ELSEVIER

Statistics & Probability Letters 58 (2002) 221–232

STATISTICS & PROBABILITY LETTERS

www.elsevier.com/locate/stapro

Survival estimation and testing via multiple imputation

Jeremy M.G. Taylor*, Susan Murray, Chiu-Hsieh Hsu

Department of Biostatistics, School of Public Health, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109-2029, USA

Received January 2001; received in revised form October 2001

Abstract

Multiple imputation is a technique for handling data sets with missing values. The method fills in each missing value several times, creating many augmented data sets. Each augmented data set is analyzed separately and the results combined to give a final result consisting of an estimate and a measure of uncertainty. In this paper we consider nonparametric multiple-imputation methods to handle missing event times for censored observations in the context of nonparametric survival estimation and testing. Two nonparametric imputation schemes are considered. In risk set imputation the censored time is replaced by a random draw of the observed times amongst those at risk after the censoring time. In Kaplan–Meier (KM) imputation the imputed time is a draw from the estimated distribution of event times amongst those at risk after the censoring time. We show that with a large number of imputes the estimates from both methods reproduce the KM estimator. In a simulation study we show that the inclusion of a bootstrap stage in the multiple imputation algorithm gives coverage rates of confidence intervals that are comparable to that from Greenwood's formula. Connections to the redistribute to the right algorithm are discussed.

Keywords: Kaplan–Meier estimate; Multiple imputation; Redistribute to the right

1. Introduction

Nonparametric survival analysis tools have earned popularity over the years in estimation and testing problems due to robustness of inference when subjected to unknown underlying data mechanisms. Among the most frequently used methods are the Kaplan–Meier (KM) survival estimate, which is a nonparametric maximum likelihood estimate for survival, and the logrank and generalized Wilcoxon test for comparing two survival curves. From the beginning these procedures have used intelligent approaches for handling censored data making them appealing for use in practical situations, such as clinical trials with finite amounts of follow-up time. Later Efron (1967) rederived

* Corresponding author.

0167-7152/02/$ - see front matter
PII: S0167-7152(02)00030-5

the KM estimate through a redistribute to the right algorithm (RRA), verifying the attractive appeal of this estimator from a different intuitive strategy of handling censored values.

There is a large literature on approaches to handling missing data outside of the field of survival analysis. However, it is only in recent years that these sets of tools are beginning to be applied in survival analysis for handling the missing failure time information in censored observations. Multiple imputation (Rubin, 1978, 1987) has become a popular strategy for analyzing data subject to various missing data mechanisms, particularly missing completely at random (MCAR) and missing at random (MAR). In the language of missing data censored data are strictly speaking nonignorably missing, albeit in a benign sense. Natural generalizations of MAR and MCAR to censoring are given by Heitjan (1994).

The theoretical underpinnings of multiple imputation are Bayesian. The central idea is to fill in the missing values by drawing from the posterior predictive distribution of the missing data given the observed data. The procedure is independently repeated M times. Each filled-in dataset is analyzed separately and the results combined following well established rules. A draw from the predictive distribution is frequently achieved in two stages. In the first the parameters are drawn from their posterior distribution and in the second the missing value is drawn conditional on the parameter and the observed data. Such two stage procedures that account for the uncertainty in the parameter estimates are sometimes refered to as proper. A bootstrap can be used as an approximation in the first stage, and has been shown to have good properties (Rubin and Schenker, 1991; Heitjan and Little, 1991).

The research in this paper describes nonparametric multiple imputation procedures for analyzing censored survival data and draws connections between these imputation-based methods, the redistribution to the right algorithm and standard KM estimates. The aim of this paper is to provide some theoretical basis and foundation for the use of multiple imputation in survival analysis. The situations we study are quite simple; the results we present do, however, provide some positive evidence regarding what might happen in more complex situations where multiple imputation would be more useful. This will be the topic of a future paper.

In Section 2, we review the redistribute to the right algorithm. In Section 3, we describe the procedure for nonparametric multiple imputation in the context of censored survival analysis. In Section 4, we show the relationship between survival estimation using nonparametric multiple imputation and the original KM estimator. In Section 5, we study properties of imputation procedures for survival analysis in finite sample sizes through simulation. A discussion follows in Section 6.

2. The redistribution of the right algorithm

Let $T_1,\ldots,T_n$ denote times to the outcome of interest for n subjects under study and $C_1,\ldots,C_n$ the corresponding potential censoring times. The observable random variables are $X_i = \min(T_i, C_i)$ and $\Delta_i = I(T_i \leqslant C_i)$. Let n_{obs} denote the number of observed failures, $n_{\text{mis}} = n - n_{\text{obs}}$ denote the number of censored values, $\{t_1,\ldots,t_n\}$ denote the ordered observed failure and censoring times and $R(j^+) = \{i\colon X_i > t_j,\ i = 1,\ldots,n\}$ denote the risk set at t_j excluding individuals with $X_i = t_j$.

In the absence of censoring, each individual is allocated a weight $w_i = 1/n,\ i = 1,\ldots,n$ and a survival estimate is $\hat{S}(t_j) = 1 - \{\sum_{i \notin R(j^+)} w_i\}$. Extending this procedure to include censored outcomes, the RRA takes weights associated with censored individuals, at say t_j, and reallocates them equally to

individuals in $R(j^+)$. The reallocation of weights begins at the smallest censored value of t_j, and is repeated for all censored observations. The survival estimate $\hat{S}(t_j)$, which corresponds to the KM estimate, is estimated as above using the redistributed weights. The RRA assumes that censored values behave similarly to uncensored counterparts in the risk set, a philosophy that is useful in considering multiple imputation approaches to handling censored survival data. An example of the RRA procedure can be found in Miller (1981, p. 53).

3. Imputation methods

In this section, we describe two strategies for nonparametric multiple imputation with censored survival data. In each imputation method we consider that censored or missing outcome values are imputed from a distribution derived from those remaining at risk. Hence, the first of two steps in each imputation method is to identify this distribution, followed by the random selection of the imputed value in the second step. Once the new data set is created, the procedure can be independently repeated M times to obtain multiple imputed data sets for use in estimation and testing.

3.1. Risk set imputation (RSI)

For each of the n_{mis} observed censored times t_j, the RSI method imputes a pair (t_j^*, δ_j^*) drawn at random from the observed pairs (X, Δ) of those individuals in $R(j^+)$. Hence for each censored time t_j the RSI method is equally likely to draw any of the observed failure or censored times from those individuals still at risk at t_j^+. This procedure begins with the smallest censored time and proceeds in order to the largest censored time. We note that each censored case is only imputed once, thus an observation which is imputed as censored is not reimputed, and that imputed data values are not included in the risk set for other censored observations. If the last observed event is censored, then it retains its value since the risk set does not contain any possible donors.

3.2. Kaplan–Meier imputation (KMI)

An alternative method draws an event time from a KM estimator of the distribution of event times among those at risk. Thus, the procedure imputes only observed failure times unless the last time is censored in which case some imputed times may include this last censored time. Specifically, for each censored time t_j, a KM survival curve, $\hat{S}_{j^+}(t)$, is estimated from among those individuals in $R(j^+)$. Then the KMI method imputes a value t_j^* from the corresponding estimated cumulative distribution function (cdf) $1 - \hat{S}_{j^+}(t)$, by simulating a uniform $(0,1)$ value and choosing t_j^* to correspond to that value of the observed cdf. Since jumps in the cdf occur only at observed failure times, they are selected for imputation with collective probability $1 - \hat{S}_{j^+}(t_n)$. Hence if the last event time is censored, it will be imputed for the censored value t_j with probability $1 - \hat{S}_{j^+}(t_n)$. No imputation occurs for the last event time if it is censored since there are no individuals from $R(j^+)$ from which to construct $\hat{S}_{j^+}(t)$. This procedure begins with the smallest censored time and proceeds to the largest. We note that imputed times t_j^* are not used in KM estimates for imputation at later times.

3.3. Bootstrap imputation procedure

The RSI and KMI procedures by themselves do not incorporate the full uncertainty in the imputes, because they do not include a first stage of an initial parameter draw, thus they would not be viewed as proper multiple-imputation schemes. The RSI and KMI procedures can be enhanced by including a Bootstrap stage in the procedure, which is designed to make them proper. Consider the bootstrap sample $\{(t_1^{(B)}, \delta_1^{(B)}), \ldots, (t_n^{(B)}, \delta_n^{(B)})\}$ selected with replacement from the original data set. The imputing risk set for the censored time t_j can be redefined as $R^{(B)}(j^+) = \{i\colon t_i^{(B)} > t_j,\ i = 1, \ldots, n\}$, i.e. those observations that are at risk at time t_j in the bootstrap sample. If $R^{(B)}(j^+)$ is empty, then the observation retains its original value. For each censored time t_j, KMI and RSI methods incorporating bootstrap methods, hereafter denoted as KMIB and RSIB methods of imputation, impute a value $t_j^{(B^*)}$ from the estimated distribution function, $(1 - \hat{S}_{j^+}^{(B)}(t))$, or draw a pair $(t_j^{(B^*)}, \delta_j^{(B^*)})$ from the risk set $R^{(B)}(j^+)$, respectively. Multiple imputations are created by independently repeating the bootstrap stage for each of the M data sets.

We note that the difference between RSIB and RSI or between KMIB and KMI is thus analagous to the difference between the approximate Bayesian bootstrap method of Rubin and Schenker (1991) and simple random sampling procedures for imputation.

3.4. Analyzing a multiply imputed data set

3.4.1. Estimation

The methods for analyzing multiply imputed data sets follow well established rules (Rubin and Schenker, 1991). With M sets of imputations from a multiple-imputation scheme, there are M enhanced data sets and hence M survival estimates at each time t with associated variances, say $\hat{S}_m(t)$ and U_m, respectively, $m = 1, \ldots, M$. Survival estimates are computed using the KM method and associated variances are based on the Greenwood formula. The final survival estimate of $S(t)$ is then the average of the M enhanced-data estimates: $\bar{S}(t) = \sum_{m=1}^{M} \hat{S}_m(t)/M$. The variability associated with $\bar{S}(t)$ is $W = \bar{U} + (1 + M^{-1})B$, which has two components: the average within-imputation variance, $\bar{U} = \sum_{m=1}^{M} U_m/M$, and the between-imputation component $B = \sum_{m=1}^{M} \{\hat{S}_m(t) - \bar{S}(t)\}^2/(M-1)$. Interval estimates and significant tests are based on a t distribution: $(\bar{S}(t) - S(t))W^{-1/2} \sim t_v$, where the degrees of freedom v is given by $v = [1 + (M/(M+1))\bar{U}/B]^2(M-1)$.

3.4.2. Testing

We consider and compare two approaches to test the equality of two survival curves with multiply imputed data. The first is a direct application of the procedure in Li et al. (1991), the second is a variant on this procedure. The procedure described by Li et al. (1991) for testing a null hypothesis that a one-dimensional parameter θ equals θ_0, is an extension of the complete data method which compares $(\hat{\theta} - \theta_0)' U^{-1} (\theta - \theta_0)$ with a chi-squared distribution with one degree of freedom, where U is an estimate of the variance of $\hat{\theta}$. In the case of missing data following multiple imputation and the construction of M complete data sets the statistic $D = (\bar{\theta} - \theta_0)' \bar{U}^{-1} (\bar{\theta} - \theta_0)/(1 + r)$ is compared with a $F_{1,w}$ distribution. In this expression $\bar{\theta} = \sum \theta_m/M$, $\bar{U}$ is the average of the M variance estimates, $r = (1 + M^{-1})BU^{-1}$ where $B = \sum (\theta_m - \bar{\theta})^2/(M-1)$, and $w = 4 + (t-4)(1 + (1 - 2t^{-1})/r)$, where

$t = M - 1$. An alternative choice for the degrees of freedom w is given by $w = (M-1)(1+1/r)^2$ (Rubin, 1987).

When comparing two survival curves, we denote the test statistic (either log-rank or Wilcoxon) by R_m and its standard error by E_m, for each completed dataset. Let $Z_m = R_m/E_m$. In the first approach we consider θ to be the parameter being estimated by R_m, so $\theta_0 = 0$ under the null hypothesis that the two curves are equal. Direct application of the Li et al. (1991) procedure gives $D = \bar{R}^2/(\bar{U} + (1 + M^{-1})B(R))$, where $\bar{U}$ is the average of E_m^2, and $B(R) = \sum (R_m - \bar{R})^2/(M-1)$. The second approach is a t-based test comparing the average of the Z_m's to their standard error. That is, it is based on $\bar{Z}/W^{1/2} \sim t_v$, where $\bar{Z} = \sum Z_m/M, W = 1 + (1 + M^{-1})B(Z)$, $v = [1 + (M/(M+1))1/V(Z)]^2(M-1)$ and $B(Z) = \sum (Z_m - \bar{Z})^2/(M-1)$. Approach 2 is therefore an application of the Li et al. (1991) procedure with θ denoting the parameter being estimated by Z_m. In summary, approach 1 is based on the ratio of the averages of the numerator and denominator of the test statistic, while approach 2 is based on the average of the ratios in the standardized test statistic from the multiply imputed data. For illustration purposes we use $M = 10$ when we apply this procedure. This makes the estimated degrees of freedom very large, for both methods of calculation of the degrees of freedom, such that the t distribution is negligibly different from a normal distribution.

4. Relationship between KMI, RSI and KM estimates

For simplicity, we assume no ties in the event times of the original data set. Let $\mathbf{Y} = \{(t_1, \delta_1), \ldots, (t_n, \delta_n)\}$, where recall t_i is the ith ordered observed time with corresponding censoring indicator δ_i. For each observed censored value t_i, we will impute a value t_i^*, and an associated censoring indicator δ_i^*.

Because the KMI and RSI survival estimates may impute censored event times differently across multiple-imputed data sets, we require notation that can be used to describe risk sets in relation to imputed values. Consider a censored time t_i and suppose we are describing the risk set at a later time t_j for an imputed dataset, where $i < j$. Behavior of the potential imputed value t_i^* in relation to t_i and t_j can take three relevant forms and it will become convenient to have indicator functions depicting the different possibilities. Let $\gamma_{1,i} = I(t_i < t_i^* < t_j)$, $\gamma_{2,i} = I(t_i^* = t_j)$ and $\gamma_{3,i} = I(t_j < t_i^*)$. Each of these three indicator functions vanish in the case t_i as an observed failure time with no need to impute. Hence $\Gamma_{1,j} = \sum_{i=1}^{j-1} \gamma_{1,i}$ counts the number of imputes $< t_j, \Gamma_{2,j} = \sum_{i=1}^{j-1} \gamma_{2,i}$ counts the number of imputes equal to t_j and $\Gamma_{3,j} = \sum_{i=1}^{j-1} \gamma_{3,i}$ counts the number of imputes $> t_j$. Therefore, the number of individuals in the risk set at time t_j after imputing censored values is $n_j^* = n_j + \Gamma_{2,j} + \Gamma_{3,j} = n_j + \Gamma_{3,j-1}$ where n_j is the number of subjects in the risk set at time t_j in the original data set, and the number of events at t_j in the imputed data set is $1 + \Gamma_{2,j}$.

Using the above notation, the KM survival estimator for the original data set at time t_j is

$$\hat{S}_{\mathrm{KM}}(t_j) = \prod_{t_i \leqslant t_j} \left(1 - \frac{1}{n_i}\right)^{\delta_i}, \quad j = 1, \ldots, n.$$

The two imputation-based methods will be altered according to the number of death times observed for each t_i and the corresponding risk set size. So, for any one particular imputed dataset, the KMI

survival estimator at time t_j can be written as

$$\hat{S}_{\mathrm{KMI}}(t_j) = \prod_{t_i \leqslant t_j} \left(1 - \frac{1 + \Gamma_{2,i}^{\mathrm{KMI}}}{n_i^{\mathrm{KMI}}}\right)^{\delta_i}, \quad j = 1, \ldots, n,$$

where $\Gamma_{2,i}^{\mathrm{KMI}}$ and n_i^{KMI} are $\Gamma_{2,i}$ and n_i^* with regard to the KMI method at time t_i.

Similarly, the RSI survival estimator at time t_j can be written as

$$\hat{S}_{\mathrm{RSI}}(t_j) = \prod_{t_i \leqslant t_j} \left(1 - \frac{1 + \Gamma_{2,i}^{\mathrm{RSI}}}{n_i^{\mathrm{RSI}}}\right)^{\delta_i}, \quad j = 1, \ldots, n,$$

where $\Gamma_{2,i}^{\mathrm{RSI}}$ and n_i^{RSI} are $\Gamma_{2,i}$ and n_i^* with regard to the RSI method at time t_i.

The properties of KMI and RSI survival estimates for a large number of imputes are summarized in the following two results. The proofs are outlined in the Appendix A:

Result 1: $E\{\hat{S}_{\mathrm{RSI}}(t_j)|\mathbf{Y}\} = \hat{S}_{\mathrm{KM}}(t_j), \quad j = 1, \ldots, n,$

Result 2: $E\{\hat{S}_{\mathrm{KMI}}(t_j)|\mathbf{Y}\} = \hat{S}_{\mathrm{KM}}(t_j), \quad j = 1, \ldots, n.$

In these expressions the expectation is with respect to the distribution of possible imputes. Since the final multiple-imputation estimates are the average of estimates from each imputed data set, the above two results show that the KMI and RSI survival point estimates will be equivalent to the KM estimator if the number of imputes is infinitely large.

5. Simulation study

We performed a small simulation study to investigate the properties of the multiple-imputation-based procedures. For the estimates of the survival distribution we investigated bias, variance and coverage rates of confidence intervals, and how these were affected by sample size, censoring rate and by the inclusion of the bootstrap stage in the multiple-imputation procedure. For the two sample test statistics we investigated size and power. In all cases we used $M = 10$ for multiple imputation, with 1000 replications.

5.1. Comparison of survival function estimates

The event and censoring times were generated from the exponential distribution with parameters chosen to give a range of censoring rates.

In Table 1 we display estimates of $S(t)$ at three times corresponding to the 35th, 50th and 75th percentiles of the survival function. The four imputation methods, KMI, RSI, KMIB, and RSIB, described in Section 3 were considered, as well as the KM estimator applied to the original data. For each method we calculate the average of the 1000 point estimates (denoted by average), the empirical standard deviation of the 1000 point estimates (denoted by SD), the average of the 1000 estimated standard errors (denoted by SE) and the fraction of 95% confidence intervals which contain

Table 1
Monte Carlo results: survival estimates of three quantities (75th, 50th, and 35th) and associated standard deviations, standard errors, and coverage rates of nominal 95% confidence intervals for the original data and for multiple imputations[a]

Method	Censoring rate	True value	Average	SD	SE	Coverage rate
KM	0.23	0.75	0.748	0.0706	0.0689	93.6
KMI			0.748	0.0707	0.0688	93.4
RSI			0.749	0.0707	0.0688	93.0
KMIB			0.748	0.0706	0.0689	93.6
RSIB			0.748	0.0706	0.0689	93.4
KM	0.23	0.50	0.501	0.0852	0.0827	91.6
KMI			0.501	0.0853	0.0818	91.4
RSI			0.501	0.0855	0.0820	91.6
KMIB			0.501	0.0850	0.0825	91.6
RSIB			0.501	0.0852	0.0825	91.2
KM	0.23	0.35	0.349	0.0813	0.0816	94.0
KMI			0.350	0.0816	0.0798	94.0
RSI			0.350	0.0816	0.0802	94.0
KMIB			0.349	0.0814	0.0812	94.0
RSIB			0.349	0.0815	0.0813	94.0
KM	0.50	0.75	0.749	0.0744	0.0726	92.7
KMI			0.749	0.0747	0.0718	91.8
RSI			0.749	0.0748	0.0720	91.9
KMIB			0.749	0.0746	0.0730	93.0
RSIB			0.749	0.0749	0.0728	92.5
KM	0.50	0.50	0.501	0.1003	0.0952	93.6
KMI			0.501	0.1014	0.0890	91.1
RSI			0.501	0.1013	0.0910	91.1
KMIB			0.500	0.1021	0.0960	92.4
RSIB			0.501	0.1024	0.0955	92.1
KM	0.50	0.35	0.350	0.1056	0.1023	92.2
KMI			0.350	0.1069	0.0878	87.1
RSI			0.350	0.1062	0.0935	89.1
KMIB			0.350	0.1089	0.1039	91.9
RSIB			0.350	0.1080	0.1025	91.4

[a]The event times $\sim$ exponential with mean $= 2.0$ and censoring times $\sim$ exponential with mean $= 6.7$ or 2.0, the sample size is 40. (KM: Kaplan–Meier estimate and Greenwood variance formula, KMI: Kaplan–Meier-based, imputation, RSI: risk set imputation, KMIB: Kaplan–Meier-based imputation using bootstrap, and RSIB: risk set imputation using bootstrap.) Results based on 1000 replications and $M = 10$.

the true value (denoted by coverage rate). Each confidence interval is calculated as estimate $+t_{\nu}^{(0.975)}$ standard error.

The results in Table 1 show that all five estimates target their quantile correctly. As expected KMI and RSI yield almost identical survival estimates and standard deviations as the KM method.

Table 2
Monte Carlo results: size and power of two-sample tests[b]

			Log-rank		Wilcoxon	
Group 1	Group 2	Method	Method 1 (%)	Method 2 (%)	Method 1 (%)	Method 2 (%)
exp(0.5)	exp(0.5)	STANDARD	5.6	5.6	5.3	5.3
$n = 25$	$n = 25$	KMI	8.0	8.0	5.4	5.4
		RSI	7.0	7.0	5.1	5.0
		KMIB	7.4	6.9	5.6	5.6
		RSIB	6.8	6.8	5.5	5.4
exp(0.5)	exp(0.5)	STANDARD	4.4	4.4	5.2	5.2
$n = 100$	$n = 100$	KMI	6.1	6.1	5.4	5.4
		RSI	5.8	5.8	5.3	5.3
		KMIB	5.5	5.3	4.8	4.8
		RSIB	5.2	5.1	5.3	5.3
exp(0.70)	exp(0.45)	STANDARD	26.1	26.1	22.4	22.4
$n = 25$	$n = 25$	KMI	30.4	30.2	23.0	22.9
		RSI	29.7	29.3	23.5	23.4
		KMIB	29.2	28.9	22.9	22.8
		RSIB	28.4	28.3	22.9	22.9
exp(0.70)	exp(0.45)	STANDARD	78.8	78.8	68.4	68.4
$n = 100$	$n = 100$	KMI	78.7	78.6	72.0	72.0
		RSI	79.2	79.1	71.1	71.1
		KMIB	76.3	76.2	70.8	70.7
		RSIB	77.7	77.4	70.3	70.3

[b]Event and censoring times are generated from exponential distributions, independently in the two groups. STANDARD: standard log-rank or Wilcoxon procedure applied to censored data, KMI: Kaplan–Meier-based imputation, RSI: risk set imputation, KMIB: Kaplan–Meier-based imputation using bootstrap, and RSIB: risk set imputation using bootstrap. Method 1 is based on the ratio of the averages of the numerator and denominator in the log-rank and Wilcoxon procedures, Method 2 is based on the average of the standardized test statistics. The mean of the censoring distribution was chosen to give censoring rates of $\sim 20\%$. $n = 25$ or 100 in each group. Results based on 1000 replications and $M = 10$.

However, these two multiple-imputation method's SEs are smaller than the SEs from Greenwood's formula applied to the original data set, especially at the higher censoring rate in the tails of the curve, resulting in lower coverage rates than the KM method. The inclusion of the bootstrap stage in the multiple procedure corrects this problem, giving coverage rates and standard errors very similar to those of Greenwood's formula on the original data.

5.2. Comparison of two sample tests

For the evaluation of two sample tests the event times were generated from separate exponential distributions and the various multiple-imputation approaches were applied separately within each group. Results in Table 2 show that the Wilcoxon test type 1 error rate based on the multiple-imputation procedures is close to the method applied to the unimputed data (denoted by

STANDARD). Imputation-based logrank tests have slightly higher nominal values than the unimputed logrank test, where this inflation in type 1 error diminishes with increasing sample size. The inclusion of the bootstrap stage slightly improves the significance level of the tests. Larger values of M (results not shown) had negligible effect on the rejection rates in Table 2. There is only a very slight difference between methods 1 and 2 of analyzing the multiply imputed datasets. The results suggest that the approach based on analyzing the standardized statistics (method 2) is slightly better than the approach based on averages of numerator and denominator over multiply imputed datasets (method 1) at giving significance levels closer to the nominal level. Power results for the four imputation-based methods are also given in Table 2, with largely similar results to unimputed test procedures.

6. Discussion

The research in this paper provides a connection between standard survival analysis methods, the RRA and multiple imputation. The fact that the KM estimator can be reproduced using multiple imputation of future event times provides a basis for the use of imputation to handle missing event times due to censoring in survival analysis. Extensions and generalizations of the one sample and two sample procedures described in this paper to more complex situations are possible through the flexibility of the multiple-imputation approach. For example, the imputations at each risk set could depend on time-independent or time-dependent covariates, or they could be based on more parametric models than we have used here. In these types of situations multiple imputation can be used to correct for bias and improve efficiency while avoiding additional development of highly specialized statistical analysis tools. An example of using multiple imputation in a more complex setting is the application described in Schenker and Taylor (1996) where event times are multiply imputed using a parametric model derived from a different data set.

There are a number of different general approaches to handling missing data, one is through weighting analogous to the Horowitz–Thompson estimator in the sampling literature, another is through multiple imputation as used here, a third is through full parametric modeling. One attractive feature of multiple imputation compared to full parametric modeling is its robustness and ease of generalizability. Any model which is used to generate the imputes is based on the observed data, once the imputes have been made then the imputation model is discarded and any final analysis is based on the original data enhanced by the imputed values. The multiple-imputation approach has the added advantage of having a standard way to obtain measures of uncertainty, which will be particularly useful in complex settings. The data analyst is now free to choose and perform, with little effort, an analysis appropriate for the goals of their study. Conditions for the appropriateness of this philosophy are discussed in Meng (1994).

The multiple-imputation schemes share some common ideas with the RRA. Both approaches essentially move censored observations to longer times, the RRA does this by reallocating weights to longer times, whereas multiple imputation achieves this by explicitly creating a longer event time.

In the Monte Carlo study we found that including the bootstrap stage in the multiple-imputation procedure improved the properties of both the estimation and testing methods, particularly in the situation of a high degree of censoring. Other than computational burden there seems no reason not to use a bootstrap.

The adequacy of imputation procedures will depend on the availability of possible donor observations, which diminishes in the tails of the survival distribution. As we have shown in Section 4, the point estimates will still be good for large M, even with a small number of possible donors. However, the adequacy of the standard errors from the multiple-imputation procedures will be questioned for imputation of censored times in the tails, where there may be only a few possible donors or even no possible donors in the risk set. We can see some suggestion of this problem in Table 1 for the estimation at the 35th percentile. For this case the multiple-imputation procedures without the bootstrap have underestimated standard errors. The two sample test results are also affected by this phenomenon. The logrank test assigns more weight to the later times points than the Wilcoxon test. We note that the multiple-imputation test gives a rejection rate of the logrank test that was slightly higher than the level from the standard procedure applied to the censored data. In contrast, the Wilcoxon test has a rejection rate close to the nominal level.

The theoretical result in Section 4.2 concerning the equivalence of the multiple-imputation estimator and the KM estimator suggests that a large number of imputes are preferable. We used $M = 10$ in the simulations, which is a little larger than the usual $M = 3$ or 5 which have been used in survey applications. Our empirical experiences also reiterate the need for a large number of imputes; we found that the statistical properties of the estimators deteriorated if a small number of imputes was used. Although the coverage rates of confidence intervals was largely unaffected by the choice of M, we found that the efficiency of the estimates, as measured by the standard deviation in Table 1 was about 3–5% worse for $M = 3$ compared to $M = 10$ or higher M.

7. Uncited References

Little and Rubin, 1987; Rubin, 1976

Appendix A.

A proof by induction gives $E\{\hat{S}_{\mathrm{RSI}}(t_j)|\mathbf{Y}\} = \hat{S}_{\mathrm{KM}}(t_j), j = 1,\ldots,n$. First when $j = 1$,

$$E\{\hat{S}_{\mathrm{RSI}}(t_1)|\mathbf{Y}\} = \left(1 - \frac{1}{n_1}\right)^{\delta_1} = \hat{S}_{\mathrm{KM}}(t_1),$$

since if the first event time is censored a later event time will be imputed and $\delta_1 = 0$ will reflect the correct result at time t_1 and if $\delta_1 = 1$ no imputation takes place. For the inductive step assume that $E\{\hat{S}_{\mathrm{RSI}}(t_j)|\mathbf{Y}\} = \hat{S}_{\mathrm{KM}}(t_j)$. We need to show the result will also hold for $j+1$. We rewrite $\hat{S}_{\mathrm{RSI}}(t_{j+1})$ as

$$\hat{S}_{\mathrm{RSI}}(t_{j+1}) = \prod_{t_i \leqslant t_{j+1}} \left(1 - \frac{1 + \Gamma^{\mathrm{RSI}}_{2,i}}{n^{\mathrm{RSI}}_i}\right)^{\delta_i} = \hat{S}_{\mathrm{RSI}}(t_j)\left(1 - \frac{1 + \Gamma^{\mathrm{RSI}}_{2,j+1}}{n^{\mathrm{RSI}}_{j+1}}\right)^{\delta_{j+1}}.$$

It is convenient to separate the proof of the inductive step into two cases depending on the value of δ_{j+1}. In the first case, $\delta_{j+1} = 0$. Any earlier censored time which was replaced by (t_{j+1}, δ_{j+1}), which would still be censored. Also imputation would replace t_{j+1} with a larger event time so that

$\hat{S}_{\mathrm{RSI}}(t_{j+1}) = \hat{S}_{\mathrm{RSI}}(t_j) * 1$. Hence

$$E\{\hat{S}_{\mathrm{RSI}}(t_{j+1})|\mathbf{Y}\} = E\{\hat{S}_{\mathrm{RSI}}(t_j)|\mathbf{Y}\} = \hat{S}_{\mathrm{KM}}(t_j) = \hat{S}_{\mathrm{KM}}(t_{j+1})$$

so that the inductive step holds in this case.

In the second case, $\delta_{j+1} = 1$. Let $\mathbf{Y}^*_{t^-_{j+1}}$ denote the filtration relating to the imputed survival and censoring information prior to time t^-_{j+1}, i.e. $\mathbf{Y}^*_{t^-_{j+1}}$ is the set of new imputed observations less than t_{j+1} in the imputed data set. Then

$$E\{\hat{S}_{\mathrm{RSI}}(t_{j+1})|\mathbf{Y}\} = E\left\{\hat{S}_{\mathrm{RSI}}(t_j)\left(1 - \frac{1 + \Gamma^{\mathrm{RSI}}_{2,j+1}}{n^{\mathrm{RSI}}_{j+1}}\right)\middle|\mathbf{Y}\right\}$$

$$= E\left[E\left\{\hat{S}_{\mathrm{RSI}}(t_j)\left(1 - \frac{1 + \Gamma^{\mathrm{RSI}}_{2,j+1}}{n^{\mathrm{RSI}}_{j+1}}\right)\middle|\mathbf{Y}, \mathbf{Y}^*_{t^-_{j+1}}\right\}\middle|\mathbf{Y}\right].$$

In this expression the outer expectation is with respect to the distribution of $\mathbf{Y}^*_{t^-_{j+1}}$ conditional on $\mathbf{Y}$. Conditional on $\mathbf{Y}$ and $\mathbf{Y}^*_{t^-_{j+1}}, n^{\mathrm{RSI}}_1, \ldots, n^{\mathrm{RSI}}_{j+1}, \Gamma^{\mathrm{RSI}}_{1,1}, \ldots, \Gamma^{\mathrm{RSI}}_{1,j+1}$, and $\Gamma^{\mathrm{RSI}}_{i,1}, \ldots, \Gamma^{\mathrm{RSI}}_{i,j}$, for $i = 2, 3$ are nonrandom, so that $\hat{S}_{\mathrm{RSI}}(t_j)$ is also nonrandom. So the above becomes

$$E\left[\hat{S}_{\mathrm{RSI}}(t_j)\left\{1 - \frac{1 + E(\Gamma^{\mathrm{RSI}}_{2,j+1}|\mathbf{Y}, \mathbf{Y}^*_{t^-_{j+1}})}{n^{\mathrm{RSI}}_{j+1}}\right\}\middle|\mathbf{Y}\right].$$

But $\Gamma^{\mathrm{RSI}}_{2,j+1}$ given $\mathbf{Y}$ and $\mathbf{Y}^*_{t^-_{j+1}}$ is binomial $(\Gamma^{\mathrm{RSI}}_{3,j}, 1/n_{j+1})$ and recall n^{RSI}_{j+1} can be rewritten as $n_{j+1} + \Gamma^{\mathrm{RSI}}_{3,j}$. So the above becomes

$$E\left[\hat{S}_{\mathrm{RSI}}(t_j)\left\{1 - \frac{1 + \Gamma^{\mathrm{RSI}}_{3,j}/n_{j+1}}{n_{j+1} + \Gamma^{\mathrm{RSI}}_{3,j}}\right\}\middle|\mathbf{Y}\right] = E\left\{\hat{S}_{\mathrm{RSI}}(t_j)\left(1 - \frac{1}{n_{j+1}}\right)\middle|\mathbf{Y}\right\}$$

$$= E\{\hat{S}_{\mathrm{RSI}}(t_j)|\mathbf{Y}\}\left(1 - \frac{1}{n_{j+1}}\right)$$

$$= \hat{S}_{\mathrm{KM}}(t_j)\left(1 - \frac{1}{n_{j+1}}\right) = \hat{S}_{\mathrm{KM}}(t_{j+1}).$$

So the inductive step holds in this case as well. Hence, $E\{\hat{S}_{\mathrm{RSI}}(t_j)|\mathbf{Y}\} = \hat{S}_{\mathrm{KM}}(t_j), j = 1, \ldots, n$. The proof for the KMI-based estimate is similar.

References

Efron, B., 1967. The two sample problem with censored data. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. IV, University of California Press, Berkeley CA, pp. 831–853.

Heitjan, D.F., 1994. Ignorability in general incomplete-data models. Biometrika 81, 701–710.

Heitjan, D.F., Little, R.J.A., 1991. Multiple imputation for the fatal accident reporting system. Appl. Statist. 40, 13–29.

Li, K.H., Raghunathan, T.E., Rubin, D.B., 1991. Large-sample significance levels from multiply imputed data using moment-based statistics and an F reference distribution. J. Amer. Statist. Assoc. 86, 1065–1073.

Little, R.J.A., Rubin, D.B., 1987. Statistical Analysis with Missing Data. Wiley, New York.

Meng, X.L., 1994. Multiple-imputation inferences with uncongenial sources of input (with discussion). Statist. Sci. 9, 538–573.

Miller, R.G., 1981. Survival Analysis. Wiley, New York.

Rubin, D.B., 1976. Inference and missing data. Biometrika 63, 581–592.

Rubin, D.B., 1978. Multiple imputations in sample surveys—A phenomenological Bayesian approach to nonresponse. Proceedings of the Survey Research Methods Section; J. Amer. Statist. Assoc. 20–34.

Rubin, D.B., 1987. Multiple Imputation for Nonresponse in Surveys. Wiley, New York.

Rubin, D.B., Schenker, N., 1991. Multiple imputations in health-care database: an overview and some applications. Statist. Medicine 10, 585–598.

Schenker, N., Taylor, J.M.G., 1996. Partially parametric techniques for multiple imputation. Comput. Statist. Data Anal. 22, 425–446.