Computational Statistics and Data Analysis 84 (2015) 41–53

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

A Gaussian pseudolikelihood approach for quantile regression with repeated measurements

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

[a] School of Mathematics and Statistics, Xi'an Jiaotong University, China

[b] Centre for Applications in Natural Resource Mathematics, School of Mathematics and Physics, The University of Queensland, St Lucia, QLD 4072, Australia

[c] Business School, Queensland University of Technology, Brisbane QLD 4001, Australia

ARTICLE INFO

Article history:
Received 3 June 2014
Received in revised form 31 October 2014
Accepted 6 November 2014
Available online 13 November 2014

Keywords:
Gaussian estimation
Induced smoothing method
Pseudolikelihood
Repeated measurements
Working covariance matrix

ABSTRACT

To enhance the efficiency of regression parameter estimation by modeling the correlation structure of correlated binary error terms in quantile regression with repeated measurements, we propose a Gaussian pseudolikelihood approach for estimating correlation parameters and selecting the most appropriate working correlation matrix simultaneously. The induced smoothing method is applied to estimate the covariance of the regression parameter estimates, which can bypass density estimation of the errors. Extensive numerical studies indicate that the proposed method performs well in selecting an accurate correlation structure and improving regression parameter estimation efficiency. The proposed method is further illustrated by analyzing a dental dataset.

1. Introduction

One characteristic of longitudinal data is that the measurements collected from the same subject are correlated (Diggle et al., 2002). To account for the correlations, Liang and Zeger (1986) developed the well-known generalized estimating equations (GEE) approach by incorporating a working correlation matrix. The GEE approach assures consistency of regression parameter estimators even when the correlation structure is misspecified. Widely used correlation structures include exchangeable, MA(1), and AR(1).

Quantile regression has also become a powerful alternative technique for analyzing repeated measurements, partly due to its flexibility and ability to describe the entire conditional distribution of a response variable (Koenker and D'Orey, 1987; Koenker, 2005). However, modeling the correlation in quantile regression for repeated measurements is challenging. A naive approach is to simply assume an independence working model (Chen et al., 2004; Yin and Cai, 2005; Wang and Zhu, 2011). This approach is simple and has some desirable properties but it may result in a great efficiency loss of parameter estimators when a high correlation exists (Tang and Leng, 2011; Fu and Wang, 2012; Leng and Zhang, 2014).

To improve the parameter estimates in quantile regression for repeated measurements, Jung (1996) introduced the quasi-likelihood method for median regression, which requires specifying and estimating a correlation matrix. Koenker (2004) considered a random effects model and carried out statistical inferences based on penalized L_1-statistics. Tang and Leng (2011) proposed a novel approach to incorporate the within-subject correlation. However, this approach requires

* Corresponding author.
E-mail address: you-gan.wang@uq.edu.au (Y.-G. Wang).

http://dx.doi.org/10.1016/j.csda.2014.11.002

specifying the conditional mean model. Fu and Wang (2012) presented a combination of between-subject and within-subject estimating functions under an exchangeable correlation structure assumption. Leng and Zhang (2014) combined sets of estimating functions based on distinct working correlation matrix by the quadratic inference function method to produce more efficient estimates. However, the parameter estimates derived from combined estimating functions are not guaranteed to perform as well as those derived from estimating equations with an accurate correlation structure (Westgate, 2014).

White (1961) and Crowder (1985) introduced the Gaussian estimation procedure which uses the Gaussian likelihood without assuming that the data are normally distributed. This method has good properties and has been investigated for mean regression with longitudinal data (Crowder, 2001; Wang and Zhao, 2007; Carey and Wang, 2011; Zhang and Paul, 2013). In this paper, we utilize the Gaussian pseudolikelihood approach to simultaneously estimate correlation parameters and select a working correlation structure for the correlated binary error terms in quantile regression with repeated measurements. This is achieved by estimating parameters and calculating the corresponding Gaussian pseudolikelihood for all the plausible candidate working correlation models. If one model is dominant based on the Gaussian pseudolikelihood criterion, the correlation structure and parameter estimates obtained from this procedure can probably be trusted for presentation. The proposed method is easy to implement. As demonstrated by the extensive simulation studies, the selected correlation structure by the proposed method is closest to the true structure in the sense that the corresponding estimates have the smallest mean squared errors. Furthermore, the induced smoothing method (Brown and Wang, 2005) is used to estimate the asymptotic covariance matrix of regression parameter estimates, which bypasses density function estimation of the errors and greatly reduces computational costs arising from other intensive resampling methods.

The rest of the paper is organized as follows: Section 2 presents the Gaussian pseudolikelihood approach. Extensive numerical studies are carried out in Section 3. Section 4 illustrates the use of the proposed method using a dental dataset. Section 5 presents conclusions.

2. Gaussian pseudolikelihood approach

Suppose that $y_i = (y_{i1}, \ldots, y_{in_i})^{\mathrm{T}}$ are measurements collected at times $(t_{i1}, \ldots, t_{in_i})$ for the ith subject, where $i = 1, \ldots, m$. Let $X_i = (x_{i1}, \ldots, x_{in_i})^{\mathrm{T}}$ be the corresponding covariate vector, where 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 parametric vector. Suppose that measurements from different subjects are independent, and those from the same subject are dependent. Let $\epsilon_{ik} = y_{ik} - x_{ik}^{\mathrm{T}}\beta_\tau$, which is a continuous error term satisfying $p(\epsilon_{ik} \le 0) = \tau$ and with an unspecified density function $f_{ik}(\cdot)$. What is of interest is to find an efficient estimate of β_τ for a particular τ.

Under an independence working model assumption, we can 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 (Koenker and Bassett, 1978). Koenker and D'Orey (1987) developed an efficient algorithm to optimize $L_\tau(\beta)$, which is available in the statistical software R (package quantreg).

2.1. Gaussian estimation

Define $Z_{ik} = I(y_{ik} \le x_{ik}^{\mathrm{T}}\beta)$ and let $Z_i = (Z_{i1}, \ldots, Z_{in_i})^{\mathrm{T}}$, hence Z_i is a correlated binary vector. Define $\mu_i = E(Z_i) = (\mu_{i1}, \ldots, \mu_{in_i})^{\mathrm{T}}$. The variance of Z_{ik} is $\mu_{ik}(1 - \mu_{ik})$. Define $A_i = \mathrm{diag}(\mu_{i1}(1 - \mu_{i1}), \ldots, \mu_{in_i}(1 - \mu_{in_i}))$. Suppose that $V_i = A_i^{1/2} R_i(\alpha) A_i^{1/2}$ is a working covariance matrix and α is a q-dimension parameter vector. The working Gaussian log-likelihood for $(Z_1, \ldots, Z_m)$ is

$$l(\alpha, \beta) = -\frac{1}{2} \sum_{i=1}^{m} \left[\log(|2\pi V_i|) + (Z_i - \mu_i)^{\mathrm{T}} V_i^{-1} (Z_i - \mu_i)\right].$$

The score function for correlation parameter α has the lth component

$$\frac{\partial l(\alpha, \beta)}{\partial \alpha_l} = \frac{1}{2} \mathrm{tr} \left\{ \sum_{i=1}^{m} [V_i^{-1}(Z_i - \mu_i)(Z_i - \mu_i)^{\mathrm{T}} - I_{n_i}] V_i^{-1} \frac{\partial V_i}{\partial \alpha_l} \right\}.$$

When V_i is appropriately specified, the score functions are unbiased. However, when V_i is misspecified, the unbiasedness is not satisfied because bias always exists when the parameter is used in a wrong family.

The regression parameter β can be estimated by the following estimating functions

$$\sum_{i=1}^{m} X_i^{\mathrm{T}} \Lambda_i A_i^{-1/2} R_i^{-1}(\alpha) A_i^{-1/2} \psi_i(\beta), \tag{2}$$

where $\Lambda_i = \mathrm{diag}(f_{i1}(0), \ldots, f_{in_i}(0))$, $A_i = \mathrm{diag}(\tau(1 - \tau), \ldots, \tau(1 - \tau))$, and $\psi_i(\beta) = (I(y_{i1} \le x_{i1}^{\mathrm{T}}\beta) - \tau, \ldots, I(y_{in_i} \le x_{in_i}^{\mathrm{T}}\beta) - \tau)^{\mathrm{T}}$. Although it is possible to estimate $f_{ik}(0)$ if some conditions for the error distributions are imposed, we will not