Journal of Multivariate Analysis 155 (2017) 334–343

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Multiple quantile regression analysis of longitudinal data: Heteroscedasticity and efficient estimation

Hyunkeun Cho [a], Seonjin Kim [b,*], Mi-Ok Kim [c]

[a] *Department of Statistics, Western Michigan University, Kalamazoo, MI 49008, United States*
[b] *Department of Statistics, Miami University, Oxford, OH 45056, United States*
[c] *Department of Epidemiology and Biostatistics, University of California, San Francisco, CA 94143, United States*

ARTICLE INFO

Article history:
Received 24 May 2016
Available online 20 January 2017

AMS subject classification:
62G10

Keywords:
Asymptotic efficiency
Empirical likelihood
Heteroscedasticity test
Longitudinal data
Multiple quantiles

ABSTRACT

The objective of this paper is two-fold: to propose efficient estimation of multiple quantile regression analysis of longitudinal data and to develop a new test for the homogeneity of independent variable effects across multiple quantiles. Estimating multiple regression quantile coefficients simultaneously entails accommodating both association among the multiple quantiles and association among the repeated measurements of the response within subjects. We formulate simultaneous estimating equations using basis matrix expansion which accounts for the above-mentioned associations. The empirical likelihood method is adopted to estimate multiple regression quantile coefficients. Theoretical results show that the proposed simultaneous estimation is asymptotically more efficient than separate estimation of individual regression quantiles or ignoring the within-subject dependency. The proposed method also offers an empirical likelihood ratio test examining the homogeneity of the independent variable effects across the multiple quantiles. The Wilk's theorem holds for the test statistic, and thus the test is easy to implement. Simulation studies and real data example of a multi-center AIDS cohort study are included to illustrate the proposed estimation and testing methods, and empirically examine their properties.

1. Introduction

In longitudinal data, the temporal relationship between the response and explanatory variables can vary across the conditional distribution of the response when heteroscedasticity exists. For this reason, quantile regression has recently attracted attention; see Jung [6], He et al. [3], Koenker [10], Tang and Leng [18], Wang and Zhu [20], Tang et al. [19], and others. Quantile regression enables one to postulate varying effects of the independent variables across the conditional distribution. Given $\tau_1, \ldots, \tau_K \in (0, 1)$, the quantiles of repeatedly measured response variables conditioning on the independent variables are given, for each $i \in \{1, \ldots, n\}$, by

$$Q_{\tau_k}(y_i|x_i) = x_i\beta_{\tau_k},$$

[*] Corresponding author.
E-mail addresses: hyunkeun.cho@wmich.edu (H. Cho), kims20@miamioh.edu (S. Kim), miok.kim@ucsf.edu (M.-O. Kim).

http://dx.doi.org/10.1016/j.jmva.2017.01.009

where n is the number of subjects, $y_i = (y_{i1}, \ldots, y_{im_i})^\top$ denotes the m_i response measurements observed within subject i, $x_i = (x_{i1}, \ldots, x_{im_i})^\top$ is an $(m_i \times p)$-dimensional matrix of the independent variables collected from the ith subject, and $\beta_{\tau_k} = (\beta_{\tau_k,1}, \ldots, \beta_{\tau_k,p})^\top$ is a p-dimensional parameter vector at the τ_kth quantile level. While the quantile specific regression coefficients $\beta_{\tau_k}, \ldots, \beta_{\tau_k}$ collectively capture the relationships with the independent variables, most of the existing work estimate the individual regression quantiles β_{τ_k} separately for each $k \in \{1, \ldots, K\}$.

For each $k \in \{1, \ldots, K\}$, let $\varphi_{\tau_k}(u)$ be the derivative of the so-called check function $\rho_{\tau_k}(u) = u\{\tau_k - \mathbf{1}(u < 0)\}$ and set $\varphi(u) = (\varphi_{\tau_1}(u)^\top, \ldots, \varphi_{\tau_K}(u)^\top)^\top$. In this paper, we estimate $\beta_{\tau_1}, \ldots, \beta_{\tau_K}$ simultaneously using generalized estimating equations [12] as follows:

$$\sum_{i=1}^{n} X_i^\top A_i^{-1/2} R_i(\delta)^{-1} A_i^{-1/2} \varphi(Y_i - X_i\beta) = 0. \tag{1}$$

Here $Y_i = (y_i^\top, \ldots, y_i^\top)^\top$, $X_i = I_K \otimes x_i$ defined by a left Kronecker product operator $\otimes$ and a $(K \times K)$-dimensional identity matrix I_K, $\beta = (\beta_{\tau_1}^\top, \ldots, \beta_{\tau_K}^\top)^\top$, $\varphi(Y_i - X_i\beta) = (\varphi_{\tau_1}(y_i - x_i\beta_1)^\top, \ldots, \varphi_{\tau_K}(y_i - x_i\beta_K)^\top)^\top$, and A_i and $R_i(\delta)$ are an $(m_iK \times m_iK)$-dimensional diagonal variance matrix and working correlation matrix of $\varphi(Y_i - X_i\beta)$, respectively.

The term $R_i(\delta)$ in (1) is used to approximate the true correlation matrix of $\varphi(Y_i - X_i\beta)$, denoted by Π_i; it usually contains a low dimension of nuisance parameters δ associated with the within-subject correlation. The proposed simultaneous estimating equations accommodate not only the within-subject dependency commonly present in longitudinal data, but also a cross-correlation among the multiple quantiles, thereby providing a more efficient estimation. The efficiency gain, however, comes with an additional requirement of estimating nuisance parameters δ in $R_i(\delta)$. When a single quantile is concerned, Yi and He [23] obtained a more efficient estimator by estimating directly the true correlation matrix Π_i. Albeit possible, reliable estimation of Π_i is non-trivial, especially when m_i is large with a large number of nuisance parameters associated with the dependency structure, or a low or high quantile is of interest. This difficulty is exacerbated with multiple quantiles.

We propose to represent an inverse of $R_i(\delta)$ in (1) using the basis matrix expansion of Qu et al. [17] and avoid estimating the additional nuisance parameters associated with the correlation structure. The estimating equations (1) are expanded by basis matrices appropriately chosen for $R_i(\delta)$ and we use the empirical likelihood method [14] for the estimation of the regression coefficients β. Qin and Lawless [16] discussed the empirical likelihood method for generalized estimating equations; Yang and He [22] showed that the method similarly applies to quantile regression for independent data; Cho et al. [1] extended it to longitudinal data in the single quantile regression model. We show that the proposed simultaneous multiple quantile estimation approach is more efficient than the one either ignoring the inter-quantile correlation and estimating β_{τ_k} individually, or ignoring both the inter-quantile and within-subject correlation. As shown herein, simulation studies exhibit efficiency gain in finite samples with meaningful effects in a concrete application.

The proposed empirical likelihood approach also provides a test for the homogeneity of the independent variable relationship with the response across the multiple quantiles using the likelihood ratio statistic. Similar to its parametric counterpart, the test does not require estimating the covariance matrix of the quantile regression coefficient estimator. This is a highly desired property as the covariance matrix of the quantile regression coefficient estimator involves the densities of the conditional distribution of the response at the quantiles of interest. Inference for this purpose is surprisingly less developed, even though quantile regression analysis and heteroscedasticity are closely associated. For situations involving independent data, Koenker and Bassett [11] and Furno [2] proposed tests. The tests require estimation of the covariance matrix of quantile regression estimators and the performance hinges on reliable estimation of the densities at the quantiles of interest. As for the inference of individual coefficients, we adopt the random perturbation approach [5] to approximate the empirical distributions of the regression quantile estimators.

There has been growing interest in properly aggregating information across multiple quantiles under the homogeneity assumption of quantile coefficients in order to yield more efficient estimators; see Koenker [9], Portnoy and Koenker [15], Zou and Yuan [26], Xiao and Koenker [21], Kai et al. [7], and Zhao and Xiao [24]. The proposed empirical likelihood test can be used to validate this assumption. When the homogeneity of a quantile coefficient cannot be rejected, the proposed empirical likelihood estimation may further improve the estimation by constraining the common coefficient to be the same across multiple quantiles.

The paper is organized as follows. Section 2 proposes efficient estimation and statistical inference in the multiple quantile regression model. Sections 3 and 4 illustrate the proposed procedure with various simulation studies and an application to an HIV data set, respectively. All proofs of theorems are provided in the Appendix.

2. Methodology

2.1. Estimation of multiple quantile regression

The working correlation structure $R_i(\delta)$ in (1) plays an important role in increasing estimation efficiency. It involves two pieces of informative associations, a within-subject correlation, denoted by $C_i(\delta)$, and cross-correlation among K quantiles, denoted by G. Accordingly, the working correlation structure can be expressed in block matrix form as $R_i(\delta) = G \otimes C_i(\delta)$.