Journal of Multivariate Analysis 124 (2014) 10–24

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Asymptotically efficient estimation under semi-parametric random censorship models

Gerhard Dikta

Fachhochschule Aachen, Heinrich-Mußmann Strasse 1, D-52428 Jülich, Germany

ARTICLE INFO

Article history:
Received 7 May 2013
Available online 29 October 2013

AMS 2010 subject classifications:
62n01
62n02
62E20

Keywords:
Survival analysis
Censored data
Semi-parametric random censorship model
Asymptotic efficiency

ABSTRACT

We study the estimation of some linear functionals which are based on an unknown lifetime distribution. The observations are assumed to be generated under the semi-parametric random censorship model (SRCM), that is, a random censorship model where the conditional expectation of the censoring indicator given the observation belongs to a parametric family. Under this setup a semi-parametric estimator of the survival function was introduced by the author. If the parametric model assumption is correct, it is known that the estimated functional which is based on this semi-parametric estimator is asymptotically at least as efficient as the corresponding one which rests on the nonparametric Kaplan–Meier estimator.

In this paper we show that the estimated functional which is based on this semi-parametric estimator is asymptotically efficient with respect to the class of all regular estimators under this semi-parametric model.

1. Introduction

To analyze failure-time or lifetime data, one often has to handle incomplete observations which are caused by some type of censoring. One type of censoring, which is widely accepted in practice, is described by the random censorship model (RCM). Under this model one has two independent sequences of independent and identically distributed (IID) random variables: the survival times $X_1, \ldots, X_n$ and the censoring times $Y_1, \ldots, Y_n$. These sequences define the observations: (Z_1, δ_1), $\ldots, (Z_n, \delta_n)$, where $Z_i = \min(X_i, Y_i)$ and δ_i indicates whether the observation time Z_i is a survival time ($\delta_i = 1$) or a censoring time ($\delta_i = 0$). We assume that all these sequences are defined over some probability space $(\Omega, \mathcal{A}, \mathbb{P})$.

Denote the distribution functions (DF) of X, Y, and Z by F, G, and H, respectively, and assume that they are continuous. Nonparametric statistical inference of F based on this type of observations is usually based on the time-honored Kaplan–Meier (KM) or product limit estimator, see Kaplan and Meier [12], defined by

$$F_n^{KM}(t) = 1 - \prod_{i:\, Z_i \le t} \left(1 - \frac{\delta_i}{n - R_{i,n} + 1}\right),$$

where $R_{i,n}$ denotes the rank of Z_i among the Z-sample.

The KM-estimator can be derived as a nonparametric maximum likelihood estimator (MLE), see Johansen [11]. Wellner [24] proved that it also retains some asymptotic optimality properties one usually expects for a MLE to have in a parametric setup. Among other things, he obtained a functional Hájek–Le Cam like convolution result, see Theorem 1 in Wellner [24], to find the optimal centered Gaussian process of any regular estimator $\hat{F}_n$ of F. Since the limiting process of $\sqrt{n}(F_n^{KM}(t) - F(t))$, derived by Breslow and Crowley [4], has the same covariance structure as the optimal centered Gaussian process, the

E-mail address: dikta@fh-aachen.de.

0047-259X/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jmva.2013.10.002

KM-estimator is asymptotically optimal with respect to all regular estimators of F under SRCM. It may be noted, that the processes considered in [24,4] are restricted to the interval $[0, \tau]$ such that $\tau < \tau_H$, where $\tau_H = \inf\{t : H(t) = 1\}$.

The analysis of lifetime data is often focused on the estimation of some characteristic parameters of the underlying distribution which can be represented by an integral of a Borel-measurable function φ with respect to F, that is, by $\int_0^\infty \varphi dF$. The DF F at a fixed point t itself, the expectation, the variance, or the mean residual lifetime are some typical examples, see Section 1 in Stute and Wang [16].

Susarla and Van Ryzin [13] started the analysis of KM-integrals to estimate $\int \varphi dF$, where φ can be of unbounded variation. In particular, they studied the special case of the identity function, $\varphi(x) = x$, and obtained the asymptotic normality of the KM-integral $\int x 1\,(0 \le x \le M_n)\, F_n^{KM}(dx)$, where $M_n \to \infty$ and $1\,(x \in A)$ denotes the indicator function of the set A. Among other things, Gill [10] studied the weak convergence of KM-processes on the interval $[0, \tau_H]$ and KM-integrals for nonnegative, continuous, and non-increasing φ. If φ is a monotone function, Schick, Susarla, and Koul [14] generalized Gill's weak convergence result. They provided an asymptotically linear representation of $\sqrt{n}(\int \varphi dF_n^{KM} - \int \varphi dF)$. Furthermore, they deduced from a Hájek–Le Cam like convolution result for regular estimators of $\int \varphi dF$ under RCM the asymptotic efficiency of $\int \varphi dF_n^{KM}$, see Theorem 1.1 and Theorem 1.2 of Schick et al. [14]. For arbitrary Borel-measurable φ, Stute [15] derived, under quite general assumptions, an asymptotic linear representation of $\int \varphi dF_n^{KM}$ which implies, according to the central limit theorem (CLT), the weak convergence of

$$\sqrt{n}\Big(\int \varphi(x)\, F_n^{KM}(dx) - \int_0^{\tau_H} \varphi(x)\, F(dx)\Big)$$

to a centered normal variable with variance $\sigma^2_{KM}(\varphi) \equiv \sigma^2_{KM}$.

The KM-estimator is the first choice under the RCM. Let $Z_{1:n}, \ldots, Z_{n:n}$ denote the ordered observations. F_n^{KM} attaches mass only to the uncensored observations and the amount of attached mass increases from the smallest to the largest uncensored observation. Furthermore, the increase depends on the number of censored observations between two uncensored ones, see Efron [9]. Therefore, if we have to analyze a heavily censored dataset, the KM-estimator will have only a few jumps which might lead to a rather sketchy result. In such a situation we can try to improve our data analysis by tightening the general model assumption of RCM and by using a different type of estimator which is better adapted to the restricted RCM model than the KM-estimator. For example, if it is guaranteed that F belongs to a parametric family, a maximum likelihood approach will be used to find the corresponding parametric estimator of F. Or, if a parametric model is appropriate but the data might be contaminated, the minimum Hellinger distance estimation (MHDE) of the parameter is suitable to get an estimator close to F which is a member of the parametric family. When there is no contamination, the MHDE is known to be asymptotically efficient among the class of regular estimators, see Yang [25,26].

If we have good reasons to assume that the censoring DF G is linked to the DF of the survival times F, the semi-parametric random censorship model (SRCM) is a possible restriction of RCM. Following Dikta [5,6], we assume under SRCM, additionally to RCM, that the conditional expectation of the indicator δ given the observation time $Z = z$,

$$m(z) = \mathbb{E}\big(\delta \,\big|\, Z = z\big) = \mathbb{P}\big(\delta = 1 \big| Z = z\big),$$

belongs to a parametric family

$$m(z) = m(z, \theta_0),$$

where $\theta_0 = (\theta_{0,1}, \ldots, \theta_{0,k}) \in \Theta \subset \mathbb{R}^k$. Together with SRCM, the semi-parametric (SE) estimator

$$F_n^{SE}(t) = 1 - \prod_{i:\, Z_i \le t} \Big(1 - \frac{m(Z_i, \hat{\theta}_n)}{n - R_{i,n} + 1}\Big)$$

of F was proposed in Dikta [5,6], where $\hat{\theta}_n$ is the MLE of θ_0, that is, the maximizer of the (partial) likelihood function

$$L_n(\theta) = \prod_{i=1}^{n} m(Z_i, \theta)^{\delta_i} (1 - m(Z_i, \theta))^{1-\delta_i}.$$

A discussion of possible parametric models for m can be found in these two papers. Since the parametric model of m is based on a parametric binary model, we can test the validity of this assumption by some goodness-of-fit tests. A general bootstrap based approach for such tests is given in Dikta, Kvesic, and Schmidt [8]. Note that F_n^{SE} attaches mass to every observation time Z_i and not only to the uncensored ones, like F_n^{KM} does.

The semi-parametric approach together with the corresponding estimator F_n^{SE} has been shown to be applicable and flexible enough for extensions to other statistical scenarios. For example, Subramanian [17] introduced an extension to the missing censoring-indicator model. He proved that the extended semi-parametric estimator is at least as good as an efficient nonparametric one under the missing censoring-indicator model if the parametric model assumption is correct. Sun and Zhu [20] applied this approach to truncated and censored data and Subramanian [18] for multiple imputations in the missing censoring-indicator model. Recently, Subramanian and Zhang [19] introduced a two-stage bootstrap procedure to construct confidence bands for the survival function under the semi-parametric model with and without missing censoring-indicators.