COMSTA: 6442 Model 3Gsc pp. 1–16 (col. fig: NIL)

ARTICLE IN PRESS

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

A moving average Cholesky factor model in covariance modeling for composite quantile regression with longitudinal data

Jing Lv [a], Chaohui Guo [b,*], Hu Yang [c], Yalian Li [c]

[a] School of Mathematics and Statistics, Southwest University, Chongqing, 400715, China
[b] College of Mathematics Science, Chongqing Normal University, Chongqing, 401331, China
[c] College of Mathematics and Statistics, Chongqing University, Chongqing, 401331, China

ARTICLE INFO

Article history:
Received 1 October 2016
Received in revised form 15 February 2017
Accepted 19 February 2017
Available online xxxx

Keywords:
Composite quantile regression
Longitudinal data
Modified Cholesky decomposition
Moving average
Robustness
Smoothed estimating equation

ABSTRACT

It is well known that the composite quantile regression is a very useful tool for regression analysis. In longitudinal studies, it requires a correct specification of the covariance structure to obtain efficient estimation of the regression coefficients. However, it is a challenging task to specify the correlation matrix in composite quantile regression with longitudinal data. In this paper, we develop a new regression model to parameterize covariance structures by utilizing the modified Cholesky decomposition. Then, based on the estimated covariance matrix, efficient composite quantile estimating functions are constructed to produce more efficient estimates. Since the proposed estimating functions are discrete and non-convex, we apply the induced smoothing approach to achieve fast and accurate estimation of the regression coefficients. Furthermore, we derive the asymptotic distributions of the parameter estimations both in mean and covariance models. Finally, simulations and a real data analysis have demonstrated the robustness and efficiency of the proposed approach.

1. Introduction

Longitudinal data arise commonly in many fields, such as biological and medical research, sociology and other fields. A significant characteristic of this data is that observations collected from the same subject at different times are correlated. Thus, it is a hot topic that how to make effectively use of the correlation to obtain efficient estimation of the regression coefficients. The well-known generalized estimating equation (GEE) is one of the most popular methods, which is proposed by Liang and Zeger (1986). Xie and Yang (2003) studied the asymptotic results of the GEE when either the number of sample size or cluster size goes to infinity. Balan and Schiopu-Kratina (2005) developed a theory of statistical inference of the GEE, which includes the weak consistency and asymptotic normality. Wang and Qu (2009) developed a novel Bayesian information criterion type model selection procedure by combining the GEE and quadratic inference function. Ueki (2009) proposed smooth-threshold estimating equations to select significance variables for longitudinal data. Wang et al. (2012) proposed new variable selection procedure for longitudinal data by combining the GEE with penalty functions. Xu et al. (2014) proposed a novel GEE-based screening procedure for longitudinal data. Other related methods can refer to Qu et al. (2000), Chaganty and Joe (2004), Li and Pan (2013) and so on.

* Corresponding author.
E-mail address: guochaohui2010@126.com (C. Guo).

http://dx.doi.org/10.1016/j.csda.2017.02.015
0167-9473/

Please cite this article in press as: Lv, J., et al., A moving average Cholesky factor model in covariance modeling for composite quantile regression with longitudinal data. Computational Statistics and Data Analysis (2017), http://dx.doi.org/10.1016/j.csda.2017.02.015

However, the GEE approach is very sensitive to outliers or heavy-tailed distributions. Quantile regression is a powerful alternative technique to analyze longitudinal data. Quantile regression not only can achieve robust estimators but also describe the entire conditional distribution of response variable. But it is a challenging work to estimate the correlation matrix in quantile regression with longitudinal data (Leng and Zhang, 2014; Fu et al., 2015). A naive approach is to estimate the regression coefficients by ignoring the correlations between repeated measurements, which results in a loss of efficiency. So far, some authors have made some contributions to improve estimation efficiency in quantile regression with longitudinal data. Under the assumption of exchangeable correlation structure, Fu and Wang (2012) combined the between and within subject estimating functions to incorporate the correlations between repeated measurements. Leng and Zhang (2014) applied the quadratic inference functions (Qu et al., 2000) to handle or incorporate the correlations and proposed more efficient unbiased estimating equations. Lu and Fan (2015) developed new estimating equations by utilizing the general stationary autocorrelation structure. Fu et al. (2015) proposed a method to estimate correlation parameters and chose the most appropriate working correlation matrix simultaneously based on the Gaussian pseudolikelihood approach. Fu and Wang (2016) constructed new unbiased estimating functions by combining the Gaussian copulas and empirical likelihood method (Owen, 2001). Other related literature includes Mu and Wei (2009), Wang and Zhu (2011) and Tang and Leng (2011).

Although quantile regression has many advantages, it may result in an arbitrarily small relative efficiency when compared with the least-squares regression. Recently, Zou and Yuan (2008) proposed a new regression method which is called composite quantile regression (CQR). The CQR not only possesses all merits of quantile regression but also provides estimation efficiency gain over a single quantile regression. Thus, it becomes a popular approach and has been extended to many fields, such as semi-parametric models, censored data and missing data. Tang et al. (2015) proposed weighted CQR estimators to improve estimation efficiency by combining the empirical likelihood and the quadratic inference function. So far, the CQR estimations with longitudinal data are not well studied in the literature.

Although the papers mentioned above had improved estimation efficiency when compared with the conventional quantile regression or composite quantile regression, they only focused on some specific correlation structures such as the AR(1), compound symmetry when estimating the correlation coefficient matrix. These approaches result in inefficient estimators when the true correlation structure is misspecified. Thus, it is still blank in the development of a general regression approach to model the covariance matrix in composite quantile regression with longitudinal data. Recently, the modified Cholesky decomposition has become a popular approach to parameterize the covariance matrix (Ye and Pan, 2006; Leng et al., 2010; Mao et al., 2011; Zhang and Leng, 2012; Yao and Li, 2013; Liu and Zhang, 2013; Liu and Li, 2015; Qin et al., 2016; Guo et al., 2016). The main merits of decomposition include the following aspects: (i) it guarantees the positive definiteness of estimated covariance matrix, (ii) the parameters of covariance matrix are related to well-founded statistical concepts. However, the literature mentioned above only focused on mean regression, and no one utilized it to research composite quantile regression. In this paper, we propose a new regression model to parameterize covariance structures in composite quantile regression with longitudinal data. Compared with existing quantile methods, the proposed approach not only can take the within correlations into consideration but also permit more general forms of the covariance structures. Thus the new method avoids to specify the working correlation structure to improve estimation efficiency, which is flexible and highly efficient.

The rest of this article is organized as follows. In Section 2, within the frame of independent working model assumption, we give detailed discussions on composite quantile regression. In Section 3, we utilize the modified Cholesky decomposition to decompose the covariance matrix as moving average coefficients and innovation variances. Then estimating equations are constructed to estimate the covariance parameters, and the corresponding theoretical properties are also established in this section. In Section 4, we develop efficient estimating equations for the regression coefficients based on the estimated covariance matrix. In Section 5, extensive simulation studies are carried out to investigate the performances of our method. In Section 6, the ophthalmology data set is used to illustrate the proposed method. In Section 7, we derive some conclusions and give some remarks. Finally, the proofs of the main results are provided in the Appendix.

2. Independent working model based on composite quantile regression

We consider the linear model

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

where y_{ij} is the jth measurement on the ith subject, $\boldsymbol{x}_{ij}$ is the p-vector covariates, $\boldsymbol{\beta}$ is $p \times 1$ unknown regression coefficient vector, and ε_{ij} is the random error. Note that $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{im_i})^T$ are correlated in the same subject but independent across the subjects. For independent data, when the slopes $\boldsymbol{\beta}$ are invariant across quantiles, Zou and Yuan (2008) proposed composite quantile regression, and the relative efficiency with respect to least squares regression is bounded below, and can be significantly more efficient when the errors are heavy tailed. Assume that $0 < \tau_1 < \tau_2 < \cdots < \tau_K < 1$. In particular, we use the equally spaced quantiles, i.e., $\tau_k = \frac{k}{K+1}$ for $k = 1, 2, \ldots, K$. Let $\rho_\tau(u) = u(\tau - I(u < 0))$ be the check function at $\tau \in (0, 1)$. Then, under an independent working model, we can estimate the regression coefficients $\boldsymbol{\beta}$ by the standard composite quantile regression

$$\left(\hat{b}_1, \ldots, \hat{b}_K, \hat{\boldsymbol{\beta}}^{CQR}\right) = \arg\min_{b_1,\ldots,b_K,\boldsymbol{\beta}} \sum_{k=1}^{K} \left\{ \sum_{i=1}^{n} \sum_{j=1}^{m_i} \rho_{\tau_k}\left(y_{ij} - b_k - \boldsymbol{x}_{ij}^T\boldsymbol{\beta}\right) \right\}, \tag{2}$$