Jordan Journal of Mathematics and Statistics (JJMS) 9(3), 2016, pp 203 - 215

ON COMPARISON OF SURVIVAL CURVES WITH INTERVAL CENSORED DATA

REZA PAKYARI [1] AND DANIAL HABIBI [2]

Abstract. Survival comparison is one of the main goals and interesting problems in most survival studies such as clinical studies. In this paper, we compare through a Monte Carlo study, several tests for comparison of survival functions for interval censored failure time data. In particular, three nonparametric generalized log-rank tests, a parametric score test and an imputation-based test have been considered. It is observed that the parametric score test and the imputation test outperform the nonparametric generalized log-rank tests in most cases. Finally, a real dataset is studied for illustrative purposes.

1. Introduction

Interval censored data often arise in medical periodic follow-up studies where the survival time of interest is observed only to belong to an interval rather than being exactly known. For example, consider a patient who is monitoring weekly or monthly for a clinically observed change and is missing some visits, and returns with a changed in observed response status, thus producing an interval censored observation. Another example occurs in breast cancer studies, where early breast cancer patients were supposed to be seen at clinic for breast retraction every four to six months. Here, the event of interest is the time to breast retraction, however, no exact time is available.

1991 *Mathematics Subject Classification.* 62N01, 92B15.

Key words and phrases. Exponential distribution, Generalized log-rank test, Interval censoring, Monte Carlo simulation, Survival comparison.

Received: Nov. 27, 2015 Accepted: Aug. 3, 2016 .

In this case, the interval censored observation consists the interval from the time of the last retraction negative test and the time of the first retraction positive test (Finkelstein and Wolfe [1]).

One of the main problems in survival studies is the comparison of survival functions between different treatment groups. The well known log-rank test is used when the available data are right censored (see e.g. Fleming and Harrington [2] and Kalbfleisch and Prentice [3]). However, with interval censored data there are several generalizations of the log-rank tests provided by the authors. Peto and Peto [4] studied the problem for comparing two groups of independent interval censored observations. Their method allow the data to be exact or interval censored. Finkelstein [5] used a proportional hazards regression model to develop a parametric score test for this problem.

Zhao and Sun [6] improved the nonparametric test studied by Sun [7] and proposed a generalized log-rank test that allows data to have both interval censored and exactly observed observations.

Sun and Zhao [8] proposed a generalized log-rank test when all the data are of the form of interval censored and there is no exact observations. Zhao et. al. [9] developed the test procedure of Sun and Zhao [8] to allow data with exact observations as well as the interval censored. Whence, their method reduces to the Sun and Zhao [8] when there is no exact observations in the interval censored data. Recently Zhao et. al. [10] proposed a new class of generalized log-rank tests. They also provided the asymptotic distribution of the test statistics under both null and alternative distributions. For an excellent overview on interval censoring, one may refer to Sun [11], Sun [12] and Chen et. al. [13].

Another approach to handle interval censored data is to use an imputation-based inference. Imputation for interval censored data means to consider one single data representing the whole interval data. One common choice is to consider the middle

point of the data, however one may consider the left or the right end points of the interval. The resulting data will be a set of complete data and then the usual log-rank test can be employed to carry out the survival comparison. The big merit of using the imputation method is its simplicity and if all the intervals are relatively narrow the resulting inference will be reasonably well.

The rest of this article is organized as follows. In Section 2, we review some of the existing parametric and nonparametric generalized log-rank tests in the literature for interval censored data. Section 3 is devoted to a simulation study of the tests studied in Section 2. The empirical powers of the test are evaluated by means of a Monte Carlo simulation and thereby the best tests are identified. Finally, in Section 4, we study a real dataset.

2. Tests for comparing of several survival functions

Consider a survival study consists of a total of n independent subjects from k different treatment groups. Also, suppose that there are n_l subjects from group l, for $l = 1, \ldots, k$, such that $\sum_{l=1}^{k} n_l = n$. Let T_i denotes the survival time of interest for subject i, $i = 1, \ldots, n$, and that only interval censored data are available of the form

$$\{(L_i, R_i], z_i; i = 1, \ldots, n\}$$

where $(L_i, R_i]$ denotes the interval to which T_i belongs and z_i represents the k vector of treatment indicators for the ith subject. Note that, if $L_i = R_i$, then T_i is exactly observed, whereas $L_i = 0$ and $R_i = \infty$ indicates left and right censored, respectively. Let $S_l(t)$ denotes the survival function corresponds to the l-th treatment group. We are interested to test the assumption that the k treatment groups have identical survival functions, i.e.

$$H_0 : S_1(t) = \ldots = S_k(t)\,. \tag{2.1}$$

Several parametric and nonparametric tests are available in the literature. Zhao and Sun [6] proposed a nonparametric generalized log-rank test statistic given by

$$T_{\text{glrt1}} = U_I^T \, V_I^- \, U_I \,, \tag{2.2}$$

where $U_I = \sum_{j=1}^{m}(d_{jl} - n_{jl}d_j/n_j)$ and d_{jl}'s and n_{jl}'s are the natural estimates of the numbers of failures and risks. The V^- is the generalized inverse of the estimate of the covariance matrix of U_I. The test statistic T_{glrt1} under the null hypothesis follows approximately a χ^2 distribution with $(k-1)$ degrees of freedom. See Zhao and Sun [6] for details.

Sun and Zhao [8] proposed another generalized log-rank test procedure based on the link function ζ and the statistic

$$U_{II} = \sum_{i=1}^{n} z_i \, \frac{\zeta\{\hat{S}(L_i)\} - \zeta\{\hat{S}(R_i)\}}{\hat{S}(L_i) - \hat{S}(R_i)} \,.$$

Several link functions may be used, however, a general link function in the form $\zeta(x) = (x \log z)x^{\rho}(1-x)^{\gamma}$ has the advantages that the users can select their own constants ρ and γ based on the application. They showed that the test statistic

$$T_{\text{glrt2}} = U_{II,0}^T \, V_{II,0}^{-1} \, U_{II,0}/n \,, \tag{2.3}$$

has asymptotically a χ^2 distribution with $(k-1)$ degree of freedom, where $U_{II,0}$ denotes the first $(k-1)$ components of U_{II} and $V_{II,0}$ is the matrix after deleting the last row and column of the estimated covariance matrix of $U_{II}/\sqrt{n}$.

Zhao et. al. [9] proposed another generalized log-rank testing procedure for interval censored data. The test statistic is given by

$$\begin{aligned} U_{III} =& A \sum_{i=1}^{n} z_i \, e_i \, \frac{\zeta\{\hat{S}(R_i-)\} - \zeta\{\hat{S}(R_i)\}}{\hat{S}(R_i-) - \hat{S}(R_i)} \\ &+ B \sum_{i=1}^{n} z_i \, (1-e_i) \, \frac{\zeta\{\hat{S}(L_i)\} - \zeta\{\hat{S}(R_i)\}}{\hat{S}(L_i) - \hat{S}(R_i)} \,, \end{aligned}$$

where e_i is the indicator for the exactly observed observation, and A and B are the diagonal coefficient matrices. They showed that the test statistic

$$T_{\text{glrt3}} = U_{III,0}^T \, V_{III,0}^{-1} \, U_{III,0}/n \,, \tag{2.4}$$

has asymptotically a χ^2 distribution with $(k-1)$ degrees of freedom. See Zhao et. al. [9] for details.

Note that U_I and U_{III} support data to be both interval censored and exactly observed observations, whilst U_{II} is devoted to just interval censored data. Moreover, U_{III} reduces to U_{II} when there is only interval censored data and no exact observed observations.

A parametric method known as score test was proposed by Finkelstein [5] using the proportional hazards regression model. The test statistic is given by

$$U_{IV} = \sum_{i=1}^{n} \sum_{j=1}^{m+1} \{ \frac{z_i \log \hat{p}_j \sum_{r=j}^{m+1} \alpha_{ir} \hat{g}_r}{\sum_l \alpha_{il} \hat{g}_l} - z_i \frac{\log \hat{p}_j}{1-\hat{p}_j} \frac{\alpha_{ij} \hat{g}_j}{\sum_l \alpha_{il} \hat{g}_l} \} \,,$$

where $\hat{p}_j = \hat{S}(s_j)/\hat{S}(s_j - 1)$ and $\hat{g}_j = \hat{S}(s_{j-1})/\hat{S}(s_j)$. She showed that under the null hypothesis of no differences between the survival curves test statistic

$$T_{\text{score}} = U_{IV}^T \, V_{IV}^{-1} \, U_{IV} \,, \tag{2.5}$$

has asymptotically a χ^2 distribution with $(k-1)$ degrees of freedom. Note that when there is no exact observed observations, U_{IV} reduces to U_{II} with link function $\zeta(x) = x \log x$.

A simple procedure to handle the interval censored data is to transform the interval censored data to a set of complete data by means of an imputation approach. One common choice is to let $T_i = \frac{L_i + R_i}{2}$, that is the mid-point imputation. Alternatively, one may consider $T_i = L_i$, or $T_i = R_i$ corresponding to the left end and right end imputation methods, respectively. Upon imputation the interval censored data, the existing methods for complete data may be employed to test the null hypothesis.

In the next section, we compare the power of the above tests through a Monte Carlo study.

3. Monte Carlo Power Study

In this section, we assess the the power of the tests discussed in the last section by means of Monte Carlo simulations. In particular, we compare the power of the Finkelstein [5], Zhao and Sun [6] and Sun and Zhao [8] test procedures as well as the imputation method. The empirical significance level is also calculated to check the validity of the test procedures. All the simulations were carried out in R using the pseudo-random generator in that software package.

We used the algorithm proposed by Kiani and Arasan [14] to generate interval censored data. A two-sample comparison with $n = 50$ and $n = 100$ subjects in each population was considered. The interval censored survival times $(L_i, R_i]$ for $i = 1, \ldots, n$ were generated from the exponential distributions with means (hazards) $\exp(\alpha)$ and $\exp(\alpha + \beta)$ corresponding to the first and second population, respectively. The exponential parameters were set to be $\alpha = 2.0$ and $\beta = -0.8, -0.4, 0.0, 0.2, 0.4$ and 0.8. Note that β represents the difference between the two populations and $\beta = 0$ will assess the empirical size of the tests. This is the underlying distribution considered by Zhao and Sun [6], Sun and Zhao [8] and Zhao et. al. [9].

The significance level of all tests was set at 0.05 and 5000 replications were considered in the Monte Carlo simulations.

We used the glrt R package to perform the nonparametric generalized log-rank tests and the parametric score test. In the tables we have denoted the generalized log-rank tests proposed by Zhao and Sun [6] and Sun and Zhao [8] by "glrt1" and "glrt2", respectively, and the score test proposed by Finkelstein [5] by "score".

Tables 1 and 2 present empirical rejection probabilities for testing the equality of the two exponential survival models, $\exp(\alpha)$ and $\exp(\alpha+\beta)$ for various subject attendance

TABLE 1. Estimated power and size when $n = 50$ and $\alpha = 0.05$ for various subjects attendance probabilities.

β	Test	ρ	γ	Subjects attendance probability (q) 0.5	0.75	0.9
-0.8	glrt1	-	-	0.9636	0.9676	0.9728
	glrt2	0	0	0.9666	0.9660	0.9708
			1	0.9322	0.9298	0.9266
		1	0	0.6792	0.6700	0.6802
			1	0.8478	0.8376	0.8436
	Score	-	-	0.9738	0.9720	0.9760
	Imputation	-	-	0.9700	0.9732	0.9746
-0.4	glrt1	-	-	0.4810	0.5078	0.5112
	glrt2	0	0	0.4867	0.4968	0.4982
			1	0.4123	0.4150	0.4132
		1	0	0.2228	0.2234	0.2290
			1	0.3181	0.3182	0.3180
	Score	-	-	0.5166	0.5270	0.5278
	Imputation	-	-	0.4804	0.4876	0.4938
0.0	glrt1	-	-	0.0542	0.0558	0.0556
	glrt2	0	0	0.0562	0.0534	0.0518
			1	0.0526	0.0526	0.0510
		1	0	0.0468	0.0486	0.0474
			1	0.0570	0.0518	0.0518
	Score	-	-	0.0628	0.0636	0.0634
	Imputation	-	-	0.0506	0.0520	0.0516
0.4	glrt1	-	-	0.4756	0.4864	0.4860
	glrt2	0	0	0.4696	0.4762	0.4730
			1	0.3932	0.3986	0.3866
		1	0	0.2192	0.2142	0.2136
			1	0.3062	0.2970	0.2934
	Score	-	-	0.5044	0.5048	0.5026
	Imputation	-	-	0.5070	0.5086	0.5096
0.8	glrt1	-	-	0.9734	0.9742	0.9756
	glrt2	0	0	0.9722	0.9728	0.9732
			1	0.9320	0.9314	0.9284
		1	0	0.6716	0.6634	0.6640
			1	0.8388	0.8310	0.8308
	Score	-	-	0.9784	0.9784	0.9784
	Imputation	-	-	0.9718	0.9714	0.9757

TABLE 2. Estimated power and size when $n = 100$ and $\alpha = 0.05$ for various subjects attendance probabilities.

β	Test	ρ	γ	Subjects attendance probability (q) 0.5	0.75	0.9
-0.8	glrt1	-	-	0.9998	0.9996	1.0000
	glrt2	0	0	0.9998	0.9996	1.0000
			1	0.9984	0.9980	0.9988
		1	0	0.9196	0.9230	0.9258
			1	0.9858	0.9868	0.9880
	Score	-	-	0.9998	0.9996	1.0000
	Imputation	-	-	1.0000	0.9996	1.0000
-0.4	glrt1	-	-	0.7702	0.7894	0.7934
	glrt2	0	0	0.7802	0.7854	0.7872
			1	0.7054	0.7056	0.7026
		1	0	0.3820	0.3740	0.3744
			1	0.5300	0.5304	0.5282
	Score	-	-	0.7904	0.7976	0.7890
	Imputation	-	-	0.7852	0.7904	0.7962
0.0	glrt1	-	-	0.0508	0.0506	0.0514
	glrt2	0	0	0.0518	0.0500	0.0502
			1	0.0510	0.0508	0.0492
		1	0	0.0526	0.0546	0.0526
			1	0.0494	0.0516	0.0516
	Score	-	-	0.0562	0.0552	0.0534
	Imputation	-	-	0.0536	0.0546	0.0536
0.4	glrt1	-	-	0.7974	0.8022	0.8026
	glrt2	0	0	0.7964	0.7982	0.7988
			1	0.7046	0.7084	0.7094
		1	0	0.3996	0.3902	0.3900
			1	0.5398	0.5342	0.5280
	Score	-	-	0.8086	0.8112	0.8116
	Imputation	-	-	0.7928	0.8006	0.8018
0.8	glrt1	-	-	1.0000	1.0000	1.0000
	glrt2	0	0	1.0000	1.0000	1.0000
			1	0.9996	0.9990	0.9992
		1	0	0.9290	0.9252	0.9268
			1	0.9840	0.9822	0.9830
	Score	-	-	1.0000	1.0000	1.0000
	Imputation	-	-	0.9998	1.0000	1.0000

probabilities, q when $n = 50$ and $n = 100$, respectively. We considered $q = 0.5, 0.75$ and 0.9. For example, $q = 0.5$ means that a subject has a $50 - 50$ chance to attend the next visit, whilst $q = 0.9$ shows a high probability for the subject to attend the next visit. Two values zero and one were considered for the link function parameters ρ and γ. Note that since all of the generated data are in the form of interval censored and there is no exactly observed observations, the "glrt3" method is equivalent to the "glrt2" and hence is omitted from the study. We found that the three imputation methods, left end, middle and right end points produce very close powers, and the mid-point method shows a little better performance in compare to the other two methods. Hence, we only give the results of the mid-point imputation approach. Obviously, for large values of q, the subject attendance probability, the difference in power for the three imputation approaches is negligible since in this case the finite intervals are narrow.

Note that when $\beta = 0$, the two models are the same and the rejection probabilities give the empirical significance level of the tests. It is observed that all tests maintain the level of significance at the nominal level. Moreover, as one would expect all the estimated powers for $n = 100$ in Table 2 are greater than their corresponding powers for $n = 50$ in Table 1. Also, the empirical powers in both tables, increases as the value of the subject attendance probability increases. This is due to narrower intervals when the value of q is high and hence more information will be available of the true value of the survival times T_i within each interval.

The values in Tables 1 and 2 reveal that the Finkelstein score test outperform the other tests almost in all cases. However, with larger value of the sample size, say $n = 100$, "glrt1" test due to Zhao and Sun [6] performs as good as the score test specifically when the distance between the two models becomes larger. In all cases, "glrt2" test with the link function parameters $(\rho, \gamma) = (1, 0)$ give the worst result in compare to the other tests. Surprisingly, the simple mid-point imputation method

performs very well and give the empirical powers that are very closed to the score test specifically with $n = 100$.

4. Illustrative Example

In this section, we study a real dataset to illustrate the generalized log-rank tests discussed in Section 2. The data was first considered by Pakyari and Abolhassani [15] and consist of a group of 207 children that were examined periodically to determine the time of their first permanent moral tooth decay in Urmia at the north-east of Iran. Using the Palmer notation they have studied the first permanent molar tooth number 6 in each quadrant. This tooth is usually emerges at the age of 6 to 7. The age of the patients approximately ranged from 3 to 14 years of which 46.8% were male and 53.2% were female.

Comparison of the time of the first permanent molar tooth decay between boys and girls are of interest. Table 3 gives the p-values of the tests discussed in Section 2. It is observed that all the tests support the null hypothesis of no difference between the time of the first permanent moral tooth decay in boys and girls.

Figure 1 shows the nonparametric maximum likelihood estimate of the survival functions for each treatment group of the tooth data. This figure also supports the equality of survival function of boys and girls.

Concluding Remarks and Future works

As one of the anonymous reviewers suggests it would be nice to do a simulation study when the observation process is a time sequence with certain probability of missing. In this way one may mimic the clinical trial setting.

TABLE 3. p-values of the equality of survival curves tests for the first moral tooth data.

Test	ρ	γ	p-value
glrt1	-	-	0.5457
glrt2	0	0	0.4805
		1	0.5543
	1	0	0.5587
		1	0.4804
Score	-	-	0.4676
Imputation	-	-	0.3422

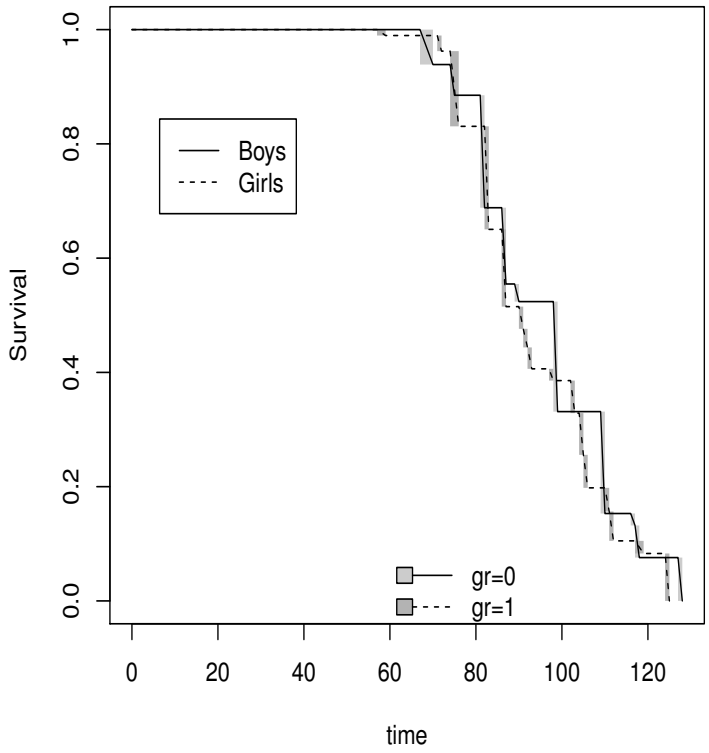

FIGURE 1. NPLME survival curve estimates from the first moral tooth data.

Acknowledgement

We would like to thank the editor and anonymous referees for making some valuable comments on an earlier version of this manuscript.

References

[1] D.M. Finkelstein, R.A. Wolfe. A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics* **41**(1985), 933-945.

[2] T.R. Fleming, D.P. Harrington. *Counting Process and Survival Analysis*. Wiley: New York, 1991.

[3] J.D. Kalbfleisch, R.L. Prentice. *The Statistical Analysis of Failure Time Data, 2nd Edition*. Wiley: New York, 2002.

[4] R. Peto, J. Peto. Asymptotically efficient rank invariant test procedures. *Journal of Royal Statistical Society A.* **135**(1972), 185-207.

[5] D.M. Finkelstein. A proportional hazard model for interval-censored failure time data. *Biometrics* **42**(1986), 845-854.

[6] Q. Zhao, J. Sun. Generalized log-rank test for mixed interval-censored failure time data. *Statistics in Medicine* **23**(2004), 1621-1629.

[7] J. Sun. A nonparametric test for interval-censored failure time data with application to AIDS studies. *Statistics in Medicine* **15**(1996), 1387-1395.

[8] J. Sun, Q. Zhao. Generalized log-rank test for interval-censored failure time data. *Scandinavian Journal of Statistics* **32**(2005), 49-57.

[9] X. Zhao, Q. Zhao, J. Sun, J.S. Kim, Generalized log-rank tests for partly interval-censored failure time data. *Biometrical Journal* **3**(2008), 375-385.

[10] X. Zhao, R. Duan, Q. Zhao, J. Sun. A new class of generalized log-rank tests for interval-censored failure time data. *Computational Statistics and Data Analysis* **60**(2013), 123-131.

[11] J. Sun. *Interval Censoring. Encyclopedia of Biostatistics*, 2090-2095. Wiley, New York, 1998.

[12] J. Sun. *Statistical Analysis of Interval-Censored Failure Time Data*. Springer, New York, 2006.

[13] D.G. Chen, J. Sun, K. Peace. *Interval-Censored Time-to-Event Data*. CRC Press, 2013.

[14] K. Kiani, J. Arasan. Simulation of interval censored data in medical and biological studies. *International Conference Mathematical and Computational Biology* **9**(2012), 112-118.

[15] R. Pakyari, S. Abolhassani. A new goodness of fit test for interval censored data. *Submitted for Publication* 2015.

(1) Department of Mathematics, Faculty of Science, Arak University, Arak 38156-8-8349, Iran.

E-mail address: r-pakyari@araku.ac.ir

(2) Department of Epidemiology, Arak University of Medical Sciences, Arak, Iran.

E-mail address: dhabibi67@gmail.com