Computational Statistics and Data Analysis 56 (2012) 785–796

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Variance estimation in censored quantile regression via induced smoothing

Lei Pang, Wenbin Lu, Huixia Judy Wang *

Department of Statistics, North Carolina State University, Raleigh, NC 27606, USA

ARTICLE INFO

Article history:
Received 9 May 2010
Received in revised form 9 October 2010
Accepted 14 October 2010
Available online 30 October 2010

Keywords:
Censored quantile regression
Smoothing
Survival analysis
Variance estimation

ABSTRACT

Statistical inference in censored quantile regression is challenging, partly due to the unsmoothness of the quantile score function. A new procedure is developed to estimate the variance of the Bang and Tsiatis inverse-censoring-probability weighted estimator for censored quantile regression by employing the idea of induced smoothing. The proposed variance estimator is shown to be asymptotically consistent. In addition, a numerical study suggests that the proposed procedure performs well in finite samples, and it is computationally more efficient than the commonly used bootstrap method.

1. Introduction

Consider a random sample subject to random right censoring: $\{Y_i = \min(T_i, C_i), \Delta_i, Z_i\}_{i=1}^n$, where T_i and C_i denote the failure time and the censoring time (or some monotone transformations thereof), respectively, $\Delta_i = I(T_i \le C_i)$ is the censoring indicator, and Z_i is the p-dimensional covariate vector with 1 as the first component corresponding to the intercept. We assume the following quantile regression model at a fixed quantile level $\tau \in (0, 1)$,

$$T_i = Z_i'\beta_0(\tau) + \epsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where $\beta_0(\tau)$ is the p-dimensional quantile coefficient at the τth quantile, and ϵ_i is the random error whose τth conditional quantile, given Z_i, equals 0. In the rest of the article, we will suppress τ in $\beta_0(\tau)$ for notational simplicity.

The quantile regression model, first introduced by Koenker and Bassett (1978), is a valuable alternative to the Cox proportional hazards model (Cox, 1972) and the accelerated failure time (AFT) model (Cox and Oaks, 1984) in survival analysis. Compared to the Cox and AFT models, quantile regression offers an automatic and flexible way to capture the heterogeneity in the data. Moreover, the relations between the conditional quantiles of the failure time and the covariates are directly interpretable.

Considerable research efforts have been devoted to the estimation of regression parameters in censored quantile regression. Early works of Powell (1984, 1986) focus on fixed censoring where the censoring variable C_i is always observable. For quantile regression with random censoring, recent developments include Ying et al. (1995), Bang and Tsiatis (2002), Honoré et al. (2002), Portnoy (2003), Peng and Huang (2008), Wang and Wang (2009), Huang (2010) and Portnoy and Lin (2010). However, relatively fewer methods are available for estimating the variance of censored quantile regression estimators and as such the associated inference becomes more difficult. In general, a plug-in estimator of the variance can be obtained based on the asymptotical normality of the corresponding estimator, while the asymptotical variance–covariance matrix usually involves $f_i(0|Z_i)$, the unknown conditional error density function $f_i(\cdot|Z_i)$ evaluated at zero. Nonparametric

* Corresponding author. Tel.: +1 919 513 1661; fax: +1 919 515 7591.
E-mail addresses: judy_wang@ncsu.edu, wang@stat.ncsu.edu (H.J. Wang).

0167-9473/$ – see front matter
doi:10.1016/j.csda.2010.10.018

approaches can be used to estimate $f_i(0|Z_i)$, but most of them require moderate to large sample size and choosing proper tuning parameters, such as the bandwidth parameter in kernel smoothing. Previous research shows that inference based on nonparametric density estimation is very sensitive to the choice of tuning parameters in finite samples; see Chen and Wei (2005) and Kocherginsky et al. (2005) for comparison of different inference methods in quantile regression without censoring. Some other researchers considered computationally intensive methods for variance estimation, e.g. Portnoy (2003) and Wang and Wang (2009) used bootstrap, and Peng and Huang (2008) employed the resampling approach of Jin et al. (2001). To bypass the variance estimation in the context of censored quantile regression, an alternative inference approach via the inversion of score tests was also considered by Ying et al. (1995) and Bang and Tsiatis (2002). However, the construction of confidence intervals is cumbersome, because one needs to perform hypothesis testing repeatedly on a fine grid in the p-dimensional parameter space.

The difficulty in variance estimation for censored quantile regression arises partly from the unsmoothness of the corresponding estimating function that involves indicator functions. The unsmoothness issue can be overcome by approximating the indicator functions with some suitable smooth functions. A common smoothing method is the kernel smoothing; see Chen and Hall (1993), Horowitz (1998), Heller (2007) and Whang (2006). Using kernel smoothing, if the bandwidth parameter converges to zero at a proper rate, the smoothed estimating function is asymptotically equivalent to the original estimating function, and the covariance can be consistently estimated by the sandwich formula derived from the smoothed estimating equation. Though conceptually simple, kernel smoothing usually needs a well-chosen bandwidth parameter to achieve good finite sample performance.

In this paper, we employ the induced smoothing idea of Brown and Wang (2005) and develop a variance estimation procedure for the censored quantile coefficient estimator of Bang and Tsiatis (2002). The induced smoothing method smoothes the estimating function by taking its expectation under the distribution of a random perturbation of the unknown parameter. Induced smoothing is practically convenient, as it does not require selecting any tuning parameters. Recently, Wang et al. (2009) demonstrated that induced smoothing provides reliable variance estimation for quantile regression with uncensored data. Other notable applications of induced smoothing in different contexts include Brown and Wang (2006), Johnson and Strawderman (2009), Fu et al. (2010) and Wang and Fu (2010). In this paper, we will show that for censored quantile regression, our proposed variance estimator via induced smoothing is asymptotically consistent. In addition, it has good finite sample performance and is computationally efficient, requiring only a fraction of the computational cost as compared to the bootstrap method.

The rest of this article is organized as follows. In Section 2, we present the variance estimation procedure. In Section 3, we establish the asymptotic property of the variance estimator. The finite sample performance of the proposed procedure is investigated through a simulation study in Section 4, and an analysis of a multiple myeloma data set in Section 5. We conclude the paper with some discussions in Section 6. Theoretical proofs are relegated to the Appendix.

2. Induced smoothing and computational algorithm

When there is no censoring, the quantile coefficient β_0 in model (1) can be consistently estimated using the solution of the following estimating equation

$$n^{-1}\sum_{i=1}^{n} Z_i\{I(Y_i - \beta' Z_i < 0) - \tau\} = 0.$$

In the presence of random right censoring, a number of estimation methods have been proposed. One popular estimator is the inverse-censoring-probability-weighted (ICPW) estimator proposed by Bang and Tsiatis (2002). Specifically, the ICPW estimator $\hat{\beta}$ is defined as the solution of the estimating equation

$$U_n(\beta) = n^{-1}\sum_{i=1}^{n} \frac{Z_i \Delta_i}{\hat{G}(Y_i)}\{I(Y_i - \beta' Z_i < 0) - \tau\} = 0, \tag{2}$$

where $\hat{G}(\cdot)$ is the Kaplan–Meier estimate of the survival function of the censoring variable C_i. Note that, solving (2) is equivalent to minimizing a weighted objective function

$$n^{-1}\sum_{i=1}^{n} \frac{\Delta_i}{\hat{G}(Y_i)}\rho_\tau(Y_i - \beta' Z_i), \tag{3}$$

where $\rho_\tau(t) = t\tau - tI(t < 0)$. The minimization can be solved efficiently by using existing linear programming algorithms, for instance, the *rq* function in R package *quantreg*.

In the Bang and Tsiatis method, censoring times are assumed to be independent and identically distributed, and independent of covariates, which we also assume throughout the rest of this paper. Bang and Tsiatis (2002) proved that under some regularity conditions, $\hat{\beta}$ is consistent and asymptotically normal:

$$n^{1/2}(\hat{\beta} - \beta_0) \xrightarrow{d} \mathcal{N}(0, \mathbf{\Gamma}),$$

where $\mathbf{\Gamma} = \mathbf{A}^{-1}\mathbf{\Sigma}\mathbf{A}^{-1}$, $\mathbf{A} = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} Z_i Z_i' f_i(0|Z_i)$, and $\mathbf{\Sigma} = \lim_{n\to\infty} \mathrm{Var}\{n^{1/2}U_n(\beta_0)\}$. It is well known that the direct estimation of $\mathbf{\Gamma}$ is difficult, since the matrix $\mathbf{A}$ depends on the unknown error densities $f_i(\cdot|Z_i)$, $i = 1, \ldots, n$.