Journal of Multivariate Analysis 150 (2016) 169–182

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Single index quantile regression for heteroscedastic data

Eliana Christou [a,*], Michael G. Akritas [b]

[a] The Pennsylvania State University, 325 Thomas Building, University Park, PA 16802, United States

[b] The Pennsylvania State University, 414 Thomas Building, University Park, PA 16802, United States

ARTICLE INFO

Article history:
Received 11 December 2015
Available online 7 June 2016

AMS subject classification:
62G05
62G08
62G20
62G35

Keywords:
Dimension reduction
Index model
Nadaraya–Watson estimator
Quantile regression

ABSTRACT

Quantile regression (QR) is becoming increasingly popular due to its relevance in many scientific investigations. Linear and nonlinear QR models have been studied extensively, while recent research focuses on the single index quantile regression (SIQR) model. Compared to the single index mean regression (SIMR) problem, the fitting and the asymptotic theory of the SIQR model are more complicated due to the lack of closed form expressions for estimators of conditional quantiles. Consequently, the proposed methods are necessarily iterative. We propose a non-iterative estimation algorithm, and derive the asymptotic distribution of the proposed estimator under heteroscedasticity. For identifiability, we use a parametrization that sets the first coefficient to 1 instead of the typical condition which restricts the norm of the parametric component. This distinction is more than simply cosmetic as it affects, in a critical way, the correspondence between the estimator derived and the asymptotic theory.

1. Introduction

Ordinary least squares regression plays a prominent role in a wide variety of fields and is a very popular method for modeling the relationship between a d-dimensional vector of covariates $\mathbf{X}$ and the conditional mean of the response variable Y given $\mathbf{X} = \mathbf{x}$. There are cases, however, where interest lies in certain conditional quantiles; see Koenker and Hallock [14] for a practical implementation. When the error term is heteroscedastic, a direct approach for estimating conditional quantiles has a number of advantages. Koenker and Bassett [13] proposed such a direct approach. Letting

$$Q_\tau(Y|\mathbf{x}) \equiv Q_\tau(Y|\mathbf{X}=\mathbf{x}) = \inf\{y : \Pr(Y \le y|\mathbf{X}=\mathbf{x}) \ge \tau\}$$

denote the τth conditional quantile of Y given $\mathbf{X} = \mathbf{x}$, they considered the linear quantile regression (QR) model

$$Q_\tau(Y|\mathbf{x}) = \boldsymbol{\beta}^\top \mathbf{x} \tag{1.1}$$

and used the representation

$$Q_\tau(Y|\mathbf{x}) = \arg\min_q \mathrm{E}\{\rho_\tau(Y-q)|\mathbf{X}=\mathbf{x}\}, \tag{1.2}$$

where, for $0 < \tau < 1$, the function $\rho_\tau(\cdot)$ is the loss function, also known as check function, defined as $\rho_\tau(u) = \{\tau - I(u < 0)\}u$, to define the estimator $\widehat{\boldsymbol{\beta}}$ as

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\mathbf{b}\in\mathbb{R}^d} \sum_{i=1}^n \rho_\tau(Y_i - \mathbf{b}^\top \mathbf{X}_i).$$

* Corresponding author.
E-mail addresses: exc277@psu.edu (E. Christou), mga@stat.psu.edu (M.G. Akritas).

http://dx.doi.org/10.1016/j.jmva.2016.05.010
0047-259X/

Thus, $\widehat{\boldsymbol{\beta}}^\top \mathbf{x}$ gives the estimator of the τth conditional quantile under the linear QR model. Observe that for $\tau = 1/2$, the objective function is the L_1 norm, that is

$$\arg\min_{\mathbf{b}\in\mathbb{R}^d} \sum_{i=1}^{n} \frac{1}{2}|Y_i - \mathbf{b}^\top \mathbf{X}_i|,$$

which gives the estimated conditional median. It turns out that QR inherits the well known robustness properties of the median regression; see Pollard [22]. Koenker and Bassett [13] studied the asymptotic statistical behavior of the estimated conditional regression quantiles, while Koenker [11] studied confidence intervals for the regression quantiles, based on the asymptotic theory.

Because the linearity assumption of model (1.1) is quite strict, several authors considered the completely flexible nonparametric estimation of the conditional quantiles. Truong [27] showed that, under conditions, local median estimators achieve the global optimal rates of Stone [24] with respect to L_m norms, $0 < m \le \infty$. Chaudhuri [2] constructed local polynomial estimators for conditional quantile functions and their derivatives, and also showed that they achieve the optimal nonparametric rates of convergence of Stone [24] under mild conditions. A local Bahadur type representation was also established by Chaudhuri [2] when the kernel function is uniform, and this result was later extended to general kernels by Hong [9]. Fan et al. [4] considered a general convex loss function, that includes the mean, median, quantiles, and other robust functionals, and constructed local linear estimators. See also Yu and Jones [29] who proposed inverting a local linear conditional distribution estimator. Takeuchi et al. [26] presented a nonparametric version of a quantile estimator, which can be obtained by solving a simple quadratic programming problem and provide uniform convergence results. Kong et al. [15] extended Chaudhuri's [2] and Hong's [9] pointwise Bahadur representation result by deriving a strong uniform (with respect to $\mathbf{x}$) Bahadur representation also for dependent observations. Guerre and Sabbah [5] investigated the bias and the weak Bahadur representation of a local polynomial estimator of the conditional quantile function and its derivatives uniformly with respect to the quantile level, the covariates and the smoothing parameter. Also, they showed that the local polynomial quantile estimator achieves the global optimal rates of Stone [24] for the L_m and uniform norms.

The rate of convergence of completely nonparametric estimators of conditional quantiles, however, decreases with increasing dimensionality of the covariate vector. This motivated the study of a number of semiparametric models, and of variable selection methods, for QR. Koenker [12] considered the additive model for QR which includes both parametric and nonparametric components. Lin et al. [18] considered variable selection for nonparametric QR via smoothing spline ANOVA (SS-ANOVA). See Su and Zhang [25] for a literature review.

The single index quantile regression (SIQR) model has received particular attention. It specifies that

$$Q_\tau(Y|\mathbf{x}) = Q_{\tau,\boldsymbol{\beta}_1}(Y|\boldsymbol{\beta}_1^\top \mathbf{x}), \tag{1.3}$$

where for any d-dimensional vector $\mathbf{b}_1$,

$$Q_{\tau,\mathbf{b}_1}(Y|\mathbf{b}_1^\top \mathbf{x}) = \inf\{y : \Pr(Y \le y|\mathbf{b}_1^\top \mathbf{X} = \mathbf{b}_1^\top \mathbf{x}) \ge \tau\}. \tag{1.4}$$

For identifiability one imposes certain conditions on $\boldsymbol{\beta}_1$, the most common of which is to assume that $\|\boldsymbol{\beta}_1\| = 1$, with its first coordinate positive. In this paper, we propose the parametrization which assumes that $\boldsymbol{\beta}_1 = (1, \boldsymbol{\beta}^\top)^\top$, $\boldsymbol{\beta} \in \mathbb{R}^{d-1}$; this parametrization is also used in the R package np for the single index mean regression (SIMR) model. This distinction is more than simply cosmetic as it affects, in a critical way, the correspondence between the estimator derived and the asymptotic theory. The advantages of the proposed parametrization are demonstrated in the simulations.

Existing literature considers the SIQR model under homoscedasticity (Wu et al. [28]), restricted heteroscedasticity (Chaudhuri et al. [3]) and general heteroscedasticity (Kong and Xia [16]). Chaudhuri et al. [3] considered the average derivative quantile regression estimator which, under a SIQR model where the variance function depends only on $\boldsymbol{\beta}_1^\top \mathbf{x}$, estimates the direction of $\boldsymbol{\beta}_1$. Wu et al. [28] and Kong and Xia [16] estimate $\boldsymbol{\beta}_1$ by minimizing an objective function. Compared with SIMR, (see Li and Racine [17, Chapter 8]), the main complication faced by this approach lies in the lack of a closed form expression for the estimator of conditional quantiles. Thus, the proposed methods are necessarily iterative. Wu et al. [28] proposed an algorithm which, starting from an initial value $\mathbf{b}_1^0$ for the parametric component, iteratively estimates the nonparametric component and its derivative using local linear QR, and the parametric component using (essentially) linear QR. Kong and Xia [16] criticized the convergence properties of the algorithm in Wu et al. [28] and proposed an improved iterative algorithm by introducing a penalty term that assures its almost sure convergence. In addition, Kong and Xia [16] allowed general heteroscedasticity, but the covariance function of the limiting normal distribution they obtained depends on the true value of the parametric component in an explicit manner.

In this paper we propose a *non-iterative* method, based on minimization of an objective function, for estimating the parametric component of the SIQR model. The proposed estimator is shown to have an asymptotically normal distribution, with a simple expression for the covariance matrix, under general heteroscedasticity.

In Section 2 we present the proposed estimator, while in Section 3 we present the main results, that include the $\sqrt{n}$-consistency and the asymptotic normality of the estimated parametric component. In Section 4 we present results from several simulation examples and a real data application on the Boston housing data. A discussion including conclusions and future work is given in Section 5.