Paper 257-2010

Analyzing Interval-Censored Survival Data with SAS® Software

Ying So and Gordon Johnston, SAS Institute Inc., Cary, NC
Se Hee Kim, University of North Carolina, Chapel Hill, NC

ABSTRACT

Survival data analysis is traditionally focused on analyzing lifetimes by using time that is measured to an event of interest, or the latest time available if the event did not occur during the observation period. Data measured in this way are called right-censored data. Many methods (nonparametric, semiparametric, and fully parametric) have been developed over the years to deal with this type of data. But what methods are available if the event time is not directly observed and the event is known only to have occurred within some interval of time? Data measured in this way are called interval-censored survival data, and the use of SAS® software to analyze this type of data is the focus of this paper.

INTRODUCTION

Interval-censored data are often found in longitudinal studies in which subjects are assessed only periodically for the response of interest. The time when the event of interest occurs is not directly observed but is known to take place within some time interval. For example, in a clinical trial subjects might visit a clinic for assessment at predetermined times. The onset of a condition of interest is known only to have occurred at some time between visits; the exact time of onset is not known. The times of occurrence of these events are said to be interval-censored. For example, the breast cancer data presented in Finkelstein and Wolfe (1985) consist of 94 breast cancer patients who were given radiation therapy (RT) or radiation therapy plus chemotherapy (RCT). Patients were supposed to be seen at clinic visits every four to six months. However, actual visit times vary from patient to patient, and times between visits also vary. At clinic visits, physicians evaluated the cosmetic appearance of patients, such as breast retraction. The data for the interval-censored event time of breast retraction are reproduced in the following table.

Therapy	Event Intervals								
RT	(45,]	(25, 37]	(37,]	(6, 10]	(46,]	(0, 5]	(0, 7]	(26, 40]	(18,]
	(46,]	(46,]	(24,]	(46,]	(27, 34]	(36,]	(7, 16]	(36, 44]	(5, 11]
	(17,]	(46,]	(19, 35]	(7, 14]	(36, 48]	(17, 25]	(37, 44]	(37,]	(24,]
	(0, 8]	(40,]	(32,]	(4, 11]	(17, 25]	(33,]	(15,]	(46,]	(19, 26]
	(11, 15]	(11, 18]	(37,]	(22,]	(38,]	(34,]	(46,]	(5, 12]	(36,]
	(46,]								
RCT	(8, 12]	(0, 5]	(30, 34]	(0, 22]	(5, 8]	(13,]	(24, 31]	(12, 20]	(10, 17]
	(17, 27]	(11,]	(8, 21]	(17, 23]	(33, 40]	(4, 9]	(24, 30]	(31,]	(11,]
	(16, 24]	(13, 39]	(14, 19]	(13,]	(19, 32]	(4, 8]	(11, 13]	(34,]	(34,]
	(16, 20]	(13,]	(30, 36]	(18, 25]	(16, 24]	(18, 24]	(17, 26]	(35,]	(16, 60]
	(32,]	(15, 22]	(35, 39]	(23,]	(11, 17]	(21,]	(44, 48]	(22, 32]	(11, 20]
	(14, 17]	(10, 35]	(48,]						

There are 38 patients who did not experience breast retraction. These right-censored data are represented by intervals without the right endpoint. The SAS data set BreastCancer that contains these data is created using the SAS statements in the section “APPENDIX: BREAST CANCER DATA” and is used to illustrate the methods described in this paper. The section “APPENDIX: SAS MACROS FOR INTERVAL-CENSORED DATA” describes SAS macros that are used in the examples in this paper.

If the interval-censored time for each subject is a member of a collection of nonoverlapping intervals, the interval-censored data become grouped failure-time data. A multinomial distribution can be used on the number of subjects in the given intervals (Lawless 2003). Prentice and Gloeckler (1978) derived the likelihood for the grouped-data proportional hazards model. This paper focuses on interval-censored data that are not grouped failure-time data.

Interval-censored event time data arise in many area of studies. Lately there is a growing interest in progression free survival (PFS) in studies of diseases that are slow growing and difficult to cure, such as low-grade lymphomas. The PFS time is usually defined as the time from randomization to either disease progression or death. Patients are assessed periodically for a possible change of disease status.

STATISTICAL ISSUES

In general, the analysis of failure-time data addresses three issues:

- estimation of the survival functions
- comparison of survival functions
- assessment of the effects of covariates on survival

For right-censored data, standard nonparametric and semiparametric methodologies include the Kaplan-Meier estimates of the survival function, the log-rank test for comparing survival functions, and Cox regression analysis for assessing covariates. Parametric methods are often used too, especially in the study of product reliability.

Parametric methods are also available for interval-censored data. The LIFEREG and RELIABILITY procedures fit popular lifetime distributions, such as the Weibull and lognormal, to interval-censored data by maximum likelihood estimation of distribution parameters. However, this paper concentrates on nonparametric methods.

Common methods for dealing with interval-censored data are midpoint imputation and right imputation. Midpoint imputation assigns the midpoint of the censoring interval as the failure time. Right imputation assigns the time when the event of interest is first noticed as the failure time. Right-censored data methodologies are then applied to the imputed data.

In the last two decades, many new methods for analysis of interval-censored failure time data have been proposed. These methods are more complex and harder to apply than their right-censored counterparts. However, many have demonstrated that the conventional imputation approach is biased and less efficient than new methodologies, especially for infrequent or imbalanced assessment. For example, the simulation results of Chen (2009) favor the proportional hazards regression model of Finkelstein (1986) over imputation-based analysis for interval-censored data.

ANALYSIS OF INTERVAL-CENSORED DATA

Regardless of the actual failure time, the failure time T of a subject is only observed to lie in the interval $(L, R]$—that is, after the last assessment with a negative identification of the event and at or before the first positive assessment. If there is no positive assessment at the end of the observation period, the failure time T is considered to be right-censored at the latest assessment time L and $R = \infty$. For data that have only one assessment at t^* per subject, then $(L = 0, R = t^*]$ for a positive identification of event and $(L = t^*, R = \infty]$ for a negative identification. Such data are also known as current status data. Methods for this type of data are much better developed and simpler than methods for the general case of interval censoring. For example, Sun (2006) describes methods for current status data.

Parametric analysis of interval-censored data can be carried out using the LIFEREG procedure in SAS/STAT software and the RELIABILITY procedure in SAS/QC software. These procedures also provide the NPMLE, which is computed by using the EM algorithm of Turnbull (1976) with the method of Gentleman and Geyer (1994) to ensure the global maximum. Variance estimates are computed by inverting the negative of the Hessian matrix at the NPMLEs.

SAS macros for analyzing interval-censored data are shown in the section "APPENDIX: SAS MACROS FOR INTERVAL-CENSORED DATA" and are available at `http://support.sas.com/kb/24980`. The macros will continue to be improved and have added features in the future.

Nonparametric Maximum Likelihood Estimator

Consider a sample of n subjects from a homogeneous population with survival function $S(t)$. Let T_i denote the survival time of interest for subject $i, 1 \le i \le n$, and let $(L_i, R_i]$ be the interval for which T_i is observed. The likelihood function for the set of observed intervals $\{(L_i, R_i], \ \ i = 1, \ldots, n\}$ is

$$\mathcal{L} = \prod_{i}^{n} \{S(L_i) - S(R_i)\}$$

From the data $\{(L_i, R_i], \ \ i = 1, \ldots, n\}$, a set of nonoverlapping intervals $\{(q_1, p_1], \ldots, (q_m, p_m]\}$ is generated over which the survival curve $S(t) = \Pr(T_i > t)$ is estimated. The nonparametric maximum likelihood estimator (NPMLE) of the survival function can decrease only on the smaller number of nonoverlapping intervals $(q_1, p_1], \ldots, (q_m, p_m]$, so the jump probabilities need be estimated only on these intervals (Peto 1973; Turnbull 1976).

The survival curve decreases in some or all of these intervals and is assumed to be constant everywhere except these intervals. Assuming the censoring mechanism is independent of the response time distribution and that each subject eventually fails, the likelihood of the data $\{T_i \in (L_i, R_i],\ \ i = 1, \ldots, n\}$ can be constructed from the pseudoparameters $\{\theta_j = \Pr(q_j < T \le p_j),\ \ j = 1, \ldots, m\}$. The vector parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ can be estimated by maximizing, with respect to $\theta_1, \ldots, \theta_m$, the likelihood $\mathcal{L}(\boldsymbol{\theta})$ under the constraint $\sum_{j=1}^{m} \theta_j = 1$,

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \sum_{j=1}^{m} z_{ij}\theta_j$$

where z_{ij} is 1 if $(q_j, p_j]$ is contained in $(L_i, R_i]$ and 0 otherwise. The maximum likelihood estimates $\{\hat{\theta}_1, \ldots, \hat{\theta}_m\}$ then yield the NPMLE of the survival function:

$$\hat{S}(t) = \begin{cases} 1 & t < q_1 \\ \sum_{k=j+1}^{m} \hat{\theta}_k & p_j \le t \le q_{j+1} \\ 0 & t \ge p_m \end{cases}$$

Peto (1973) suggested using a constrained Newton-Raphson search to locate the maximum of the log likelihood, but the optimization might not be feasible when the number of pseudoparameters is large. Also, the Newton-Raphson method does not guarantee a global maximum. Turnbull (1976) proved that the maximization of the likelihood function is equivalent to the solution of the following self-consistency equation and can be solved using the expectation-maximization (EM) algorithm (Dempster, Laird, and Rubin 1977):

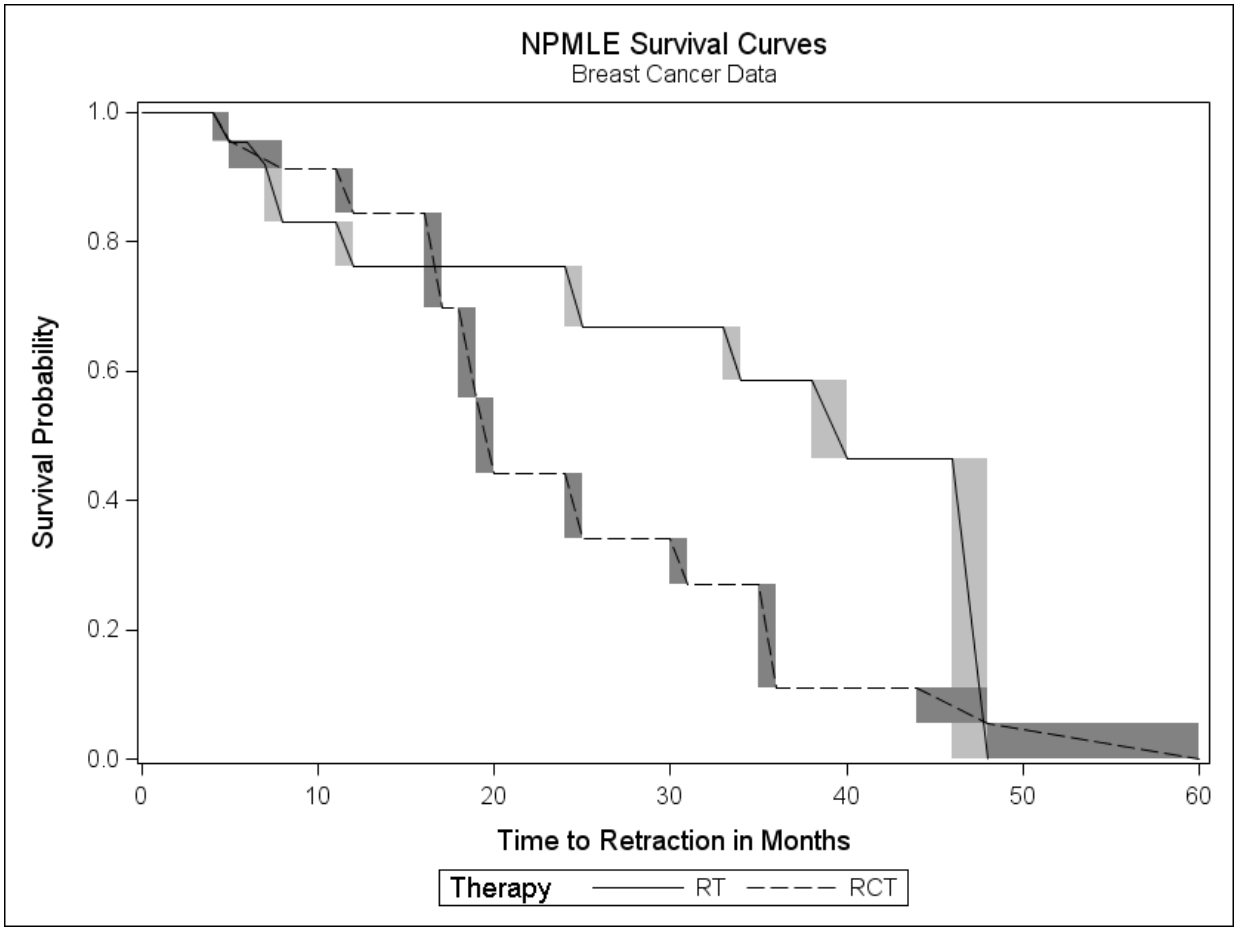
$$\theta_j = \frac{1}{n} \sum_{i=1}^{n} \frac{z_{ij}\theta_j}{\sum_{k=1}^{m} z_{ik}\theta_k}$$

Gentleman and Geyer (1994) introduced a method to ensure that the solution is a global maximum. Even with a moderate number of parameters, the EM algorithm is very slow. The iterative convex minorant (ICM) algorithm of Groeneboom and Wellner (1992) and the EM iterative convex minorant algorithm (EM-ICM) of Wellner and Zhan (1997) are much more efficient methods of computing the NPMLE than the EM algorithm. The latter algorithm converges to the NPMLE if it exists and is unique.

The following call to the macro %EMICM uses the EM-ICM algorithm to compute the NPMLE of the survival functions. The macro uses the BreastCancer data set described in the section “APPENDIX: BREAST CANCER DATA” and creates Figure 1.

```
************************************************
* NPMLE survival curves using the EM-ICM alorithm *
************************************************;
%EMICM(data=BreastCancer,
       left=lTime,
       right=rTime,
       group=Therapy,
       options=plot,
       title="NPMLE Survival Curves",
       title2="Breast Cancer Data",
       timelabel="Time to Retraction in Months"
      );
```

Figure 1 NPMLE of Survival Function for the RT and RCT Groups



The %EMICM macro also creates a tabular listing of the NPMLE of the survival functions for the RT and RCT therapies, as shown in Figure 2.

Figure 2 Listing of the NPMLE of the Survival Functions for the RT and RCT Groups

Therapy	Lower	Upper	Probability	Cumulative_ Probability	Survival_ Probability	Var_ Survival
RT	4	5	0.04635	0.04635	0.95365	.001224914
RT	6	7	0.03336	0.07971	0.92029	.002074818
RT	7	8	0.08867	0.16838	0.83162	.003369853
RT	11	12	0.07075	0.23913	0.76087	.003955371
RT	15	16	0.00000	0.23913	0.76087	.003955371
RT	17	18	0.00000	0.23913	0.76087	.003955371
RT	24	25	0.09265	0.33178	0.66822	.004985073
RT	25	26	0.00000	0.33178	0.66822	.004985073
RT	33	34	0.08179	0.41356	0.58644	.005459871
RT	34	35	0.00000	0.41356	0.58644	.005459871
RT	36	37	0.00000	0.41356	0.58644	.005459871
RT	38	40	0.12088	0.53444	0.46556	.005787789
RT	40	44	0.00000	0.53444	0.46556	.005787789
RT	46	48	0.46556	1.00000	0.00000	0
RCT	4	5	0.04328	0.04328	0.95672	.001163146
RCT	5	8	0.04328	0.08657	0.91343	.001694096
RCT	8	9	0.00000	0.08657	0.91343	.001694096
RCT	11	12	0.06921	0.15577	0.84423	.003229329
RCT	12	13	0.00000	0.15577	0.84423	.003229329
RCT	16	17	0.14540	0.30117	0.69883	.005688908
RCT	18	19	0.14109	0.44226	0.55774	.006726530
RCT	19	20	0.11575	0.55801	0.44199	.006002243
RCT	21	22	0.00000	0.55801	0.44199	.006002243
RCT	22	23	0.00000	0.55801	0.44199	.006002243
RCT	23	24	0.00000	0.55801	0.44199	.006002243
RCT	24	25	0.09987	0.65787	0.34213	.005339539
RCT	30	31	0.07088	0.72876	0.27124	.004596136
RCT	31	32	0.00000	0.72876	0.27124	.004596136
RCT	33	34	0.00000	0.72876	0.27124	.004596136
RCT	34	35	0.00000	0.72876	0.27124	.004596136
RCT	35	36	0.16083	0.88959	0.11041	.002305587
RCT	44	48	0.05521	0.94479	0.05521	.001257847
RCT	48	60	0.05521	1.00000	0.00000	0

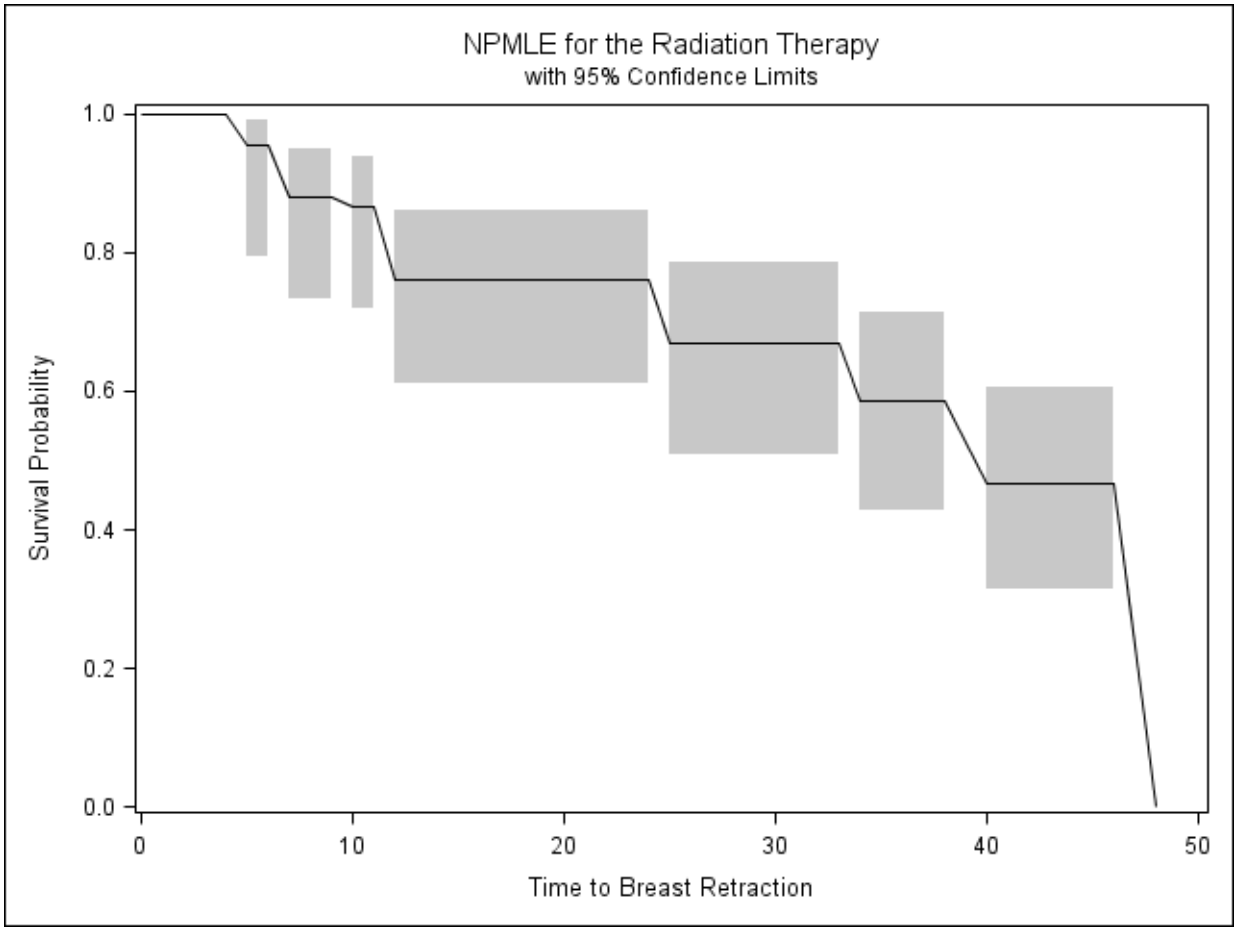
Peto (1973) and Turnbull (1976) both suggest inverting the Hessian matrix (the matrix of second derivatives of the negative log-likelihood) evaluated at the maximum likelihood estimates to yield the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$. However, this method is not reliable because the asymptotics might fall apart for a large number of parameters. You can always use a simple bootstrap method (Efron and Tibshirani 1993) for the variance estimation. For each replication k $(1 \le k \le M)$, a sample of size n is drawn with replacement from the data $\{(L_i, R_i], 1 \le i \le n\}$, and the NPMLE $\hat{S}_k(t)$ is calculated. The bootstrap variance of $\hat{S}(t)$ is the sample variance of $\{\hat{S}_1(t), \ldots, \hat{S}_M(t)\}$.

Sun (2001) suggests a generalization of the Greenwood formula for variance estimation based on resampling. For each replication $k(1 \le k \le M)$, an independent right-censored sample $\{(T_i^k, \delta_i^k), i = 1, \ldots, n\}$ of size n is obtained from the data $\{(L_i, R_i], 1 \le i \le n\}$ and the NPMLE $\hat{S}(t)$ as follows: $\delta_i^k = 0$ and $T_i^k = L_i$ if $(L_i, Ri]$ represents a right-censored observation; otherwise, $\delta_i^k = 1$ and T_i^k is drawn from the conditional survival function

$$\frac{\hat{S}(t)}{\hat{S}(L_i) - \hat{S}(R_i)}, \quad t \in (L_i, R_i]$$

The Kaplan-Meier estimate of the survival function is calculated for each sample. The proposed variance estimate is the sum of the sample variance of the Kaplan-Meier estimates and a variance term that is calculated in the same fashion as the Greenwood formula. Figure 3 shows the NPMLE for the RT group with confidence intervals derived from variance estimates produced by the %EMICM macro.

Figure 3 NPMLE for the RT Group with Confidence Intervals



Nonparametric Comparison of Survival Functions

The goal of a k-sample test is to test the null hypothesis H_0 that the k survival functions that correspond to k different samples are identical. For right-censored data, a common approach is to use the rankings within the combined sample to test whether the different samples come from the same population. The test that is most commonly used to compare survival functions is the log-rank test.

Suppose that each of n subjects receives one of the k treatments. The data for the k samples can be represented as $\{((L_i, R_i], \mathbf{x}_i), 1 \leq i \leq n\}$, where $\mathbf{x}_i$ is the $k \times 1$ vector of treatment indicators that are associated with subject i with interval-censored time $(L_i, R_i]$ whose l element is 1 if it is from the l population, and 0 otherwise. Let $\hat{S}_0(t)$ be the NPMLE of the common survival function $S_0(t) = \Pr(T_i > t)$ under H_0, and let $0 = s_0 < s_1 < \ldots < s_{m+1} = \infty$ be the ordered distinct time points of $\{(L_i, R_i], 1 \leq i \leq n\}$ at which $\hat{S}_0$ has jumps.

Score-Function-Based Test Procedures

For right-censored data, the log-rank test can be obtained as a score test on the proportional hazards regression model. One way to compare survival functions for interval-censored data is to perform a score test on a regression model for interval-censored data.

Various regression models for interval-censored data have been proposed: the grouped proportional hazards model of Finkelstein (1986), the discrete logistic model of Sun (1996), and the proportional odds model of Fay (1996). The survival functions are then compared by performing the score test for $\boldsymbol{\beta} = 0$, where $\boldsymbol{\beta}$ is the vector of regression coefficients for $\mathbf{x}_i$.

Let $S(t|\mathbf{X})$ be the survival function given the covariates $\mathbf{X}$, which depends on the model chosen and the value of $\mathbf{X}$. The likelihood is

$$L = L(\boldsymbol{\beta}, S_0(s_1), \ldots, S_0(s_m)) = \prod_{i=1}^{n} \sum_{j=1}^{m+1} u_{ij} [S(s_{j-1}|\mathbf{x}_i) - S(s_j|\mathbf{x}_i)]$$

where $u_{ij} = 1$ or 0, depending on whether $s_j \in (L_i, R_i]$. The score statistic for testing $\boldsymbol{\beta} = \mathbf{0}$ is

$$U = \left. \frac{\partial \log(L(\boldsymbol{\beta}, \hat{S}_0(s_1), \ldots, \hat{S}_0(s_m)))}{\partial \boldsymbol{\beta}} \right|_{\boldsymbol{\beta}=0}$$

The score statistic for each model can be expressed in the same form as the weighted log-rank statistic for right-censored data. Fay (1999) shows by simulation that tests of Finkelstein (1986) and Sun (1996) are similar, giving constant weights to differences in survival distribution over time, whereas the test of Fay (1996) gives more weight to early differences. The test of Sun (1996) is closest to the log-rank test for right-censored data, but it does not reduce to the log-rank test for right-censored data. One important drawback of these score-function-based tests is that it is hard to justify the assumptions needed for the regularity conditions of maximum likelihood.

Generalized Log-Rank Test I

Zhao and Sun (2004) generalized the log-rank test of Sun (1996) to include exact failure times in the interval-censored data. They also use an imputation approach to compute the variance of this generalized log-rank statistic. For each pair of (i, j), define α_{ij} to be the indicator of the event $s_j \in (L_i, R_i]$, $1 \le i \le n, 1 \le j \le m$. The log-rank statistic $\mathbf{U} = (U_1, \ldots, U_k)'$ is

$$U_l = \sum_{j=1}^{m} \left(d_{jl} - \frac{n_{jl} d_j}{n_j} \right)$$

where

$$\begin{aligned} d_j &= \sum_{i=1}^{n} \delta_i \frac{\alpha_{ij}[\hat{S}_0(s_j-) - \hat{S}_0(s_j)]}{\sum_{u=1}^{m+1} \alpha_{iu}[\hat{S}_0(s_u-) - \hat{S}_0(s_u)]} \\ n_j &= \sum_{r=j}^{m+1} \sum_{i=1}^{n} \delta_i \frac{\alpha_{ir}[\hat{S}_0(s_r-) - \hat{S}_0(s_r)]}{\sum_{u=1}^{m+1} \alpha_{iu}[\hat{S}_0(s_u-) - \hat{S}_0(s_u)]} + \sum_{i=1}^{n} \rho_{ij} \end{aligned}$$

and d_{ij} and n_{ij} are defined as d_i and n_i, respectively, with $\sum_{i=1}^{n}$ replaced by the summation over all subjects in population j, $\rho_{ij} = I(\delta_i = 0, L_i \ge s_j)$, and $\delta_i = 0$ for right-censored subjects, $\delta_i = 1$ otherwise.

Let M be the number of imputations for the variance estimation. For each r $(1 \le r \le M)$, an independent right-censored sample $\{(T_i^k, \delta_i^k), i = 1, \ldots, n\}$ of size n is obtained from the data $\{(L_i, R_i], 1 \le i \le n\}$ and the NPMLE $\hat{S}(t)$ as follows: $\delta_i^k = 0$ and $T_i^k = L_i$ if $(L_i, Ri]$ represents a right-censored observation; otherwise, $\delta_i^k = 1$ and T_i^k is drawn from the conditional survival function

$$\frac{\hat{S}_0(t-) - \hat{S}_0(t)}{\hat{S}_0(L_i) - \hat{S}_0(R_i)}, \quad t \in (L_i, R_i]$$

The log-rank statistic and its covariance matrix are calculated for this right-censored sample. The covariance matrix $\boldsymbol{\Sigma}$ of the generalized log-rank test is the sum of the within-imputation covariance and the between-imputation covariance. The within-imputation covariance matrix is the mean of the covariance matrices for the M imputations, and the between-imputation covariance matrix is the sample covariance matrix of the log-rank statistics for the M imputations.

Let $\boldsymbol{\Sigma}^-$ be a generalized inverse of $\boldsymbol{\Sigma}$. To test the null hypothesis that the k samples come from the same population, the test statistic $T = \mathbf{U}'\boldsymbol{\Sigma}^-\mathbf{U}$ is compared to a χ^2 distribution with $k - 1$ degrees of freedom.

The following call to the macro %ICSTEST uses the BreastCancer data set described in the section "APPENDIX: BREAST CANCER DATA" and creates Figure 4.

```
************************************************
*          Generalized Logrank Test I          *
************************************************;
%ICSTEST(data=BreastCancer,
         left=lTime,
         right=rTime,
         group=Therapy,
        );
```

The results in Figure 4 show the components of the vector $\mathbf{U}$, their covariance matrix $\boldsymbol{\Sigma}$, the test statistic T, and the associated p-value for the breast cancer data. The p-value of 0.04 indicates a statistically significant difference between the two groups, and examination of the plots in Figure 1 of the survival function estimates indicates that the RT group generally has longer times to retraction than the RCT group.

Figure 4 Generalized Log-Rank Test I for the Breast Cancer Data

```
Generalized Log-rank Test (Zhao & Sun, 2004)

    Test Statistic and Covariance Matrix

Therapy            U       cov(U)

      1      -9.1418      18.9658     -18.9658
      2       9.1418     -18.9658      18.9658

   Chi-Square        DF Pr > Chi-Square

       4.4065         1          0.0358
```

Generalized Log-Rank Test II

Sun, Zhao, and Zhao (2005) propose a new class of k-sample test for interval-censored data and develop the asymptotics. Consider a combined sample of n subjects from k populations with n_l subjects in the lth sample; that is, $n_1+\ldots+n_k = n$. Let $\mathbf{x}_i$ be the $k \times 1$ vector of treatment indicators that are associated with subject i with the interval-censored time $(L_i, R_i]$ whose lth element is 1 if it is from the lth population, and zero otherwise. Let $\xi(u)$ be a known function over (0,1) such that $\lim_{u\to 0}\xi(u) = \lim_{u\to 1}\xi(u) = c_0$ for some constant c_0; typically, $\xi(u) = u\log(u)$ is used. Denote

$$K_i(L_i, R_i) = \frac{\xi[\hat{S}_0(L_i)] - \xi[\hat{S}_0(R_i)]}{\hat{S}_0(L_i) - \hat{S}_0(R_i)}$$

The k-sample test statistic proposed by Sun, Zhao, and Zhao (2005) is

$$\mathbf{U}_n = \sum_{i=1}^{n} \mathbf{x}_i K_n(L_i, R_i)$$

This test statistic includes the score test statistic of Finkelstein (1986) as a special case when $\xi(u) = u\log(u)$ and is also asymptotically equivalent to the score statistic of Sun (1996) for the same ξ. Under the null hypothesis of no treatment differences, the $\frac{1}{n}\mathbf{U}_n$ has an asymptotically normal distribution with covariance matrix $\mathbf{\Sigma} = (\sigma_{lr})$ which can be consistently estimated by $\hat{\mathbf{\Sigma}}_n = (\hat{\sigma}_{lr})$ given by

$$\hat{\sigma}_{lr} = \begin{cases} \frac{n_l(n-n_l)}{n^2}\hat{Q}_n & \text{if } l = r \\ \frac{-n_l n_r}{n^2}\hat{Q}_n & \text{otherwise} \end{cases}$$

where $\hat{Q}_n = \frac{1}{n}\sum_{i=1}^{n} K_n^2(L_i, R_i)$. Let $\mathbf{U}_n^*$ be the first $k-1$ components of $\mathbf{U}_n$, and let $\hat{\mathbf{\Sigma}}_n^*$ be the matrix that is derived by deleting the last row and column of $\hat{\mathbf{\Sigma}}$. The null hypothesis of the homogeneity of the k populations can be tested by comparing the statistic $\frac{1}{n}\mathbf{U}_n^{*\prime}\hat{\mathbf{\Sigma}}_n^{*-1}\mathbf{U}_n^*$ to a χ^2 distribution with $k-1$ degrees of freedom.

The following call to the macro %ICE uses the BreastCancer data set described in the section "APPENDIX: BREAST CANCER DATA" and creates Figure 5.

```
************************************************
*          Generalized Logrank Test II         *
************************************************;
%ICE(data=BreastCancer,
     time=(lTime,rTime),
     group=Therapy,
     options=notable);
```

The results in Figure 5 compare the two therapies (RT and RCT) for the breast cancer data. The graph of the NPMLE survival functions in Figure 1 for the two therapies provides a visual indication that the RT group has a longer time to retraction; the two-sample generalized log-rank test results in Figure 5 confirm a statistically significant difference between the two groups.

Figure 5 Generalized Log-Rank Test II for the Breast Cancer Data

```
                    Number of Observations by Group

                           Therapy          N

                           RCT             48
                           RT              46


              Generalized Log-Rank Test (Sun, Zhao, and Zhao, 2005)
                            xi(x)=xlog(x)

                     Test Statistic and Covariance Matrix

                  Therapy          U      cov(U)

                  RCT          9.9442    13.5820   -13.5820
                  RT          -9.9443   -13.5820    13.5820

                        ChiSquare    DF Pr>ChiSquare

                           7.2807     1       0.0070
```

Regression Model

The Cox (1972) proportional hazards model is the most commonly used regression model for survival data with right-censoring. Many have attempted to fit a proportional hazards model to interval-censored data. Finkelstein (1986) uses a discrete baseline survival, and the estimation is based on a full likelihood under the proportional hazards model. The number of parameters might increase with the number of event times, rendering numerically unstable optimization. Goggins et al. (1998) propose a Monte Carlo EM algorithm to fit the proportional hazard model. Goetghebeur and Ryan (2000) use a different approach which uses an EM algorithm. Software is not currently available for these approaches.

CONCLUSION

Methods for the analysis of right-censored survival data are well developed, and software is widely available to implement the methods. The analysis of interval-censored survival data is of growing importance. Methods for analysis of interval-censored survival data have been developed over the past two decades, but they are more complicated and harder to implement than their right-censored counterparts. In this paper, nonparametric methods for the analysis of interval-censored survival data have been surveyed, and SAS macros have been presented to address two important statistical issues: the estimation of survival functions, and comparison of survival functions from multiple populations.

APPENDIX: BREAST CANCER DATA

The breast cancer data (Finkelstein and Wolfe 1985) is used to illustrate the methodologies presented in this paper. The following SAS statements create the data set BreastCancer which is used to illustrate the methods and SAS macros presented in this paper.

```
************************************************
*         Radidation Therapy (RT)              *
*                                              *
* lTime and rTime represent the left and right *
* endpoint of the interval time, respectively  *
************************************************;
data RT;
   input lTime rTime @@;
   datalines;
45 . 25 37 37 .
6 10 46 . 0 5
0 7 26 40 18 .
46 . 46 . 24 .
46 . 27 34 36 .
7 16 36 44 5 11
17 . 46 . 19 35
7 14 36 48 17 25
37 44 37 . 24 .
0 8 40 . 32 .
4 11 17 25 33 .
```

```
15 . 46 . 19 26
11 15 11 18 37 .
22 . 38 . 34 .
46 . 5 12 36 .
46 .
;

************************************************
*     Radidation and Chemoherapy (RCT)         *
*                                              *
* lTime and rTime represent the left and right *
* endpoints of the interval time, respectively *
************************************************;
data RCT;
   input lTime rTime @@;
   datalines;
8 12 0 5 30 34
0 22 5 8 13 .
24 31 12 20 10 17
17 27 11 . 8 21
17 23 33 40 4 9
24 30 31 . 11 .
16 24 13 39 14 19
13 . 19 32 4 8
11 13 34 . 34 .
16 20 13 . 30 36
18 25 16 24 18 24
17 26 35 . 16 60
32 . 15 22 35 39
23 . 11 17 21 .
44 48 22 32 11 20
14 17 10 35 48 .
;

proc format;
   value Rx 1="RT" 2="RCT";
   run;

data BreastCancer;
   set RT (in=ina) RCT;
   if ina then Therapy=1;
   else        Therapy=2;
   format Therapy Rx.;
   run;
```

APPENDIX: SAS MACROS FOR INTERVAL-CENSORED DATA

The macros described in this appendix are available at `http://support.sas.com/kb/24980`.

%EMICM Macro

The %EMICM macro computes the NPMLE of the survival function:

```
%EMICM(
   DATA=      /* Input SAS data set. Default is _last_.                          */

   LEFT=      /* Variable name of the left endpoint of the time interval.       */

   RIGHT=     /* Variable name of the right endpoint of the time interval.      */

   GROUP=     /* Variable identifying different treatment groups. A separate NPMLE is  */
              /* computed for each value of the GROUP= variable.                */

   METHOD=    /* Select the method to compute the NPMLEs. Default is METHOD= EMICM.    */
              /* User can choose METHOD= EM or METHOD= ICM.                     */

   OUT=       /* Name of an output data set that contains the NPMLE.            */

   OUTITER=   /* Name of an output data set that contains the history of iterations.   */
              /* Variables 'ERROR1', 'ERROR2', 'ERROR3', and 'ERROR4' correspond to    */
```

```
              /* ERRORTYPE=1, ERRORTYPE=2, ERRORTYPE=3, and ERRORTYPE=4, respectively.     */

   ERRORTYPE= /*  Convergence criterion used.                                              */
              /*  1 -- The maximum of the closeness of consecutive estimates,             */
              /*  2 -- The closeness of the log likelihood function,                      */
              /*  3 -- The gradient of the log likelihood function,                       */
              /*  4 -- The maximum measures of ERRORTYPE=1, ERRORTYPE=2, and ERRORTYPE=3. */
              /*  Default is ERRORTYPE=1.

   RATECONV=  /*  Rate of convergence. Default is RATECONV=1e-7.                          */

   mRS=       /*  Number of resampling used for the generalized Greenwood formula.        */
              /*  Default is mRS = 50.                                                    */

   TITLE=     /*  Primary title for plot.                                                 */

   TITLE2=    /*  Secondary title for plot.                                               */

   TIMELABEL= /*  Label for the time axis.                                                */

   OPTIONS=   /* Display options (separated by blanks):                                   */
              /* NOTABLE -- Suppresses printing the tables of estimated survival curves   */
              /* PLOT    -- displays the estimated survival curves using ODS graphics.    */

);
```

You can choose between the EM algorithm of Turnbull (1976), the ICM algorithm of Groeneboom and Wellner (1992), and the EM-ICM algorithm of Wellner and Zhan (1997). The estimated variance is computed based on the generalized Greenwood formula (Sun 2001). A plot of the survival curves can be displayed using ODS Graphics. Figure 1 is an example of the NPMLE of the survival function computed using %EMICM for the breast cancer data.

%ICSTEST Macro

The %ICSTEST macro computes the generalized log-rank test of Zhao and Sun (2004), which is described in the section "Generalized Log-Rank Test I":

```
%ICSTEST(
   DATA=      /* Input SAS data set. Default is _last_.                                   */

   LEFT=      /* Variable name of the left endpoint of the time interval.                */

   RIGHT=     /* Variable name of the right endpoint of the time interval.               */

   GROUP=      /* Variable identifying different treatment groups for comparison         */

   ERRORTYPE= /*  Convergence criterion used.                                            */
              /*  1 -- The maximum of the closeness of consecutive estimates,           */
              /*  2 -- The closeness of the log likelihood function,                    */
              /*  3 -- The gradient of the log likelihood function,                     */
              /*  4 -- The maximum measures of ERRORTYPE=1, ERRORTYPE=2, and ERRORTYPE=3. */
              /*  Default is ERRORTYPE=1.

   RATECONV=  /*  Rate of convergence. Default is RATECONV=1e-7.                        */

   mRS=       /*  Number of resampling used for the generalized Greenwood formula.      */
              /*  Default is mRS = 50.                                                  */
);
```

The results in Figure 4 were computed using the %ICSTEST macro.

%ICE Macro

The %ICE macro computes the NPMLE of survival function and the log-rank test, which is described in the section "Generalized Log-Rank Test II":

```
%ICE(
   DATA=      /* Input  SAS data set.

   GROUP=     /* Variable identifying different treatment groups. A separate NPMLE is      */
              /* computed for each value of the GROUP= variable. A k-sample test          */
              /* comparing the treatment groups is also conducted.                        */

   TIME=      /* Two variables (separated by blanks) representing the left and right      */
              /* endpoints of the time interval. You may enclose these variable names     */
              /* by a pair of parentheses, but a comma should not be used to separate the */
              /* names.                                                                   */

   FREQ=      /* A single numeric variable whose values represent the frequency of        */
              /* occurrence of the observations.                                          */

   TECH=      /* Optimization technique for maximizing the likelihood. Valid values are:  */

              /* NRA --  Newton-Raphson Ridge                                             */
              /* QN  --  Quasi-Newton                                                     */
              /* CG  --  Conjugate Gradient                                               */
              /* EM  --  Self-Consistency Algorithm of Turnbull                           */

              /* NRA, QN and CG are NLP optimization routines. EM is the self-consistency */
              /* algorithm. With m as the number of estimated parameters, the default     */
              /* technique is

              /*      NRA  if m <=30                                                      */
              /*      QN   if 30 < m <= 200                                               */
              /*      CG   if m > 200                                                     */


   LBOUND=    /* Lower bound for the estimated parameters. The default is 1e-6. Only used */
              /* in the NRA, QN and CG techniques.                                        */

   ALPHA=     /* A number between 0 and 1 that sets the level of the confidence intervals */
              /* for the survival curve. The confidence level for the intervals is 1-ALPHA.*/
              /* The default is .05.                                                      */

   OPTIONS=   /* List of display options (separated by blanks):

              /* NOTABLE  Suppress printing of the parameter estimates, the survival curve */
                          estimates and confidence limits for the survival curve.          */
              /* PLOT     Graphical display of the estimated survival curve.              */

   NLPOPT=    /* An IML row vector to be passed into the OPT argument of the NLP          */
              /* optimization routines. This vector controls the option vector of the     */
              /* NLP optimization routine. The default is {1 0}.                          */

   NLPTC=     /* An IML row vector to be passed into the TC argument of the NLP           */
                 optimization routine. This vector controls the termination criteria of   */
                 the NLP optimization routine. The default is {2000 5000}.                */

   EMCONV=    /* Convergence criterion for the EM technique. Convergence is declared if   */
              /* the increase in the log-likelihood is less than the convergence criterion.*/
              /* The default is 1e-8.                                                     */

   OUTE=      /* A SAS data set name containing the parameter estimates.                  */

   OUTS=      /* A SAS data set name containing the estimates of the survival curve and   */
              /* the corresponding confidence limits.                                     */

);
```

The EM algorithm of Turnbull (1976) and various Newton methods for maximizing the log-likelihood are available for computing the NPMLE. The estimated variance is computed by inverting the negative of the Hessian matrix evaluated at the NPMLE. The generalized log-rank test of Sun, Zhao, and Zhao (2005) described in the section "Generalized Log-Rank Test II" has been added to the macro. The results in Figure 5 were computed using the %ICE macro.

REFERENCES

Chen, C. (2009), "Empirical Comparison between Conventional Approach and Finkelstein's Method," DIA/FDA/PHRMA PFS Workshop, Oct. 7–9, Bethesda, MD.

Cox, D. R. (1972), "Regression Models and Life Tables," *Journal of the Royal Statistical Society, Series B*, 20, 187–220, with discussion.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society, Series B*, 39, 1–38.

Efron, B. and Tibshirani, R. (1993), *An Introduction to the Bootstrap*, New York: Chapman & Hall.

Fay, M. P. (1996), "Rank Invariant Tests for Interval Censored Data under the Grouped Continuous Model," *Biometrics*, 52, 811–822.

Fay, M. P. (1999), "Comparing Several Score Tests for Interval Censored Data," *Statistics in Medicine*, 18, 273–285.

Finkelstein, D. M. (1986), "A Proportional Hazards Model for Interval-Censored Failure Time Data," *Biometrics*, 42, 845–854.

Finkelstein, D. M. and Wolfe, R. A. (1985), "A Semiparametric Model for Regression Analysis of Interval-Censored Failure Time Data," *Biometrics*, 41, 933–945.

Gentleman, R. and Geyer, C. J. (1994), "Maximum Likelihood for Interval Censored Data: Consistency and Computation," *Biometrika*, 81, 618–623.

Goetghebeur, E. and Ryan, L. (2000), "Semiparametric Regression Analysis of Interval-Censored Data," *Biometrics*, 56, 1139–1144.

Goggins, W. B., Finklestein, D. M., Schoenfeld, D. A., and Zaslavsky, A. M. (1998), "A Markov Chain Monte Carlo EM Algorithm for Analyzing Interval-Censored Data under the Cox Proportional Model," *Biometrics*, 54, 1498–1507.

Groeneboom, P. and Wellner, J. A. (1992), *Information Bounds and Nonparametric Maximum Likelihood Estimation*, New York: Birkhauser.

Lawless, J. F. (2003), *Statistical Model and Methods for Lifetime Data*, Second Edition, New York: John Wiley & Sons.

Peto, R. (1973), "Experimental Survival Curves for Interval-Censored Data," *Applied Statistics*, 22, 86–91.

Prentice, P. L. and Gloeckler, L. A. (1978), "Regression Analysis of Grouped Survival Data with Applications to Breast Cancer Data," *Biometrics*, 34, 57–67.

Sun, J. (1996), "A Nonparametric Test for Interval-Censored Failure Time Data with Application to AIDS Studies," *Statistics in Medicine*, 15, 1387–1395.

Sun, J. (2001), "Variance Estimation of a Survival Function for Interval-Censored Survival Data," *Statistics in Medicine*, 20, 1249–1257.

Sun, J. (2006), *The Statistical Analysis of Interval-Censored Failure Time Data*, New York: Springer.

Sun, J., Zhao, Q., and Zhao, X. (2005), "Generalized Log-Rank Test for Interval-Censored Failure Time Data," *Scandinavian Journal of Statistics*, 32, 49–57.

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.

Wellner, J. A. and Zhan, Y. (1997), "A Hybrid Algorithm for Computation of the Nonparametric Maximum Likelihood Estimator from Censored Data," *Journal of the American Statistical Association*, 92, 945–959.

Zhao, Q. and Sun, J. (2004), "Generalized Log-Rank Test for Mixed Interval-Censored Failure Time Data," *Statistics in Medicine*, 23, 1621–1629.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author:

Ying So	Gordon Johnston
SAS Institute Inc.	SAS Institute Inc.
SAS Campus Drive	SAS Campus Drive
Cary, NC 27513	Cary, NC 27513
ying.so@sas.com	gordon.johnston@sas.com