II PhDay-EIO
Santiago de Compostela, 23 de xuño do 2021

Bootstrap bandwidth selector for Beran's estimator of the conditional survival function

Rebeca Peláez[1], Ricardo Cao[1], Juan M. Vilar[1]

[1]Department of Mathematics. Universidade da Coruña.

ABSTRACT

Let $\{(X_i, Z_i, \delta_i)\}_{i=1}^n$ be a simple random sample of (X, Z, δ) with X being the covariate, $Z = \min\{T, C\}$ the observed variable and $\delta = I_{T \leq C}$ the uncensoring indicator. Usually, T is the time until the occurrence of an event and C is the censoring time. The survival function of T is denoted by $S(t)$ and $S(t|x)$ is the conditional survival function of T given $X = x$ evaluated at t. The conditional survival function estimator proposed by Beran (1981) is given by

$$\widehat{S}_h^B(t|x) \quad = \quad \prod_{i=1}^{n} \left(1 - \frac{I_{\{Z_i \leq t,\ \delta_i = 1\}} w_{n,i}(x)}{1 - \sum_{j=1}^{n} I_{\{Z_j < Z_i\}} w_{n,j}(x)} \right) \tag{1}$$

where

$$w_{n,i}(x) = \frac{K\big((x - X_i)/h\big)}{\sum_{j=1}^{n} K\big((x - X_j)/h\big)}$$

with $i = 1, ..., n$ and $h = h_n$ is the bandwidth for the covariable.

First, finding a method for automatic selection of the smoothing parameter h is interesting. Secondly, the issue of confidence intervals of $S(t|x)$ by means of Beran's estimator is addressed. Bootstrap has become a strong instrument in many statistical applications since it was first introduced by Efron (1979). It is a suitable technique in this context.

In this work, a resampling technique to approximate the smoothing parameters involved in Beran's estimator is defined. Our approach is based on resampling by the smoothed bootstrap and minimising the bootstrap approximation of the mean integrated squared error to find the bootstrap bandwidth. Given r an appropiate pilot bandwidth, the bootstrap resampling algorithm consists of generating $U_i \sim U(0, 1)$ and $V_i \sim K$ and obtaining

$$X_i^* = X_{[nU_i]+1} + rV_i,$$

$$Z_i^* = Z_{[nU_i]+1},$$

$$\delta_i^* = \delta_{[nU_i]+1},$$

for each $i = 1, \ldots, n$. The bootstrap sample is formed as $\{(X_i^*, Z_i^*, \delta_i^*)\}_{i=1}^n$.

The optimal smoothing parameter is the bandwidth that minimizes the mean integrated squared error given by:

$$MISE_x(h) = E\bigg(\int \big(\widehat{S}_h(t|x) - S(t|x)\big)^2 dt \bigg).$$

Then, the bootstrap bandwidth is obtained by minimizing the Monte Carlo approximation of the bootstrap MISE defined as follows

$$MISE^*_x(h) \quad \simeq \quad \frac{1}{B}\sum_{j=1}^{B}\Big(\int \big(\widehat{S}^{*(j)}_h(t|x) - \widehat{S}_r(t|x)\big)^2 dt\Big),$$

where $\widehat{S}_r(t|x)$ is the Beran's survival estimation with pilot bandwidth r using the original sample $\{(X_i, Z_i, \delta_i)\}_{i=1}^n$, $\widehat{S}^{*(j)}_h(t|x)$ is the Beran's survival estimation with bandwidth h using the bootstrap resample $\{(X_i^{*(j)}, Z_i^{*(j)}, \delta_i^{*(j)})\}_{i=1}^n$, and B the number of bootstrap resamples.

A simulation study is carried out to analyse the behaviour of the bootstrap algorithm previously described. Several models with different conditional probabilities of censoring were considered. Figure 1 shows the MISE bootstrap functions in two of these scenarios: Model 1 that considers Weibull distribution for life and censoring times and Model 2 that considers exponential life and censoring times. Both models have a conditional probability of censoring equal to 0.5.

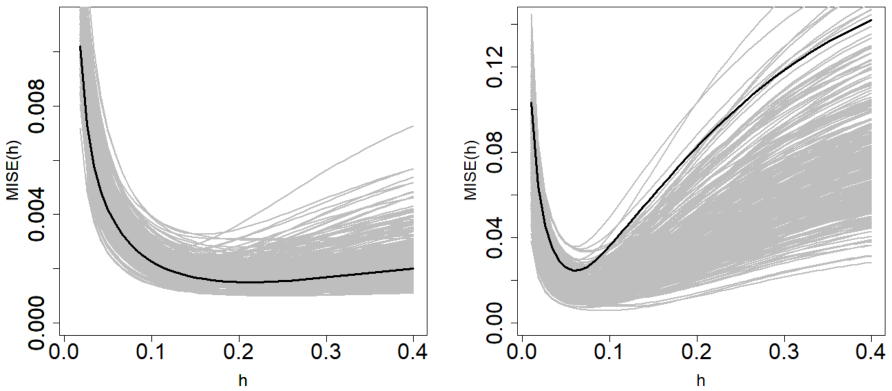

Figure 1: $MISE_x(h)$ function approximated via Monte Carlo using $N = 300$ samples and $MISE^*_x(h)$ approximated via bootstrap using $B = 500$ resamples in Model 1 (left) and Model 2 (right).

A second simulation study focuses on the calculation of confidence intervals of $S(t|x)$ for fixed values of t and x. The same resampling technique introduced above and the percentile method are used for this purpose.

Given an appropiate smoothing parameter h and fixed values of time, t, and covariate, x, the bootstrap confidence interval for a confidence level of $1-\alpha$ is given by

$$\Big(\widehat{S}_r(t|x) - \frac{\rho_{1-\alpha/2}}{\sqrt{nh}},\ \widehat{S}_r(t|x) - \frac{\rho_{\alpha/2}}{\sqrt{nh}}\Big),$$

where $\widehat{S}_r(t|x)$ is the Beran's estimation with the pilot bandwidth r that is used in the bootstrap resampling, and $\rho_{\alpha/2}$ and $\rho_{1-\alpha/2}$ are the $100\alpha/2$ and $100(1-\alpha/2)$ percentiles of the resampling distribution of $\sqrt{nh}\big(\widehat{S}^*_h(t|x) - \widehat{S}_r(t|x)\big)$, being $\widehat{S}^*_h(t|x)$ the Beran's survival estimation of the bootstrap resample.

Figure 2 shows the theoretical survival function along with the bootstrap estimation and the bootstrap confidence intervals in one sample from Models 1 and 2 with a conditional probability of censoring equal to 0.5.

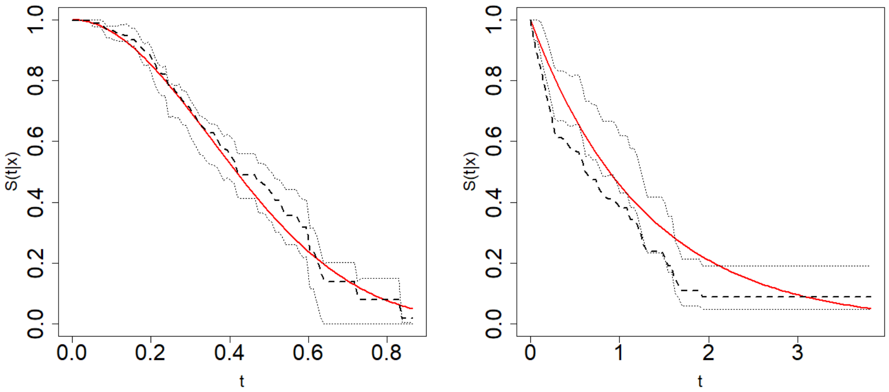

Figure 2: Theoretical survival function (solid line), Beran's estimator with bootstrap bandwidth (dashed line) and the bootstrap confidence intervals (dotted line) for each t in a grid of size $n_t = 100$ in Model 1 (left) and Model 2 (right).

A brief illustration of the use of the bootstrap technique is provided here. A dataset from the Galician Health Service (SERGAS) with times of hospitalisation and age of 2453 COVID-19 patients in Galicia (Spain) is used. The censoring rate of this dataset is 8.8%. The survival function of the time that COVID-19 patients remain hospitalised in ward is estimated by means of Beran's estimator with bootstrap bandwidth for three different ages. Figure 3 shows Beran's survival estimation with bootstrap bandwidth. Only 20% of the 40-year-old patients spend more than 15 days in ward. Meanwhile, 40% of COVID-19 positive patient of 60 or 80 years old spend more than 15 days in ward and only 20% of these patients spend more than 25 days in ward.

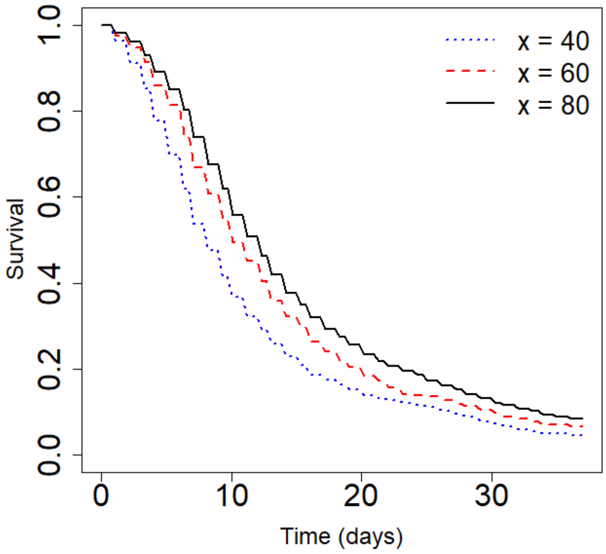

Figure 3: Estimation of $S(t|x)$ for time in ward with Beran's estimator using the optimal bootstrap bandwidth for $x = 40$ (dotted line), $x = 50$ (dashed line) and $x = 80$ (solid line).

The results of the simulations show that this bootstrap algorithm provides adequate smoothing parameters to estimate the survival function in this context. The bootstrap bandwidths obtained are similar to the optimal ones and the estimation errors of both are quite similar.

The work of Peláez et al. (2020) presents a modification of Beran's estimator that involves both a covariate smoothing and a time variable smoothing. We are currently working on an extension of the bootstrap algorithm presented here to select bootstrap bandwidths of the doubly smoothed Beran's estimator, as well as confidence intervals based on it.

References

Beran, R. (1981). Nonparametric regression with randomly censored survival data. *Technical report, University of California.*

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, (7):1–26.

Peláez, R., Cao, R., and Vilar, J. M. (2020). Nonparametric estimation of the conditional survival function with double smoothing. *Under review.*