Statistics and Probability Letters 117 (2016) 151–157

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Bandwidth selection for the presmoothed logrank test

M.A. Jácome [a,*], I. López-de-Ullibarri [b]

[a] Faculty of Science, Universidade da Coruña, 15071, A Coruña, Spain
[b] Escuela Universitaria Politécnica, Universidade da Coruña, 15471, Ferrol, Spain

ARTICLE INFO

Article history:
Received 29 December 2015
Received in revised form 20 May 2016
Accepted 20 May 2016
Available online 2 June 2016

Keywords:
Censoring
Two-sample problem
Bootstrap
Data-driven bandwidth

ABSTRACT

The use of presmoothed logrank-type tests requires that the problem of bandwidth choice be addressed. Two bandwidth selectors for the presmoothed logrank test are proposed and their performance is shown with simulations and the analysis of a dataset.

1. Introduction

The most used of all distribution-free methods for testing the equality of two survival distributions under right random censoring is the logrank (LR) test. It can be shown to have optimality properties for detecting differences in hazard rates assuming that the hazards are proportional. Even if this assumption is reasonable in many settings, situations of patent nonproportionality may still be found. This is the case if e.g. short- and long-term treatment effects have opposite signs. Weighted LR tests have been proposed in order to increase efficiency with nonproportional hazards. However, since in practice it is unknown how the survival of two groups may differ, power optimality is not guaranteed.

Since the LR test and its weighted versions are based on the comparison of the Nelson–Aalen (NA) estimates of the cumulative hazard functions, the power would increase if a more efficient estimator of the cumulative hazard were used. Cao et al. (2005) and Jácome and Cao (2007) have demonstrated the efficiency of the so-called presmoothed NA estimator. While the NA estimator is a step function jumping only at the uncensored observations, the presmoothed NA estimator jumps also at the censored observations. Thus, more information on the local behavior of the lifetime distribution is provided, and estimation accuracy is increased. Jácome and López-de-Ullibarri (2014) introduced a new LR-type test that, instead of the NA estimator, is based on its presmoothed counterpart. The presmoothed LR (PLR) test has the proper size under the null hypothesis, while at the same time improving power over the LR test for a wide range of alternatives, mainly with highly skewed distributions, large censoring rates or moderate sample sizes. The improvement comes at the price of having to choose a bandwidth.

In this paper, we address the issue of automatic bandwidth selection for the PLR test. We adapt two approaches taken from the literature, which are based on different optimality measures. In Section 2, we describe the main characteristics of the PLR test. Two bandwidth selection methods are proposed in Section 3. In Section 4, we present a simulation study investigating the power of the PLR test with the bandwidth selectors. In Section 5, the test is applied to the analysis of a real dataset.

* Corresponding author.
E-mail address: maria.amalia.jacome@udc.es (M.A. Jácome).

http://dx.doi.org/10.1016/j.spl.2016.05.015
0167-7152/

2. The presmoothed logrank test

Suppose we have independent samples of n_1 and n_2 individuals from two populations. For $j = 1, 2$ and $i = 1, 2, \ldots, n_j$ the ith subject in population j has latent lifetime T_{ij}, with survival function S_j and cumulative hazard function Λ_j. In the right censoring model, there exist censoring time variables C_{ij}, independent of the T_{ij}, with distribution function G_j. The data for sample j consist of the pairs $\{(Z_{ij}, \delta_{ij}),\ i = 1, 2, \ldots, n_j\}$, where $Z_{ij} = \min(T_{ij}, C_{ij})$ is the ith observed time, with distribution function H_j, and $\delta_{ij} = \mathbf{1}(T_{ij} \leq C_{ij})$ is the uncensoring indicator.

We want to test the null hypothesis of no difference between survival functions, i.e., $H_0 : S_1 = S_2$. The weighted LR statistic (Fleming and Harrington, 1991) is

$$G^{LR}(t) = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \int_0^t W(v) \frac{Y_1(v)Y_2(v)}{Y_1(v) + Y_2(v)} (d\Lambda_1^{NA}(v) - d\Lambda_2^{NA}(v)), \tag{1}$$

where $W(v)$ is a weight function, $Y_j(v) = \sum_{i=1}^{n_j} \mathbf{1}(Z_{ij} \geq v)$ is the size of the risk set at time v in sample j, and $\Lambda_j^{NA}(v) = \int_0^v Y_j^{-1}(u) dN_j(u)$ is the NA estimate of $\Lambda_j(v)$, with $N_j(u) = \sum_{i=1}^{n_j} \mathbf{1}(Z_{ij} \leq u, \delta_{ij} = 1)$ the number of failures in sample j occurred by time u. Theorem 7.2.1 in Fleming and Harrington (1991) states that, under H_0, if $W(t) = f[\widehat{S}^{KM}(t-)]$ or $W(t) = f[\widehat{\pi}(t)]$, where $\widehat{S}^{KM}$ is the Kaplan–Meier estimator computed with the pooled samples, and $\widehat{\pi}(t)$ is the pooled estimator of the probability that a subject is alive and uncensored at time t, and defining

$$\widehat{\sigma}^2_{LR} = \frac{n_1 + n_2}{n_1 n_2} \int_0^\tau W^2(v) \frac{Y_1(v)Y_2(v)}{(Y_1(v) + Y_2(v))^2} \left(1 - \frac{\triangle N_1(v) + \triangle N_2(v) - 1}{Y_1(v) + Y_2(v) - 1}\right) (dN_1(v) + dN_2(v)), \tag{2}$$

where $\tau = \min(Z_{(n_1)1}, Z_{(n_2)2})$ and $Z_{(n_j)j}$ is the n_jth order statistic of sample j, then, as $n \to \infty$, $G^{LR}/\widehat{\sigma}_{LR}$ converges in distribution to a standard normal distribution. So, a two-sided α-level test rejects H_0 when $|G^{LR}/\widehat{\sigma}_{LR}| > z_{\alpha/2}$, where z_α is the $(1-\alpha)100$th percentile of the standard normal.

The weighted LR test (1) can be interpreted as a weighted comparison of NA estimates computed by giving mass only to the uncensored observations. A key function in presmoothing is $p_j(t) = P(\delta_{ij} = 1 | Z_{ij} = t)$, the conditional probability that the ith observed time in sample j is not censored. The presmoothing approach consists in replacing $\mathbf{1}(\delta_{ij} = 1)$ in the expression of the NA estimator with a smooth estimate of $p_j(Z_{ij})$. Thus, all observations contribute information regardless of censoring status. Since $p_j(t) = \mathrm{E}(\delta_{ij} | Z_{ij} = t)$, it can be estimated with a regression estimator. To avoid model misspecification, we use the Nadaraya–Watson (NW) nonparametric estimator:

$$\widehat{p}_{j,b_j}(t) = \frac{\sum_{i=1}^{n_j} K_{b_j}\left(t - Z_{ij}\right) \delta_{ij}}{\sum_{i=1}^{n_j} K_{b_j}\left(t - Z_{ij}\right)},$$

where b_j is the presmoothing bandwidth, K a kernel function, and $K_{b_j}(t) = b_j^{-1} K(b_j^{-1} t)$ the rescaled kernel. We assume that K is a symmetric, twice continuously differentiable density, with bounded variation and support on the interval $[-1, 1]$. The weighted PLR test statistic (Jácome and López-de-Ullibarri, 2014) is defined by

$$G^P_{\boldsymbol{b}}(t) = \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \int_0^t W(v) \frac{Y_1(v)Y_2(v)}{Y_1(v) + Y_2(v)} (d\Lambda^P_{1,b_1}(v) - d\Lambda^P_{2,b_2}(v)),$$

where $\boldsymbol{b} = (b_1, b_2)$, $\Lambda^P_{j,b_j}(v) = \int_0^v Y_j^{-1}(u) dN^P_{j,b_j}(u)$ is the presmoothed estimate of $\Lambda_j(v)$, and $N^P_{j,b_j}(u) = \sum_{i=1}^{n_j} \mathbf{1}(Z_{ij} \leq u) \widehat{p}_{j,b_j}(Z_{ij})$ is the presmoothed estimate of the number of failures occurred by time u in sample j. Note that when $b_j \simeq 0$, then $\widehat{p}_{j,b_j}(Z_{ij}) \simeq \mathbf{1}(\delta_{ij} = 1)$, that is, the PLR test reduces to the LR test by taking $\boldsymbol{b} = (0, 0)$. Theorem 2.1 in Jácome and López-de-Ullibarri (2014) shows that if the distribution of the censoring variables is the same for both populations and the bandwidths are $b_j = C_j n^{-\beta_j} + o(n^{-\beta_j})$, with $1/4 < \beta_j < 1/2$, then LR and PLR tests have the same asymptotic normal distribution under the null hypothesis.

In estimation problems, the literature is rich in methods for bandwidth selection, which in general aim at minimizing an error criterion. Evidence from simulation studies suggests that bandwidth choice in presmoothed estimation is not crucial, since presmoothed estimators tend to outperform their classical counterparts also in a neighborhood of the optimal bandwidth (Cao et al., 2005; Cao and Jácome, 2004). The `survPresmooth` package (López-de-Ullibarri and Jácome, 2013) provides an implementation in R (R Core Team, 2016) of two bandwidth selection methods for the presmoothed estimators of the functions S, Λ, $f = -S'$ and $\lambda = \Lambda'$.

Correct bandwidth choice is also expected to increase the power of the PLR over the LR test. A simple approach is to use estimation-based optimal bandwidths. Simulations presented in Jácome and López-de-Ullibarri (2014) show that the PLR test performs well with those bandwidths. Nevertheless, since in testing problems interest is in constructing a powerful test rather than a good estimator, estimation-based optimal bandwidths may be suboptimal. Bandwidth selection methods