Journal of Multivariate Analysis 143 (2016) 492–502

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Efficient parameter estimation via Gaussian copulas for quantile regression with longitudinal data

Liya Fu [a,c,*], You-Gan Wang [b,c]

[a] *School of Mathematics and Statistics, Xi'an Jiaotong University, China*
[b] *School of Mathematical Sciences, Queensland University of Technology, Australia*
[c] *Centre for Applications in Natural Resource Mathematics, School of Mathematics and Physics, The University of Queensland, Australia*

HIGHLIGHTS

- Copulas are constructed as working correlation matrices.
- Multiple unbiased estimating functions are combined via empirical likelihood.
- Induced smoothing approach is applied to reduce computation burdens.

ARTICLE INFO

Article history:
Received 11 May 2014
Available online 21 July 2015

AMS subject classifications:
62J99
62H99

Keywords:
Empirical likelihood
Gaussian copula
Induced smoothing
Longitudinal data
Quantile regression

ABSTRACT

Specifying a correlation matrix is challenging in quantile regression with longitudinal data. A naive method is simply to adopt an independence working model. However, the efficiency of parameter estimates may be lost. We propose constructing a working correlation matrix via Gaussian copula which can handle or incorporate general serial dependence. A suit of unbiased estimating functions can be obtained by assuming the Gaussian copula with different correlation matrices, and the empirical likelihood method can then combine these unbiased estimating functions. Furthermore, the induced smoothing approach is applied to the discontinuous estimating functions to reduce computation burdens. The asymptotic normality of the resulting estimators is established. Simulation studies indicate that the proposed method is superior to the alternative estimating functions especially when the working correlation matrix is misspecified. Finally, a real dataset from forced expiratory volume study is used to illustrate the proposed method.

1. Introduction

Longitudinal data arise commonly in medical research and many other fields. Correlation usually exists when data are collected from the same subject in these longitudinal studies [6]. Various methods have developed to evaluate covariate effects on the mean of a response variable while taking account of possible correlation patterns from longitudinal data [18,23,11]. In particular, Koenker and Bassett [14] proposed a novel quantile regression model to give a global assessment about covariate effects on the distribution of the response variable, which provides more complete description of the distribution and is more robust against outliers. Therefore, it has become a more widely used technique to describe the response distribution in recent years [5,13,9,24].

* Corresponding author at: School of Mathematics and Statistics, Xi'an Jiaotong University, China.
E-mail addresses: fuliya@mail.xjtu.edu.cn (L. Fu), you-gan.wang@uq.edu.au (Y-G. Wang).

http://dx.doi.org/10.1016/j.jmva.2015.07.004
0047-259X/

For independent measurements, parameter estimation and statistical inference procedures for quantile regression have been developed by Koenker and Bassett [14]. For longitudinal data, because the underlying correlation structure is difficult to describe and specify, a naive way is to use an independence working model [5,32,30], which is simple and has certain desirable properties. However, the efficiency of parameter estimation will be lost when strong correlations exist [27,8,16]. Koenker [12] considered a random effect model for quantile regression and made inferences by a penalized objective function. Geraci and Bottai [9] presented a likelihood-based approach by assuming the response variable following an asymmetric Laplace distribution. Tang and Leng [27] incorporated the within correlation by specifying a conditional mean model. Fu and Wang [8] introduced a combination of between- and within-subject estimating functions based on an exchangeable correlation structure assumption. Leng and Zhang [16] proposed constructing estimating functions by the quadratic inference method, which produces efficient estimates. In quantile regression, parameter estimates can be obtained using linear programming techniques [12]. The variances of parameter estimates typically depend on the unknown error distribution. To this end, a variety of resampling approaches have been proposed to estimate them [21,3,1]. However, intensive resamplings add computational burdens, while with very small number of resamplings, the estimating function may not be jittered enough to yield good estimates [32].

In this paper, we model the correlation structure in quantile regression with longitudinal data via copulas. A copula is a multivariate distribution with uniform marginal distribution on the interval $(0, 1)$, and it is a useful tool to study dependence of repeated measurements and to construct families of multivariate distributions [7]. Copulas have been explored in the statistics, econometrics, finance and insurance literature. For theories and examples of copulas, see Nelsen [19]. Because a Gaussian copula has a correlation structure and can handle general serial dependence, we utilize the Gaussian copula to explore the correlations in quantile regression with longitudinal data. We construct multiple unbiased estimating functions based on the working correlation matrices derived via the Gaussian copula with different correlation matrices, such as: exchangeable and AR(1). Furthermore, we combine these multiple estimating functions by the empirical likelihood method [22,20,17]. Because the estimating functions are discontinuous, we smooth them using the induced smoothing method [2]. The asymptotic properties of the objective functions and resulting estimators are derived under some regularity conditions given in the Appendix.

The remaining part of this paper is organized as follows: In Section 2, unbiased estimating functions are constructed based on working correlation matrices induced via Gaussian copulas, and then combine these estimating functions by the empirical likelihood method. In Section 3, extensive simulation studies are carried out to demonstrate the performance of the proposed method. In Section 4, the data from the forced expiratory volume in one second study [26] are used to illustrate the proposed method. Finally, some conclusions are given.

2. A new quantile regression estimator

Let $y_i = (y_{i1}, \ldots, y_{in_i})^{\mathrm{T}}$ denote the underlying outcomes for the ith subject, where $i = 1, \ldots, m$, and $X_i = (x_{i1}, \ldots, x_{in_i})^{\mathrm{T}}$ be the corresponding covariate vector, and x_{ik} is a $p \times 1$ vector. Assume that the 100τth percentile of y_{ik} is $x_{ik}^{\mathrm{T}}\beta_\tau$, that is $Q_\tau(y_{ik}|x_{ik}) = x_{ik}^{\mathrm{T}}\beta_\tau$, where β_τ is an unknown parameter vector, and $\epsilon_{ik} = y_{ik} - x_{ik}^{\mathrm{T}}\beta_\tau$, which is a continuous error term satisfying $p(\epsilon_{ik} \le 0) = \tau$ and with an unspecified density function $f_{ik}(\cdot)$. The median regression is obtained by taking $\tau = 0.5$. Here, what of interest is to find an efficient estimate of β_τ for a particular τ.

Under an independence working model, we estimate β_τ by minimizing the following objective function:

$$L_\tau(\beta) = \sum_{i=1}^{m}\sum_{k=1}^{n_i} \rho_\tau(y_{ik} - x_{ik}^{\mathrm{T}}\beta), \tag{1}$$

where $\rho_\tau(u) = u\{\tau - I(u \le 0)\}$, and $I(\cdot)$ is an indicator function [14]. Koenker and D'Orey [15] developed an efficient algorithm to optimize $L_\tau(\beta)$, which is available in statistical software R (package quantreg). The resulting estimates $\hat{\beta}_I$ from (1) can also be derived from the following estimating functions:

$$U_\tau(\beta) = \sum_{i=1}^{m} X_i^{\mathrm{T}} S_i(\beta) \tag{2}$$

where $S_i(\beta) = (S_{i1}, \ldots, S_{in_i})^{\mathrm{T}}$ and $S_{ik} = \tau - I(\epsilon_{ik} \le 0)$ is a Bernoulli distributed random noise variable [5].

The estimating functions $U_\tau(\beta)$ are based on the independence working model assumption, hence the efficiency of $\hat{\beta}_I$ could be improved if correlations of $S_i(\beta)$ are incorporated. To take account of the within correlations, Jung [10] proposed the following optimal estimating equation by the quasi-likelihood method

$$U_w(\beta) = \sum_{i=1}^{m} X_i^{\mathrm{T}} \Lambda_i A_i^{-1/2} R_i^{-1} A_i^{-1/2} S_i(\beta) = 0,$$

where $\Lambda_i = \mathrm{diag}(f_{i1}(0), \ldots, f_{in_i}(0))$, $A_i = \tau(1-\tau)I_{n_i}$, and R_i is the correlation matrix of $S_i(\beta)$. Here I_{n_i} is an $n_i \times n_i$ identity matrix. Diagonal matrix Λ_i involves the unknown density function $f_{ik}(\cdot)$. Although it is possible to estimate $f_{ik}(0)$ if some