Journal of Multivariate Analysis 102 (2011) 105–117

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Multiple imputations and the missing censoring indicator model

Sundarraman Subramanian

Center for Applied Mathematics and Statistics, Department of Mathematical Sciences, New Jersey Institute of Technology, USA

ARTICLE INFO

Article history:
Received 20 August 2009
Available online 20 August 2010

AMS subject classifications:
62N01
62N02
62N03

Keywords:
Asymptotic normality
Functional delta method
Lindeberg's condition
Maximum likelihood
Missing at random
Model-based resampling

ABSTRACT

Semiparametric random censorship (SRC) models (Dikta, 1998) provide an attractive framework for estimating survival functions when censoring indicators are fully or partially available. When there are missing censoring indicators (MCIs), the SRC approach employs a model-based estimate of the conditional expectation of the censoring indicator given the observed time, where the model parameters are estimated using only the complete cases. The multiple imputations approach, on the other hand, utilizes this model-based estimate to impute the missing censoring indicators and form several completed data sets. The Kaplan–Meier and SRC estimators based on the several completed data sets are averaged to arrive at the multiple imputations Kaplan–Meier (MIKM) and the multiple imputations SRC (MISRC) estimators. While the MIKM estimator is asymptotically as efficient as or less efficient than the standard SRC-based estimator that involves no imputations, here we investigate the performance of the MISRC estimator and prove that it attains the benchmark variance set by the SRC-based estimator. We also present numerical results comparing the performances of the estimators under several misspecified models for the above mentioned conditional expectation.

1. Introduction

There are two important approaches for estimating survival functions from right censored data. The nonparametric and most popular approach leads to the Kaplan–Meier (KM) or product limit estimator, which has several appealing properties such as asymptotic efficiency [27]. An alternative approach is based on semiparametric random censorship (SRC) models [3] and leads to an estimator of the survival function with asymptotic variance not greater than that of the KM estimator, and potentially even smaller. The efficacy of the SRC approach, however, is rooted in the basic premise that the correct model be specified for the conditional expectation of the censoring indicator given the observed, possibly censored, event time—since, otherwise, the estimator would be inconsistent. When the censoring indicators are always available, therefore, the choice between the two approaches may present an intriguing dilemma as it represents a fundamental trade-off between semiparametric efficiency and nonparametric "robustness"—the KM estimator is consistent, if less efficient than the possibly inconsistent SRC estimator. When there are MCIs, however, the KM estimator is inapplicable, and the "robustness" advantage of nonparametric approaches is perhaps neutralized by the need for smoothing, requiring the specification of data-based optimal bandwidths for computing the estimator [24,17,20,22]. Apart from the effort needed to choose a suitable model, the SRC approach has no such frailties, which may well be a significant advantage when there are MCIs [19].

The approach of multiple imputations is useful when there are missing data [10,11,26,8,23,13,21]. In this approach, the missing components are filled in with imputed values and parameter estimates are obtained from the completed data set, treating the imputed values as though they were actually observed. Estimates from multiple completed data sets are combined in some natural way, such as averaging, to further improve their precision. Kim [7] investigated the finite sample properties of multiple imputations estimators while Schenker and Welsh [12] derived asymptotic results.

E-mail address: sundars@njit.edu.

0047-259X/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2010.08.005

In this article, we focus on multiple-imputations-based estimation of a survival function from right censored data with MCIs. For right censorship without MCIs, the observed random variables are X and δ, where $X = \min(T, C)$, $\delta = I(T \leq C)$ is the censoring indicator, T is the lifetime of interest, and C is an independent censoring variable. Dikta [3] introduced SRC models, by proposing model-based estimation of the conditional expectation $E(\delta|X = t) = p(\delta = 1|X = t) = p(t)$ and proved that, when the model for $p(t)$ was correctly specified, the SRC estimator of $S(t)$, the survival function of T, was as efficient as or more efficient than the KM estimator. The data for the MCI model of random censorship are $\{(X_i, \xi_i, \sigma_i)_{1\leq i\leq n}\}$, where $\xi_i = 1$ when δ_i is observed and is 0 otherwise, and $\sigma_i = \xi_i\delta_i$. Subramanian [19] proved that the SRC estimator for the MCI model, denoted by $\hat{S}_D(t)$, was as efficient as or more efficient than nonparametric estimators. Subramanian [21] investigated a multiple-imputations-based KM estimator (referred to henceforth as the MIKM estimator), defined as the average of many single imputation KM estimators, and proved that the MIKM estimator was asymptotically *less efficient* than $\hat{S}_D(t)$. Naturally, the question arises as to whether there are alternative multiple imputations estimators which are better than the MIKM estimator, and whether they would attain the existing benchmark variance set by the estimator $\hat{S}_D(t)$. We address this issue by proposing the multiple-imputations-based SRC estimator, called the MISRC estimator, and derive its asymptotic distribution.

Note that $\hat{S}_D(t)$ is computed without recourse to any imputations. To obtain the model-based estimate of $p(t)$ used for computing $\hat{S}_D(t)$, we choose a suitable good-fitting model $p(t, \theta)$ (from candidates such as logit, probit, generalized proportional hazards, among others; see [3]) and estimate the model parameter $\theta \in \mathbb{R}^k$ by using maximum likelihood based on only the *complete* cases. We denote the maximum likelihood estimator (MLE) by $\hat{\theta}_D$. Estimating θ in this way still produces a consistent estimate under the assumption that the MCIs are *missing at random* (MAR; see [8,23], or Subramanian [19]). Note that MAR implies that $P(\xi = 1|X = t, \delta = d) = P(\xi = 1|X = t) = \pi(t)$ (Rubin [9]), and also means that, conditional on X, the missingness and censoring indicators are independent: $P(\sigma = 1|X = t) = \pi(t)p(t)$. The multiple imputations approach involves using the estimated conditional probability $p(t, \hat{\theta}_D)$ to impute missing δ, to form $M \geq 1$ completed data sets, and then computing the SRC estimator, denoted by $\hat{S}^{(m)}(t)$. The average of the M single imputation SRC estimates $\hat{S}^{(m)}(t)$, $m = 1, \ldots, M$, provides the MISRC estimator, to be denoted henceforth by $\hat{S}(t)$. Lu and Tsiatis [8], and Tsiatis et al. [23] implemented this method for competing risks with covariates and missing cause of failure information. We prove that when the model for $p(t)$ is specified correctly, the MISRC estimator $\hat{S}(t)$ is asymptotically equivalent to the SRC estimator $\hat{S}_D(t)$ and hence asymptotically as efficient as or more efficient than the MIKM estimator. We also carried out several numerical studies to compare the performance of the estimators when $p(t)$ was misspecified. The MIKM was more robust to misspecification.

Significantly, the multiple imputations procedure has connections with the *model-based resampling* introduced by Dikta et al. [4] for model checking in the context of binary data. Dikta et al. [4] prescribe the following recipe for standard right censored data (that is, when there are no MCIs): Resample *all* the censoring indicators, on the basis of the estimated model $\hat{p}_D(t) = p(t, \hat{\theta}_D)$. Dikta and Winkler [5] implemented the extension to MCI data, resampling *only* the *non-missing* censoring indicators. *In contrast, the model-based resampling implicit in our multiple imputations procedure entails resampling (imputing) only the MCIs*. We do *not* resample (impute) the non-missing censoring indicators.

The rest of the article is organized as follows. In Section 2, we derive the asymptotic distribution of the MISRC estimator. In Section 3, we present several numerical results comparing the SRC, MIKM, and MISRC estimators. Section 4 focuses on some discussion and conclusions. Technical complements are included in an Appendix.

2. Multiple-imputations estimation

Some of the notation below is from Dikta [3]. Specify a parametric model for $p(t)$ through $p(t) = p(t, \theta)$, where $p(\cdot)$ is known up to the k-dimensional parameter θ. Define

$$q(t, \theta) = \log p(t, \theta), \qquad \bar{q}(t, \theta) = \log(1 - p(t, \theta)).$$

Let θ_0 denote the true value of θ and define $p_0(t) = p(t, \theta_0)$, $q_0(t) = q(t, \theta_0)$, and $\bar{q}_0(t) = \log(1 - p_0(t))$. Note that $q_0(t) = \log p_0(t)$. Write $D_r(p(t, \theta))$ for the partial derivative of $p(t, \theta)$ with respect to θ_r; when it is evaluated at $\theta = \theta^*$, denote it by $D_r(p(t, \theta^*))$. Write $\mathrm{Grad}(p(t, \theta)) = [D_1(p(t, \theta)), \ldots, D_k(p(t, \theta))]^T$ and $C_\theta(t) = \mathrm{Grad}(p(t, \theta))\,(\mathrm{Grad}(p(t, \theta)))^T$. When $\theta = \theta_0$, we denote the matrix $C_{\theta_0}(t)$ by $C_0(t)$. Define the information matrices

$$I(\theta_0) \doteq I_0 = E\left(\frac{C_0(X)}{p_0(X)(1 - p_0(X))}\right), \qquad J(\theta_0) \doteq J_0 = E\left(\frac{\pi(X)C_0(X)}{p_0(X)(1 - p_0(X))}\right).$$

Note that the (r, s) elements of I_0 (the case of no MCIs) and J_0 (the case with MCIs) are given by

$$i_{r,s} = E\left(\frac{D_r(p_0(X))D_s(p_0(X))}{p_0(X)(1 - p_0(X))}\right), \qquad j_{r,s} = E\left(\frac{\pi(X)D_r(p_0(X))D_s(p_0(X))}{p_0(X)(1 - p_0(X))}\right). \tag{1}$$

Also, write $\alpha(u, v) = (\mathrm{Grad}(p_0(u)))^T J_0^{-1}\mathrm{Grad}(p_0(v))$. We denote the second-order partial derivatives by $D_{r,s}(\cdot)$. We will need the following assumptions (cf. Dikta et al. [4]):

(A1) The functions $D_{r,s}(q(t, \theta))$ and $D_{r,s}(\bar{q}(t, \theta))$ are continuous with respect to θ at each $\theta \in \mathcal{D} \subset \mathbb{R}^k$ and $t \in \mathbb{R}$. Also, the functions $D_r(q(\cdot, \theta))$, $D_r(\bar{q}(\cdot, \theta))$, $D_{r,s}(q(\cdot, \theta))$ and $D_{r,s}(\bar{q}(\cdot, \theta))$ are measurable for each $\theta \in \mathcal{D}$, and there exists a