Computational Statistics and Data Analysis 84 (2015) 99–115

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

A semiparametric Bayesian approach for joint-quantile regression with clustered data

Woosung Jang [a], Huixia Judy Wang [b,*]

[a] SAS Institute Inc., 100 SAS Campus Drive, Cary, NC 27513, United States

[b] Department of Statistics, George Washington University, Washington, DC, 20052, United States

ARTICLE INFO

Article history:
Received 29 March 2014
Received in revised form 1 October 2014
Accepted 14 November 2014
Available online 25 November 2014

Keywords:
Generalized Pareto distribution
Markov chain Monte Carlo
Mixed model
Quantile regression
Random effects

ABSTRACT

Based on a semiparametric Bayesian framework, a joint-quantile regression method is developed for analyzing clustered data, where random effects are included to accommodate the intra-cluster dependence. Instead of posing any parametric distributional assumptions on the random errors, the proposed method approximates the central density by linearly interpolating the conditional quantile functions of the response at multiple quantiles and estimates the tail densities by adopting extreme value theory. Through joint-quantile modeling, the proposed algorithm can yield the joint posterior distribution of quantile coefficients at multiple quantiles and meanwhile avoid the quantile crossing issue. The finite sample performance of the proposed method is assessed through a simulation study and the analysis of an apnea duration data.

1. Introduction

Clustered data are commonly encountered in many areas of applications, for instance in medical studies with repeated measurements from the same individual and in educational studies with test scores of students in the same class. The key feature of clustered data is that measurements from the same cluster often have some common characteristics and thus tend to be correlated. One commonly used model-based analysis for analyzing clustered continuous data is the linear mixed model (Laird and Ware, 1982), where random effects are included to account for the within-cluster dependence.

Most linear mixed model analyses for clustered data also assume that both the random effects and random errors are normally distributed with constant variances. Such assumptions imply that the covariates affect only the location of the response distribution and thus cannot accommodate population heterogeneity. However, in some applications, the covariates may have different impacts at different locations of the response distribution. For instance, in a birth weight study, Abrevaya and Dahl (2008) found that some covariates such as the gender of the baby and the mother's prenatal-care visits have different effects at the lower and upper quantiles of the infant birth weight distribution. By focusing on the conditional quantiles, quantile regression (Koenker and Bassett, 1978) offers an alternative tool that can automatically capture the heterogeneity in covariate effects at different quantiles of the response distribution without modeling the heteroscedasticity.

Existing work for quantile regression with mixed effects is limited. The main challenges are that quantile regression usually does not make any parametric distributional assumptions, and unlike mean, quantiles are not additive, that is, quantiles of a sum of two random variables are often not the sum of their quantiles. To bypass these challenges, some researchers analyzed clustered data by considering marginal quantile regression models that treat the sum of random effects

* Corresponding author.
E-mail address: judywang@gwu.edu (H.J. Wang).

http://dx.doi.org/10.1016/j.csda.2014.11.008
0167-9473/

and random errors as a unit and focus on the covariate effects averaged over clusters; see for instance, Jung (1996), Wei and He (2006), Wang (2009), Mu and Wei (2009), Tang and Leng (2011), and Fu and Wang (2012).

For a conditional quantile regression model with a random intercept, Koenker (2004) proposed a regularization method, where a L_1 penalty is introduced to shrink the random effects towards a common value. Several authors proposed parametric or semiparametric Bayesian approaches for quantile regression with random effects. Geraci and Bottai (2007), Yuan and Yin (2010), Wang (2012) and Geraci and Bottai (2014) extended the asymmetric Laplace distribution idea in Yu and Moyeed (2001) for linear quantile regression to quantile regression with mixed effects, where the conditional distribution of the response is assumed to follow an asymmetric Laplace distribution. Instead of posing a parametric likelihood, Reich et al. (2010) proposed to model the likelihood nonparametrically by an infinite mixture of quantile-restricted two-component Gaussian mixtures and to accommodate error heteroscedasticity by specifying its form parametrically. Yang and He (2012) proposed the empirical likelihood as a working likelihood for Bayesian quantile regression for independent data and this method was extended to clustered data by Kim and Yang (2011). One commonality of these existing Bayesian methods is that the analysis (and modeling) is carried out at a single quantile level separately. Such separate analyses have two major limitations. First, the resulting estimated conditional quantiles are not guaranteed to be monotonically increasing in the quantile level and thus quantile crossing may be encountered. Second, for the single-quantile-analysis methods (e.g. Geraci and Bottai (2007), Yuan and Yin (2010), Geraci and Bottai (2014) and Reich et al. (2010)), the sampling distribution of the response at one quantile level is usually different from that at a different quantile level, and such inconsistency of likelihood makes it difficult to carry out inter-quantile analysis.

Using a semiparametric Bayesian framework, we propose a joint-quantile estimation method for quantile regression with random effects. Joint-quantile Bayesian analysis was also considered in Reich et al. (2011) for spatial data, where the quantile functions are modeled using basis functions, and in Tokdar and Kadane (2012) for independent data, where the quantile functions are modeled through logistic transformations of a smooth Gaussian process. Instead of making any parametric distributional assumptions on the random errors, we assume that the conditional quantiles of the response given covariates and random effects are linear. We propose to approximate the likelihood by linearly interpolating the quantile functions at multiple central quantiles and estimate the tail densities by adopting extreme value theory. A Metropolis-within-Gibbs algorithm is proposed to update fixed and random effects. With joint-quantile modeling, the proposed algorithm can avoid the quantile crossing problem, and yield the joint posterior distribution of quantile coefficients at multiple quantiles. Through simulation studies, we demonstrate that by approximating the likelihood through information-sharing across quantiles, the proposed method leads to more efficient multiple-quantile estimation than existing methods in finite samples.

2. The proposed method

2.1. Model setup

Suppose that we observe the clustered data $\{(y_{ij}, \mathbf{x}_{ij}, \mathbf{z}_{ij}),\ i = 1, \ldots, n, j = 1, \ldots, n_i\}$, where y_{ij} is the response, and $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ are respectively the p- and q-dimensional covariate vectors associated with cluster i for the jth subject (or measuring time). We assume the following conditional quantile regression model

$$Q_\tau(Y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_i) = \mathbf{x}_{ij}^T\boldsymbol{\beta}_\tau + \mathbf{z}_{ij}^T\mathbf{b}_i, \quad 0 < \tau < 1, \tag{2.1}$$

where $Q_\tau(Y_{ij}|\mathbf{x}_{ij}, \mathbf{z}_{ij}, \mathbf{b}_i)$ is the τth conditional quantile of the response given covariates and the random cluster effects $\mathbf{b}_i$, $\boldsymbol{\beta}_\tau$ is a $p \times 1$ vector of fixed effects, $\mathbf{z}_{ij}$ is a $q \times 1$ covariate vector associated with the cluster-specific random effects $\mathbf{b}_i$. By including the random effects $\mathbf{b}_i$, the conditional quantile regression model (2.1) captures the cluster-specific effects of covariates on the conditional quantile of the response distribution. The fixed effect $\boldsymbol{\beta}_\tau$ is allowed to vary with the quantile level but we assume that the random effects $\mathbf{b}_i$ are the same across all quantile levels and that $\mathbf{b}_i \overset{\text{i.i.d.}}{\sim} N(0, \Sigma)$.

Instead of making any parametric assumptions on the conditional distribution of Y_{ij}, we assume in model (2.1) that the conditional quantile functions are linear for $\tau \in (0, 1)$. Under this global linearity assumption, we propose a Bayesian approach based on an approximate likelihood for regression at multiple quantiles for clustered data. The idea of the approximate likelihood is to estimate the conditional density of Y_{ij} by linearly interpolating the conditional quantiles $Q_\tau(Y_{ij}|\mathbf{x}_{ij}, \mathbf{b}_i)$ for a sequence of central quantiles τ, and to estimate the tail density by using extreme value theory.

Before presenting the proposed procedure for clustered data, we first introduce the proposed approximate likelihood method for quantile regression with independent data.

2.2. Approximate likelihood for independent data

Let $\{(y_i, \mathbf{x}_i), i = 1, \ldots, n\}$ be a random sample of $(Y, \mathbf{X})$, where Y is the response variable and $\mathbf{X}$ is the p-dimensional design vector with the first element being one. Suppose the following linear quantile regression model holds for any $\tau \in (0, 1)$,

$$Q_\tau(Y|\mathbf{X} = \mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta}_\tau. \tag{2.2}$$