 0.1 and 0.1 were chosen as the parameters of λ and γ. By generating one random number, u_i: $u(0,1)$, t_i's are generated by,

$$t_i = \frac{1}{\gamma} ln\left[1 - \frac{\gamma ln u_i}{\lambda}\right], \qquad i = 1,2,\dots,n.$$

Three study periods, $k = 12, 24$ and 36 months and five subject's attendance probabilities, $q = 1, 0.8, 0.6, 0.4$ and 0.2 are assumed. From table 1 we can see that the percent of interval censored data increase with the increase in k and q. Also, the mean of length of intervals seems to increase with the increase in k and decrease in q. Thus, data with wider intervals contains less information regarding the actual failure times t_i. Hence, the bias, standard error and root mean square error (RMSE) will all be higher. From table 2 we can clearly see that the bias, standard error and RMSE values decrease with the increase in k, q and n. Bias, standard error and RMSE values are small enough to conclude that simulated data are produced from a stable and well designed simulation process.

4. Conclusion

In this paper we proposed a method for simulation of interval censored survival time data when the survivor function or $S(t)$ is known. It was shown that the bias, standard error and RMSE values decrease with the increase in k, q and n. Thus, these values confirm that simulated data are generated by a well designed simulation algorithm.

The work in this research was limited in testing the algorithm by the Gompertz model without covariate. The algorithm should be investigated and tested further by using other parametric models also adding different types of covariates to the models, for example fixed and time-dependent covariates. Also, q was assumed fixed, more work should be done to extend the algorithm to support different q's or a random q for each subject.

Table 1. Comparison of interval censored data in different attendance probabilities and study periods, n = 100.

Attendance	Study period	Censoring type		Mean of intervals
probability (q)	(k month)	Right %	Interval %	length (month)
1	12	36.0	64.0	1.000
	24	10.5	89.5	1.000
	36	2.8	97.2	1.000
0.8	12	36.9	63.1	1.438
	24	10.8	89.2	1.460
	36	2.9	97.1	1.470
0.6	12	38.5	61.5	2.110
	24	11.4	88.6	2.208
	36	3.0	97.0	2.225
0.4	12	42.0	58.0	3.203
	24	12.6	87.4	3.599
	36	3.4	96.6	3.669
0.2	12	50.3	49.7	4.973
	24	13.6	82.4	6.834
	36	5.2	94.8	7.444

Table 2. Bias, standard error and RMSE of the parameter estimates.

Sample	Attendance	Study	Bias		Std.		RMSE	
size (n)	probability (q)	period (k)	λ	γ	λ	γ	λ	γ
50	0.6	12	0.0041	0.0050	0.2180	0.2236	0.2180	0.2237
		24	0.0028	0.0032	0.2135	0.2201	0.2135	0.2201
		36	0.0017	0.0021	0.2094	0.2123	0.2094	0.2123
	0.8	12	0.0038	0.0047	0.2041	0.2097	0.2041	0.2098
		24	0.0023	0.0029	0.2015	0.2065	0.2015	0.2065
		36	0.0014	0.0018	0.2000	0.2015	0.2000	0.2015
100	0.6	12	0.0039	0.0045	0.1982	0.2051	0.1982	0.2051
		24	0.0025	0.0029	0.1920	0.2012	0.1920	0.2012
		36	0.0017	0.0017	0.1881	0.1976	0.1881	0.1976
	0.8	12	0.0034	0.0041	0.1878	0.1995	0.1878	0.1995
		24	0.0021	0.0026	0.1734	0.1912	0.1734	0.1912
		36	0.0013	0.0015	0.1720	0.1902	0.1720	0.1902
150	0.6	12	0.0035	0.0039	0.1852	0.1854	0.1825	0.1854
		24	0.0024	0.0025	0.1851	0.1850	0.1851	0.1850
		36	0.0015	0.0016	0.1812	0.1832	0.1812	0.1832
	0.8	12	0.0034	0.0035	0.1742	0.1801	0.1742	0.1801
		24	0.0018	0.0023	0.1715	0.1773	0.1715	0.1773
		36	0.0011	0.0012	0.1705	0.1753	0.1705	0.1753
200	0.6	12	0.0033	0.0034	0.1694	0.1724	0.1694	0.1724
		24	0.0021	0.0022	0.1653	0.1687	0.1653	0.1687
		36	0.0012	0.0012	0.1620	0.1631	0.1620	0.1631
	0.8	12	0.0030	0.0033	0.1587	0.1609	0.1587	0.1609
		24	0.0015	0.0018	0.1526	0.1534	0.1526	0.1534
		36	0.0009	0.0010	0.1478	0.1483	0.1478	0.1483

References

1. J. Huangand, J. A. Wellner, Interval Censored survival Data: A Review of Recent Progress, in Proceedings of the First Scattle Symposium in Biostatistics: Survival Analysis. (Springer-Verlag, New York, 1997) p. 123.
2. J. C. Lindsey and L. M. Ryan, Statistics in Medicine. 17, 219 (1998).
3. G. Gomez, M. L. Calle and R. Oller, Statistical Papers. 45, 139 (2004).
4. E. Lesaffre, A. Komarek and D. Declerck, Statistical Methods in Medical Research. 14, 539 (2005).
5. J. Sun. The Statistical Analysis of Interval-Censored Failure Time Data. (Springer, New York, 2006).
6. P. M. Odell, K. M. Anderson and R. B. D'Agostino, Biometrics. 48, 951 (1992).
7. J. K. Lindsey, Lifetime Data Analysis. 4, 329 (1998).
8. J. F. Lawless and D. Babineau, Biometrika. 93-3, 671 (2006).
9. C. Lee, Statistics & Probability Letters. 45, 131 (1999).
10. N. L. Johnson, S. Kotz and N. Balakrishnan. Continuous Univariate Distributions, Volume 2. (Wiley Press, New York, 1995).
11. R. Makany, Biometrical. 33, 121 (1991).
12. Z. Chen, Biometrical. 39, 117 (1997).
13. J. D. Kalbfleisch and R. L. Prentice. The Statistical Analysis of Failure Time Data. (Wiley press, New York, 2002).