Journal of Statistical Planning and Inference 191 (2017) 13–24

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Hypothesis testing for regional quantiles

Seyoung Park [a,*], Xuming He [b]

[a] School of Public Health, Yale University, New Haven, CT, 06511, USA
[b] Department of Statistics, University of Michigan, Ann Arbor, MI 48109, USA

ARTICLE INFO

Article history:
Received 9 January 2017
Received in revised form 5 June 2017
Accepted 6 June 2017
Available online 21 June 2017

Keywords:
Hypothesis test
Quantile regression
Multiple quantiles
High dimensional
B-spline

ABSTRACT

We consider the problem of testing significance of predictors in quantile regression, where the sample size n and the number of predictors are allowed to increase together. Unlike the quantile regression analysis for the τth quantile at a given $\tau \in (0, 1)$, we aim to detect any covariate that is significant for the conditional quantiles at any level of interest in a given region, $\tau \in \Delta$. We use B-splines to approximate the quantile functions as τ varies and consider the composite quantile regression to estimate the parameters. The proposed score-type test admits normal approximations even in the presence of high dimensional variables. Through numerical examples, we demonstrate that the proposed test can provide higher power than existing tests designed for single quantile levels.

1. Introduction

Quantile regression has become a widely used method in evaluating the effect of regressors on the conditional distribution of a response variable (Koenker, 2005). Compared with the least squares analysis, quantile regression is robust to the misspecification of error distributions and characterizes the relationship between the response variable and the covariates in a more comprehensive way. In quantile regression analysis, it is certainly of interest to develop methods for testing the significance of certain predictors to inform rational decisions about their effects to the response. Koenker (2005) considers Wald-type and score-type tests at a given quantile level. Kocherginsky et al. (2005) propose a time-saving resampling method based on a modification of the Markov chain marginal bootstrap to construct confidence intervals. Volgushev et al. (2013) consider the problem of significance tests in multivariate nonparametric quantile regression.

Much of the existing literature considers significance tests at a given quantile level with the number of regressors p fixed in the model. In this paper, instead of focusing on a specific quantile level τ, we consider an interval Δ of quantile levels with high dimensional variables. For example, Δ may be chosen as $[0.4, 0.6]$ instead of just $\tau = 0.5$ if we would like to detect variables that impact the center of the conditional distributions, or $[0.75, 0.9]$ if we are interested in the upper tails. This approach is partly motivated by the work by Zheng et al. (2015), where they propose a globally concerned model selection strategy that examines regression quantiles over a set of quantile levels Δ.

There are potential advantages of considering a set of quantile levels instead of a single or multiple quantile levels. First, we may gain power in the statistical analysis by polling information across quantile levels. Second, it is often difficult to justify the choice of one quantile level, whether it is $\tau = 0.75$ or 0.9, if we are interested in the upper quantiles. When we speak of upper quantiles, it can make better sense if we consider all $\tau \geq 0.75$ up to a reasonable upper bound.

For testing over a specified range of quantiles, Koenker and Machado (1999) exploit the Wald processes or Rankscore process by using the asymptotic behavior of the Bessel process under fixed p-asymptotics. We consider high-dimensional

* Corresponding author.
E-mail addresses: seyoung.park@yale.edu (S. Park), xmhe@umich.edu (X. He).

http://dx.doi.org/10.1016/j.jspi.2017.06.002
0378-3758/

settings by allowing p to increase with n and use B-splines to approximate the quantile functions. We exploit a composite quantile regression estimate (Zou and Yuan, 2008) and construct a score-type test based on the asymptotic normality of the statistic.

The rest of the paper is organized as follows. We introduce the proposed score-type test in Section 2. In Section 3, we investigate the theoretical properties of the test statistic. In Section 4, we evaluate the finite sample performance of the proposed method by Monte Carlo simulation. In numerical examples, we demonstrate that our proposed test preserves the type I error and provides higher power, compared to the existing tests. A real data example is given in Section 5. Some concluding remarks are given in Section 6. Technical proofs are deferred to Appendix.

2. The proposed method

2.1. The model

Consider the following linear quantile model: for $i = 1, \ldots, n$,

$$y_i = x_i^T \alpha(\tau) + z_i^T \beta(\tau) + \epsilon_i(\tau) \quad \text{for } \tau \in \Delta, \tag{1}$$

where $\Delta \subset (0, 1)$ is an interval of quantile levels of interest, and $x_i \in \mathbb{R}^{p_n}$ and $z_i \in \mathbb{R}^{q_n}$ are ith fixed design vectors. Here p_n and q_n can increase with n. The random variables $\epsilon_i(\tau)$s are independent across i and satisfy $P\{\epsilon_i(\tau) \le 0 \mid x_i, z_i\} = \tau$ for all $\tau \in \Delta$ and $i = 1, \ldots, n$. The vectors $\alpha(\tau) \in \mathbb{R}^{p_n}$ and $\beta(\tau) \in \mathbb{R}^{q_n}$ are the τth conditional quantile coefficients in the sense that ${x_i}^T \alpha(\tau) + z_i^T \beta(\tau)$ is the τth conditional quantile of y_i given x_i and z_i. Let

$$y = [y_1, \ldots, y_n]^T,\ X = [x_1, \ldots, x_n]^T,\ Z = [z_1, \ldots, z_n]^T,\ \epsilon(\tau) = [\epsilon_1(\tau), \ldots, \epsilon_n(\tau)]^T.$$

The above linear quantile model can be written in the following matrix form:

$$y = X\alpha(\tau) + Z\beta(\tau) + \epsilon(\tau). \tag{2}$$

We assume that each column of the matrices X and Z is normalized to mean zero and the L_2 norm $\sqrt{n}$. Throughout the paper, we are interested in testing

$$H_0 : \beta(\tau) = 0_{q_n} \text{ for all } \tau \in \Delta \quad \text{versus} \quad H_1 : \beta(\tau) \neq 0_{q_n} \text{ for some } \tau \in \Delta. \tag{3}$$

2.2. The method

To assess the hypothesis (3), we use B-spline basis functions to approximate the quantile function $\alpha(\tau)$ for τ in the smallest interval that contains Δ. Following Schumaker (1981), Pena (1997), Kim (2007), and Wang et al. (2009), let $\Pi_{m_n}(\tau) = \{\pi_1(\tau), \ldots, \pi_{m_n+l}(\tau)\}^T$ be the normalized B-spline basis functions of order l with m_n quasi-uniform knots. For simplicity, let $K_n = m_n + l$. We approximate $\alpha(\tau)$ in (2) by a linear combination of $\pi_j(\tau)$s, i.e., $\alpha(\tau) \approx \Theta \Pi_{m_n}(\tau)$, where Θ is the $p_n \times K_n$ spline coefficient matrix. Hence, under H_0, the model (2) can be approximated by

$$y \approx X\Theta\Pi_{m_n}(\tau) + \epsilon(\tau).$$

In the theoretical analysis in Section 3, we reflect the approximation errors induced by the spline approximation. Let $\tau_1, \ldots, \tau_{b_n}$ be the equally spaced quantile levels from the set $\Delta = [\Delta_1, \Delta_2]$. We fix $\tau_1 = \Delta_1$ and $\tau_{b_n} = \Delta_2$ so that the distance between neighborhood quantiles are all equal to $(\Delta_2 - \Delta_1)/(b_n - 1)$. Under H_0, the spline coefficient matrix estimate $\widehat{\Theta}$ is obtained by the following composite quantile regression (Zou and Yuan, 2008):

$$\widehat{\Theta} = \underset{\Theta \in \mathbb{R}^{p_n \times K_n}}{arg\,min} \frac{1}{nb_n} \sum_{k=1}^{b_n} \sum_{i=1}^{n} \rho_{\tau_k}\{y_i - x_i^T \Theta \Pi_{m_n}(\tau_k)\}, \tag{4}$$

where $\rho_\tau(u) = u(\tau - I_{\{u<0\}})$ is the quantile loss function (Koenker and Basset, 1978). Let

$$v_i^{(k)} = \text{vec}(\Pi_{m_n}(\tau_k) \otimes x_i) \text{ for } k = 1, \ldots, b_n\ ;\ i = 1, \ldots, n.$$

Then, (4) can be rewritten as

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^{p_n K_n}}{arg\,min} \frac{1}{nb_n} \sum_{k=1}^{b_n} \sum_{i=1}^{n} \rho_{\tau_k}\{y_i - (v_i^{(k)})^T \theta\}, \tag{5}$$

where $\hat{\theta} = \text{vec}(\widehat{\Theta})$. Let $v^{(k)} = [v_1^{(k)} \mid, \ldots, \mid v_n^{(k)}]^T$ be an $n \times p_n K_n$ matrix for $k = 1, \ldots, b_n$. Let $f_i^{(\tau)}(t)$ be the conditional density function of $\epsilon_i(\tau)$ given $w_i := (x_i^T, z_i^T)^T \in \mathbb{R}^{p_n+q_n}$ evaluated at t. Let $H^{(k)} = \text{diag}(f_1^{(\tau_k)}(0), \ldots, f_n^{(\tau_k)}(0))$ be a diagonal matrix of the conditional densities at 0. Let

$$P^{(k)} = H^{(k)} X [X^T (H^{(k)})^2 X]^{-1} X^T H^{(k)}, \quad \tilde{P}^{(k)} = I_n - P^{(k)}, \quad Z^{(k)} = \tilde{P}^{(k)} Z.$$