A BOOTSTRAP APPROACH TO NONPARAMETRIC REGRESSION FOR RIGHT CENSORED DATA

GANG LI$^1$ AND SOMNATH DATTA$^2$

$^1$Department of Biostatistics, School of Public Health, University of California, Los Angeles, CA 90095-1772, U.S.A.

$^2$Department of Statistics, Franklin College of Arts and Sciences, University of Georgia, 204 Statistics Building, Athens, GA 30602-1952, U.S.A., e-mail: datta@stat.uga.edu

(Received May 13, 1999; revised April 4, 2000)

Abstract. In this paper a two-stage bootstrap method is proposed for nonparametric regression with right censored data. The method is applied to construct confidence intervals and bands for a conditional survival function. Its asymptotic validity is established using counting process techniques and martingale central limit theory. The performance of the bootstrap method is investigated in a Monte Carlo study. An illustration is given using a real data.

Key words and phrases: Bootstrap, Beran’s estimate, censoring, confidence bands, confidence intervals, Kaplan-Meier estimate, nonparametric regression, quantile regression.

1. Introduction and a review

In lifetime data analysis, nonparametrically estimated conditional survival curves (such as the conditional Kaplan-Meier estimate) are useful for assessing the influence of risk factors, predicting survival probabilities, and checking goodness-of-fit of various survival regression models. However, it has not been an easy task to assess the variability of the estimated conditional survival curves. Consider the right censored survival regression data consisting of $n$ i.i.d. triples $(X_1, \delta_1, Z_1), \ldots, (X_n, \delta_n, Z_n)$, where $X_i = \min(T_i, C_i)$, $\delta_i = I(T_i < C_i)$, and $T_i \geq 0$, $C_i \geq 0$, and $Z_i$ represent the survival time, the censoring time, and the covariate, respectively, for the $i$-th subject under study, $i = 1, \ldots, n$. To ensure the identifiability of the model, we assume that for each $i$, $T_i$ and $C_i$ are conditionally independent given $Z_i$. Let $S(t \mid z) = P(T_i > t \mid Z_i = z)$ and $A(t \mid z) = -\int_0^t S(ds \mid z)/S(s - \mid z)$ denote the conditional survival function and the conditional cumulative hazard function of $T_i$ given $Z_i = z$, respectively. We study the problem of constructing nonparametric confidence bands and intervals for $S(t \mid z)$ and $A(t \mid z)$ using the optimal rate conditional Kaplan-Meier estimate of Beran (1981). Such confidence bands and intervals can be used to assess the variability of the estimated conditional survival probabilities and provide a useful scale against which unusual features of the estimated conditional survival curve may be evaluated.

Nonparametric estimation of the conditional survival function and its related functions was initiated by Beran (1981) and has been further studied by Dabrowska (1987, 1989, 1992), Li (1997), Li and Doss (1995), McKeague and Utikal (1990), and others. For the convenience of discussion, let us consider a simple version of Beran’s (1981) esti-
mate $\hat{S}_h(t \mid z)$ of $S(t \mid z)$ that is defined as the Kaplan-Meier estimate constructed using only those observations whose covariate $Z$ fall inside a neighborhood $(z - h_n, z + h_n)$ of $z$, where $h_n$ is called the bandwidth (the general definition of Beran's estimate is given in Section 2). Because only a portion of the data are used, the rate of convergence of $\hat{S}_h(t \mid z)$ to $S(t \mid z)$ is typically slower than the $n^{-1/2}$ rate for the ordinary Kaplan-Meier estimate. The actual rate of convergence depends on how fast $h_n$ goes to 0 as well as on the smoothness of $S$. A decrease in the bandwidth would reduce the bias but, on the other hand, increase the variance of the estimate, and vice versa. Under certain smoothness conditions, it has been shown that (cf. Dabrowska (1987) and Li (1997)) if $h_n$ is of order $n^{-1/5}$, then $\hat{S}_h(t \mid z)$ converges to $S(t \mid z)$ at an optimal rate. Moreover,

\[(nh_n)^{1/2}(\hat{S}_h(t \mid z) - S(t \mid z)) \xrightarrow{d} U(t \mid z),\]

where for fixed $z$, $U(t \mid z)$ is a continuous Gaussian martingale process with a nonzero mean (see (2.6) and (2.7) below for explicit expressions of the mean and variance function of $U$). In particular, the optimal rate of convergence for $\hat{S}_h(t \mid z)$ is $(nh_n)^{-1/2} = O(n^{-2/5})$.

It, however, remains an open problem as to how to construct confidence bands and intervals for $S(t \mid z)$ using the optimal rate weak convergence result (1.1). The major hurdles are that $U$ has an unknown nonzero mean and that the distribution of $\sup_t |U(t \mid z)|$ is intractable. One possible solution is to “undersmooth” the Beran estimate: if $nh_n^5 \rightarrow 0$, then the limiting process $U$ will have a zero mean and a Hall-Wellner (1980) type confidence band for $\hat{S}(t \mid z)$ can be constructed (cf. Li and Doss (1995)). The drawback of this approach is that undersmoothing slows down the rate of convergence, which is not desired.

The main purpose of this paper is to study a bootstrap method for censored nonparametric regression and use it to construct confidence bands and intervals for $S(t \mid z)$ from the optimal rate Beran estimate. The basic idea of bootstrap is to first resample from the observed data, then reconstruct the estimate of interest, say Beran’s estimate $\hat{S}_h^* (t \mid z)$, from the resampled data, and finally approximate the distribution of $(nh_n)^{1/2}(\hat{S}_h(t \mid z) - S(t \mid z))$ by that of its bootstrap version which can be obtained via computer simulation. An advantage of the bootstrap approach is that with an appropriately designed resampling method, the bootstrap will correctly account for the bias of the estimated survival function $\hat{S}_h(t \mid z)$. Therefore it does not require additional bias estimation or the use of suboptimal rate estimate (undersmoothing) for constructing confidence bands or intervals. It also automatically adapts to different variances of the estimated survival function at different covariate locations. However, it is not obvious what resampling scheme should be used for censored nonparametric regression as discussed below.

The idea of bootstrap was first introduced by Efron (1979) for homogeneous i.i.d. complete data setup in which the bootstrap is carried out by “resampling with replacement” from the sample data. This approach has since become a powerful tool in many statistical applications. See, for example, Efron and Tibshirani (1993) for the bootstrap method and its applications. Bootstrap for right censored data with no covariate was first studied by Efron (1981) who proposed two equivalent versions of bootstrap: a “simple” version and an “obvious” version. The “simple” bootstrap of Efron (1981) “resamples with replacement” from the observed pairs $\{(X_i, \delta_i), i = 1, \ldots, n\}$. The validity of this method was established by Akritas (1986) and Lo and Singh (1986). For