Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Q1 Composite quantile regression for correlated data

Q2 Weihua Zhao[a], Heng Lian[b,*], Xinyuan Song[c]

[a] *School of Science, Nantong University, Nantong, China*
[b] *Department of Mathematics, City University of Hong Kong, Kowloon Tong, Hong Kong*
[c] *Department of Statistics, Chinese University of Hong Kong, Shatin, Hong Kong*

ARTICLE INFO

Article history:
Received 29 April 2016
Received in revised form 14 November 2016
Accepted 27 November 2016
Available online xxxx

Keywords:
Composite quantile regression
Longitudinal data
Quadratic inference function
Weighted composite quantile regression

ABSTRACT

This study investigates composite quantile regression estimation for longitudinal data on the basis of quadratic inference functions. By incorporating the correlation within subjects, the proposed CQRQIF estimator has the advantages of both robustness and high estimation efficiency for a variety of error distributions. The theoretical properties of the resulting estimators are established. Given that the objective function is non-smooth and non-convex, an estimation procedure based on induced smoothing is developed. It is proved that the smoothed estimator is asymptotically equivalent to the original estimator. The weighted composite quantile regression estimation is also proposed to improve the estimation efficiency further in some situations. Extensive simulations are conducted to compare different estimators, and a real data analysis is used to illustrate their performances.

1. Introduction

Longitudinal data frequently arises in many economic studies and biomedical research, where measurements on the same individuals are taken repeatedly over time (Liang and Zeger, 1986; Diggle et al., 2002; Lin et al., 2015). Often, the primary goal is to characterize the dependence between the response and the factors correctly, while incorporating the correlated nature within subjects. At least three classes of important methodologies are widely used to analyze correlated longitudinal data, including marginal models, mixed-effects models and transition models, and each modeling approach serves specific analytic purposes. Among the three methods, the marginal model primarily aims to estimate the population-average effects of covariates on the response, and a well-known approach to statistical estimation is the use of generalized estimating equation (GEE) proposed by Liang and Zeger (1986). However, the application of GEE is restricted by the assumption that the correlation within subjects should be correctly specified. If the working correlation is misspecified, the estimation efficiency of GEE is lost. In addition, the GEE method is very sensitive to outliers or contaminated data (Qu and Song, 2004; Song, 2007).

To overcome some shortcomings of GEE, the quadratic inference functions (QIF) approach, which was first proposed by Qu et al. (2000), has recently received considerable attention. The QIF approach considers the within cluster correlation and is more efficient than the GEE approach when the working correlation is misspecified. The advantages of the QIF method have facilitated its use in many models, see for example Qu and Li (2006) for varying coefficient models, Bai et al. (2008) for partial linear models, Ma et al. (2014) and Lai et al. (2013) for partially linear single-index models, as well as

[*] Corresponding author.
E-mail address: henglian@cityu.edu.hk (H. Lian).

http://dx.doi.org/10.1016/j.csda.2016.11.015
0167-9473/© 2016 Elsevier B.V. All rights reserved.

Xue et al. (2010) and Wang et al. (2014) for generalized additive models and generalized additive partial linear models, respectively. For further discussions on the use of QIF over GEE, readers are referred to the review of Song et al. (2009).

Although QIF is more robust to outliers than GEE because of its bounded influence function, as shown in Qu and Song (2004), it may still be sensitive to heavy-tailed error distributions. In the extreme, when the error distribution does not have a finite second moment, the asymptotic normality of the estimator does not hold. This problem is obviously inherited from the standard least squares (LS) procedure for independent data. Median regression, as a special case of quantile regression (Koenker, 2005), is more robust to heavy-tailed error distribution but can have arbitrarily low efficiency than LS regression. To address this problem, Zou and Yuan (2008) proposed composite quantile regression (CQR) to obtain a highly efficient and robust estimator, which shows significant improvement over median regression in terms of estimation efficiency, and has also been extended to nonparametric (Kai et al., 2010) and semiparametric models (Kai et al., 2011) for cross-sectional data. Motivated by its superior performance, we investigate the performance of this method for longitudinal data by combining CQR with QIF in this study. Note that CQR is an estimation method for estimating the conditional mean of the response, by combining information from multiple quantile levels, instead of trying to estimate the conditional quantiles as its name might suggest.

Jiang et al. (2012) studied the weighted CQR (WCQR) for independent data. The motivation is that the use of the same weight at different quantile levels is generally not optimal. By minimizing the asymptotic variance of the WCQR estimator, the optimal weight can be obtained. The simulation results in Jiang et al. (2012) showed that WCQR has better performance than CQR for many error distributions. WCQR has also been applied to the double-threshold autoregressive conditional heteroscedastic (Jiang et al., 2014) and nonparametric regression models (Sun et al., 2013). However, for correlated data, whether the WCQR can still improve the estimation efficiency over CQR is unclear. To this end, on the basis of CQR and QIF (CQRQIF), we further propose the weighted CQRQIF (WCQRQIF) method for longitudinal data. Simulations demonstrated that WCQRQIF is better than CQRQIF for some error distributions.

The remainder of this paper is organized as follows. Section 2 investigates CQR combined with the QIF method and uses induced smoothing (Brown and Wang, 2005) to obtain the estimator in practice. Moreover, the large sample properties of both the CQRQIF estimator and the smoothing estimator are established. Section 3 develops WCQRQIF estimator to improve efficiency further. Section 4 presents numerical studies, including simulations and a real data analysis, to illustrate the performance of the proposed approaches. Section 5 gives some concluding remarks. Appendix contains the technical proofs.

2. CQR for correlated data

2.1. Model and estimation

We first briefly introduce CQR proposed in Zou and Yuan (2008), which deals with the standard linear regression problem

$$y_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + \epsilon_i,$$

where ϵ_i is the mean zero error independent of covariates $\boldsymbol{X}_i$. Although we are interested in estimating the mean regression function, it was noted in Zou and Yuan (2008) that the conditional quantile of y_i at level $\tau \in (0, 1)$ is $F^{-1}(\tau) + \boldsymbol{X}_i^T \boldsymbol{\beta}$, where F^{-1} is the quantile function for ϵ_i. Thus, performing a quantile regression will also produce an estimate of $\boldsymbol{\beta}$. Furthermore, it is natural to combine information from different quantile levels and solve the optimization problem

$$\min_{\boldsymbol{\beta}, e_k} \sum_{i=1}^{n} \sum_{k=1}^{q} \rho_{\tau_k}(y_i - e_k - \boldsymbol{X}_i^T \boldsymbol{\beta}),$$

where $\rho_{\tau_k}(u) = u(\tau_k - I(u < 0))$ is the check loss function, and $0 < \tau_1 < \tau_2 < \cdots < \tau_q < 1$ are q quantile levels. By combining multiple quantile levels, higher efficiency in estimation is achieved compared to a single-level quantile regression, while being more robust compared to least squares procedure.

Consider now the linear model for longitudinal data which is our focus in this study

$$y_{ij} = \boldsymbol{X}_{ij}^T \boldsymbol{\beta} + \epsilon_{ij}, \quad i = 1, \ldots, n,\ j = 1, \ldots, m_i, \tag{1}$$

where $\boldsymbol{X}_{ij} = (x_{ij1}, \ldots, x_{ijp})^T$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$, and ϵ_{ij} is the random error with mean zero, independent of covariates $\boldsymbol{X}_{ij}$. Notably, we ignored intercept in the model to reduce notation complexity slightly. In data analysis, the intercept is typically not of interest for mean regression.

Denote $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{im_i})^T$, $\boldsymbol{X}_i = (\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{im_i})^T$ and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{im_i})^T$, then model (1) can be written as

$$\boldsymbol{y}_i = \boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n. \tag{2}$$

Suppose that $\{\boldsymbol{X}_i, \boldsymbol{y}_i\}_{i=1}^n$ are generated independently from (2). According to the CQR approach (Zou and Yuan, 2008), if we ignore the correlation within subjects, we can estimate $\boldsymbol{\beta}$ from

$$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{e}}) = \operatorname{argmin}_{\boldsymbol{\beta}, e_1, \ldots, e_q} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \sum_{k=1}^{q} \rho_{\tau_k}(y_{ij} - e_k - \boldsymbol{X}_{ij}^T \boldsymbol{\beta}), \tag{3}$$