Finite mixtures of linear quantile regressions with concomitant variables: a simple solution to endogeneity in longitudinal data models

Marco Alfò[1], Maria Francesca Marino[2], and Francesca Martella[1]

[1] Sapienza Università di Roma, Piazzale Aldo Moro 5, 00185 Roma, ITALY
marco.alfo@uniroma1.it, francesca.martella@uniroma1.it
[2] Università degli Studi di Firenze, Viale Morgagni 59, 50134 Firenze, ITALY
mariafrancesca.marino@unifi.it

Abstract. Longitudinal data often give the chance to control for time-constant heterogeneity, which is added to the model formulation via individual-specific effects. Adopting a random effect specification, issues of endogeneity may arise. We discuss quantile regression models for longitudinal data and propose a concomitant variable framework to address endogeneity. Specifically, we assume that mixing proportions are unknown and depend on time-constant covariates, as well as on time-constant levels of time-varying covariates. A multinomial logit specification is considered to model the relation between such proportions and the (potentially) endogenous covariates. This provides a simple, efficient, and general solution to the aforementioned problem. The performance of the proposed model is examined using a simulation study. The results are promising and warrant additional discussion.

Keywords: Endogenous covariates, panel data, clustered observations, random effects, unobserved heterogeneity.

1 Introduction

In many instances of practical applications, dependent observations with a hierarchical structure may be encountered. Such a structure may originate from several sources, such as spatial configurations, multilevel frameworks, or longitudinal sampling designs. In the framework of regression models for longitudinal data, it is often essential to take into account the potential dependence between observations from the same individual and the potential heterogeneity between individuals participating in the sample. These two sources of extra-model variation may be due to time-constant features that are specific to sampled individuals, as well as to serial correlation between measurements recorded at different time points. See, e.g., Fitzmaurice et. al [6] for a general discussion of this and related issues. If the dependence is not adequately addressed, and potential heterogeneity is not inserted in the model specification, the model parameter estimates could exhibit significant bias or even be inconsistent. Individual-specific intercepts are often included in a regression framework to address the effect of

unobserved heterogeneity. Often, these are treated as random variables with a known, parametric (Gaussian), distribution, and they are integrated out of the likelihood to get parameter estimates. Although there is a widely accepted consensus on this objective, the naive random effect approach has two assumptions that often faced criticisms. First, the specification of a parametric (Gaussian) distribution for the random effects is generally challenging to be assessed based solely on the observed data. Further, in the presence of non-Gaussian data, the computational burden for numerical integration or related approximations may be substantial. Second, a frequent assumption is that of exogeneity of the random effects. They are assumed to capture the effect of omitted covariates on the response and often, they are assumed to convey non-overlapping signals with respect to the covariates already in the model. This is translated into linear or stochastic independence between observed (covariates) and unobserved (individual-specific effects) heterogeneity.

This paper aims to address the issue of endogeneity in the framework of linear quantile regression models for longitudinal data. These represent a valuable tool of analysis thanks to their robustness to asymmetry and outliers in the data, as well as to their ability in capturing the effects that covariates may have on different quantiles of the (conditional) response distribution. Up to our knowledge, only a few attempts have been proposed to deal with endogeneous covariates in the longitudinal quantile regression framework. Most of these fall within the fixed effect framework [11, 7, 13, 8], and, in turn, suffer from the so called incidental parameter problem [20]. Further, they do not allow for the estimation of the effect that time-invariant covariates may have on the response (conditional) quantiles. The correlated random effect estimators proposed by Abrevaya and Dahl, [1] and revisited by Bache et. al, [3] represent an interesting alternative. Here, a set of sufficient covariates, derived from time-varying endogenous covariates, is included in the model specification to obtain unbiased estimates for the parameters of interest.

In this paper, we try to address endogeneity by extending the mixture of linear quantile regression models proposed by Alfó et. al [2]. That is, we include individual-specific, time-constant, random effects in the model for the (conditional) response quantiles; these are treated as random variables, with an unspecified distribution that is estimated from the data via a nonparametric maximum likelihood (NLPM) approach, as discussed by [12, 15, 14]. Such an approach leads to a discrete estimate of the mixing distribution, which is both straightforward to handle and exhibits robustness against deviations from the standard Gaussian assumption on the random effects. On the other hand, we take advantage from the finite mixture specification, and we handle the dependence between the observed and unobserved heterogeneity by explicitly modeling the influence that the possible endogenous covariates have on the random effect distribution, through the corresponding mixing proportions. These are modeled via a multinomial logit model, as in mixtures of regression models in the presence of concomitant variables [4]. This parameterization offers a clear interpretation while accounting for general forms of dependence between the random effects

and the observed covariates. Model parameter estimation is achieved through maximum likelihood via an EM algorithm [5].

2 Finite mixtures of linear quantile regression models

Let us start considering a *continuous* longitudinal response Y_{it} observed for individuals $i = 1, \ldots, n$, at time points $t = 1, \ldots, T_i$. Further, let $\mathbf{x}_{it}$ denote a p-dimensional vector of covariates. We are interested in modeling the effect that these covariates have on different portions of the (conditional) response variable distribution. For this purpose, a regression model for the τ-th (conditional) quantile of Y_{it}, $\tau \in (0, 1)$, is defined as

$$\mathcal{Q}_\tau \left(y_{it} \mid \mathbf{x}_{it}, \alpha_{i\tau} \right) = \mathbf{x}_{it}' \boldsymbol{\beta}_\tau + \alpha_{i\tau}, \tag{1}$$

or, alternatively, as

$$Y_{it} = \mathbf{x}_{it}' \boldsymbol{\beta}_\tau + \alpha_{i\tau} + e_{it\tau},$$

under the constraint that $\mathcal{Q}_\tau(e_{it\tau} \mid \alpha_{i\tau}, \mathbf{x}_{it}) = \mathcal{Q}_\tau(e_{it\tau}) = 0$. The parameter $\boldsymbol{\beta}_\tau$ appearing in the above equations denotes a p-dimensional and τ-dependent parameter vector which summarizes the relation between the observed covariates in $\mathbf{x}$ and the τ-th (conditional) quantile of the response. On the other hand, terms $\alpha_{i\tau}$ represent the effect of *unobserved* time-constant, individual-specific, features that have not been considered in the design vector. In this framework, we assume that they are realizations of a random variable with density $g_\alpha(\cdot)$. Lastly, $e_{it\tau}$ is an error term assumed to be independent of both $\mathbf{x}_{it}$ and $\alpha_{i\tau}$. Thus, for a fixed τ, the dependence between observations from the same individual $i = 1, \ldots, n$, recorded at the different occasions $t = 1, \ldots, T_i$, arises from the shared common parameter $\alpha_{i\tau}$. Conditional on the individual-specific terms, repeated measurements recorded from the same individual become independent.

Regarding the conditional response variable distribution $f_Y \left(y_{it} \mid \mathbf{x}_{it}, \alpha_{i\tau}, \tau \right)$, a frequent assumption within the framework of parametric linear quantile regression models is that of a conditional Asymmetric Laplace (AL) density [22, 10, 9, 2, 16, 17]. The corresponding location parameter is modeled according to equation (1). The assumption of (conditional) AL responses is ancillary and it simply allows us to recast model parameter estimates within a maximum likelihood framework. Based on the above assumptions, the marginal log-likelihood is

$$\begin{aligned} \ell \left(\cdot \right) &= \sum_{i=1}^{n} \log \left\{ \int_{\mathcal{A}} \left[\prod_{t=1}^{T_i} f_Y \left(y_{it} \mid \mathbf{x}_{it}, \alpha_{i\tau}, \tau \right) \right] g_\alpha(\alpha_{i\tau} \mid \mathbf{X}_i) \mathrm{d}\alpha_{i\tau} \right\} \\ &= \sum_{i=1}^{n} \log \left\{ \int_{\mathcal{A}} f_Y \left(\mathbf{y}_i \mid \mathbf{X}_i, \alpha_{i\tau}, \tau \right) g_\alpha(\alpha_{i\tau} \mid \mathbf{X}_i) \mathrm{d}\alpha_{i\tau} \right\}, \end{aligned} \tag{2}$$

where $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT_i})'$, $\mathbf{X}_i$ is the $(T_i \times p)$-dimensional matrix of covariates recorded from the i-th individual, and and $\mathcal{A}$ denotes the support for the distribution of the random effects.

As it can be easily noticed, even if a known parametric density is chosen for the individual-specific effects $g_\alpha(\cdot)$, the previous integral for the log-likelihood contains a conditional density, $g_\alpha(\cdot \mid \mathbf{X}_i)$, which should be handled properly. A simple and typically adopted solution is based on assuming $g_\alpha(\cdot \mid \mathbf{X}_i) = g_\alpha(\cdot)$. Moreover, parametric assumptions on the random effect distribution cannot be directly tested in practical applications. Many researchers in the statistical field have investigated into this topic without finding a clear-cut solution [18, 21].

To address this issue, we opt for the nonparametric approach proposed, in the linear quantile regression context, by Alfo' et. al [2]. The idea is to approximate $g_\alpha(\cdot)$ by a discrete distribution that spans $K \leq m$ locations, $\{\zeta_{1\tau}, \ldots, \zeta_{K\tau}\}$, with associated masses defined by $\pi_{k\tau} = \Pr(\alpha_{i\tau} = \zeta_{k\tau})$, for $i = 1, \ldots, n$, and $k = 1, \ldots, K$. Here, m denotes the number of different individual profiles in the sample, which represents a bound for the total number of locations. That is, we assume that

$$g_\alpha(\alpha_{i\tau}) \sim \sum_{k=1}^{K} \pi_{k\tau} \delta_{\zeta_{k\tau}},$$

where $\delta_{\zeta_{k\tau}}$ is a one-point distribution putting a unit mass at $\zeta_{k\tau}$. Both locations and masses are directly estimated from the observed data (together with the remaining model parameters) by maximizing the approximation of the log-likelihood in the following equation

$$\ell(\cdot) \simeq \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_{k\tau} f_Y(\mathbf{y}_i \mid \mathbf{X}_i, \alpha_{i\tau} = \zeta_{k\tau}, \tau). \tag{3}$$

As it can be easily noticed, equation (3) resembles the likelihood function for a finite mixture of linear quantile regression models, where $f_Y(\mathbf{y}_i \mid \mathbf{X}_i, \alpha_{i\tau} = \zeta_{k\tau}, \tau)$ is defined as the product of T_i (conditional) AL densities for an individual i coming from the k-th mixture component, $i = 1, \ldots, n$, and $k = 1, \ldots, K$. In line with standard practice in finite mixture models, parameter estimation involves maximizing the log-likelihood in equation (3) (once K and τ have been fixed) through an EM algorithm.

3 Dealing with endogeneity

As discussed in the previous section, a frequent assumption when dealing with random effect models is that of exogeneity of the observed covariates on the random effects: $g_\alpha(\cdot \mid \mathbf{X}_i) = g_\alpha(\cdot)$. This assumption, however, rarely holds. In fact, it reduces to assuming that observed and unobserved covariates are independent, and this is often not true. In fact, $\alpha_{i\tau}$ is meant to represent the effect of omitted, time-invariant, covariates on the τ-th quantile of the individual (conditional) response distribution, and these are typically correlated with the observed ones. To address issues related to possible endogeneity of observed covariates, we extend the proposal by Alfo' et al. [2]. The idea is not to assume $g_\alpha(\alpha_{i\tau} \mid \mathbf{X}_i) = g_\alpha(\alpha_{i\tau})$ and let the (approximate) discrete distribution of the

random effects to depend on the possible endogeneous covariates. As the number of locations of this distribution (K) is bounded above by the number of distint individual profiles in the sample (m), we assume that the covariates influence the masses associated with the locations, but not the locations themselves. In other words, we assume

$$g_\alpha(\alpha_{i\tau} \mid \mathbf{X}_i) \sim \sum_{k=1}^{K} \pi_{k\tau}(\mathbf{X}_i)\, \delta_{\zeta_{k\tau}}.$$

In the spirit of Mundlak [19] for the mean regression framework and Backe et. al [3] for the quantile framework, we denote by $\bar{\mathbf{x}}_i$ the individual-specific vector of covariate means; the mixture prior weights are then modeled by means of the following multinomial logit specification:

$$\pi_{k\tau}(\mathbf{X}_i) = \frac{\exp(\bar{\mathbf{x}}_i'\boldsymbol{\eta}_{k\tau})}{1 + \sum_{h=2}^{K} \exp(\bar{\mathbf{x}}_i'\boldsymbol{\eta}_{h\tau})} \tag{4}$$

as in mixtures of regressions with concomitant variables [4]. It is also worth noticing that, in this parameterization, we can include not only $\bar{\mathbf{x}}_i$ but also any time-invariant variable, $\mathbf{w}_{it} = \mathbf{w}_i \in \mathbf{x}_{it}$, available in the dataset by substituting $[\bar{\mathbf{x}}_i, \mathbf{w}_i]$ to $\bar{\mathbf{x}}_i$ in equation (4).

The simulation findings indicate that the proposed approach consistently performs well across various scenarios, demonstrating favorable results in terms of both bias and mean squared error (MSE) of the estimated model parameters. This leads us to conclude that it can be effectively applied for the analysis of real datasets entailing the social, the behavioral, as well as the bio-medical field.

References

[1] Jason Abrevaya and Christian M Dahl. "The Effects of Birth Inputs on Birthweight: Evidence from Quantile Estimation on Panel Data". In: *Journal of Business and Economic Statistics* 26 (2008), pp. 379–397.

[2] M. Alfó, N. Salvati, and M. G. Ranalli. "Finite Mixtures of Quantiles and M-quantile Models". In: *Statistics and Computing* 27 (2017), pp. 547–570.

[3] Stefan Holst Milton Bache, Christian Møller Dahl, and Johannes Tang Kristensen. "Headlights on tobacco road to low birthweight outcomes: Evidence from a battery of quantile regression estimators and a heterogeneous panel". In: *Empirical Economics* 44 (2013), pp. 1593–1633.

[4] C. Mitchell Dayton and George B. Macready. "Concomitant-Variable Latent-Class Models". In: *Journal of the American Statistical Association* 83.401 (1988), pp. 173–178. (Visited on 02/15/2024).

[5] A.P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum Likelihood from Incomplete Data Via the EM Algorithm". In: *Journal of the Royal Statistical Society B* 39 (1977), pp. 1–38.

[6] G. M. Fitzmaurice, N. M. Laird, and J. H. Ware. *Applied longitudinal analysis*. John Wiley & Sons, 2004.

[7] A. F. Galvao. "Quantile Regression for Dynamic Panel Data with Fixed Effects". In: *Journal of Econometrics* 164 (2011), pp. 142–157.
[8] A. F. Galvao and G. V. Montes-Rojas. "Penalized Quantile Regression for Dynamic Panel Data". In: *J. Statist. Plann. Inference* 140 (2010), pp. 3476–3497.
[9] M. Geraci and M. Bottai. "Linear Quantile Mixed Models". In: *Statistics and Computing* 24 (2014), pp. 461–479.
[10] M. Geraci and M. Bottai. "Quantile Regression for Longitudinal Data Using the Asymmetric Laplace Distribution". In: *Biostatistics* 8 (2007), pp. 140–54.
[11] R. Koenker. "Quantile Regression for Longitudinal Data". In: *Journal of Multivariate Analysis* 91 (2004), pp. 74–89.
[12] N. Laird. "Nonparametric Maximum Likelihood Estimation of a Mixing Distribution". In: *Journal of the American Statistical Association* 73 (1978), pp. 805–811.
[13] C Lamarche. "Robust Penalized Quantile Regression Estimation for Panel Data". In: *J. Econometrics* 157 (2010), pp. 396–408.
[14] B. G. Lindsay. "The Geometry of Mixture Likelihoods, Part II: the Exponential Family". In: *The Annals of Statistics* 11 (1983), pp. 783–792.
[15] B. G. Lindsay. "The Geometry of Mixture Likelihoods: a General Theory". In: *The Annals of Statistics* 11 (1983), pp. 86–94.
[16] Maria Francesca Marino and Marco Alfó. "Latent drop-out based transitions in linear quantile hidden Markov models for longitudinal responses with attrition". In: *Advances in Data Analysis and Classification* 9 (2015), pp. 483–502.
[17] Maria Francesca Marino, Nikos Tzavidis, and Marco Alfò. "Mixed Hidden Markov Quantile Regression Models for Longitudinal Data with Possibly Incomplete Sequences". In: *Statistical Methods in Medical Research* 27 (2018), pp. 2231–2246.
[18] Charles E McCulloch and John M Neuhaus. "Misspecifying the Shape of a Random Effects Distribution: Why Getting It Wrong May Not Matter". In: *Statistical Science* 26.3 (2011), pp. 388–402.
[19] Yair Mundlak. "On the pooling of time series and cross section data". In: *Econometrica: journal of the Econometric Society* (1978), pp. 69–85.
[20] Jerzy Neyman and Elizabeth L Scott. "Consistent estimates based on partially consistent observations". In: *Econometrica: Journal of the Econometric Society* (1948), pp. 1–32.
[21] Dimitris Rizopoulos, Geert Verbeke, and Geert Molenberghs. "Shared parameter models under random effects misspecification". In: *Biometrika* 95.1 (2008), pp. 63–74.
[22] K. Yu and R. A. Moyeed. "Bayesian Quantile Regression". In: *Statistics and Probability Letters* 54 (2001), pp. 437–447.