International Journal of Trend in Scientific Research and Development (IJTSRD)

International Open Access Journal

ISSN No: 2456 - 6470 | www.ijtsrd.com | Volume - 2 | Issue – 4

Data Imputation Methods and Technologies

Ritesh Kumar Pandey
M. Tech Scholar, Dept. of CSE, Kalinga University,
Naya Raipur, Chhattisgarh, India

Dr Asha Ambhaikar
Ph.D., Dept. of CSE, Kalinga University,
Naya Raipur, Chhattisgarh, India

ABSTRACT

We introduce a class of linear quantile regression estimators for panel data. Our framework contains dynamic autoregressive models, models with general predetermined regressors, and models with multiple individual effects as special cases. We follow a correlated random-effects approach, and rely on additional layers of quantile regressions as a flexible tool to model conditional distributions. Conditions are given under which the model is nonparametrically identified in static or Markovian dynamic models. We develop a sequential method-of-moment approach for estimation, and compute the estimator using an iterative algorithm that exploits the computational simplicity of ordinary quantile regression in each iteration step. Finally, a Monte-Carlo exercise and an application to measure the effect of smoking during pregnancy on children's birthweights complete the paper.

K-means and K-medoids clustering algorithms are widely used for many practical applications. Original k-mean and k-medoids algorithms select initial centroids and medoids randomly that affect the quality of the resulting clusters and sometimes it generates unstable and empty clusters which are meaningless. The original k-means and k-mediods algorithm is computationally expensive and requires time proportional to the product of the number of data items, number of clusters and the number of iterations. The new approach for the k mean algorithm eliminates the deficiency of exiting k mean. It first calculates the initial centroids k as per requirements of users and then gives better, effective and stable cluster. It also takes less execution time because it eliminates unnecessary distance computation by using previous iteration. The new approach for k- medoids selects initial k medoids systematically based on initial centroids. It generates stable clusters to improve accuracy.

Keywords: *Panel data, quantile regression, expectation-Maximization*

INTRODUCTION

Nonlinear panel data models are central to applied research. However, despite some recent progress, it is fair to say that we are still short of answers for panel versions of many models commonly used in empirical work.[1] In this paper we focus on one particular nonlinear model for panel data: quantile regression.

Since Koenker and Bassett (1978), quantile regression has become a prominent methodol-ogy for examining the effects of explanatory variables across the entire outcome distribution. Extending the quantile regression approach to panel data has proven challenging, however, mostly because of the difficulty to handle individual-specific heterogeneity. Starting with Koenker (2004), most panel data approaches to date proceed in a quantile-by-quantile fash-ion, and include individual dummies as additional covariates in the quantile regression. As shown by some recent work, however, this fixed-effects approach faces special challenges when applied to quantile regression. Galvao, Kato and Montes-Rojas (2012) develop the large-N, T analysis of the fixed-effects quantile regression estimator, and show that it may suffer from large asymptotic biases. Rosen (2010) shows that the fixed-effects model for a single quantile is not point-identified.[2]

$$Q\ (Y_{it} \mid X_i, \eta_i, \tau\ ) = X_{it}{}'\beta\ (\tau\ ) + \eta_i\gamma\ (\tau\ )\ ,$$

$$\text{for all } \tau \in (0, 1). \qquad (1)$$

We depart from the previous literature by proposing a random-effects approach for quantile models from panel data. This approach treats individual unobserved heterogeneity as time-invariant missing data. To describe the model, let i = 1, ..., N denote individual units, and let t = 1, ..., T denote time periods. The random-effects quantile regression (REQR) model specifies the τ -specific conditional quantile of an outcome variable Y_{it}, given a se-quence of strictly exogenous covariates $X_i = (X_{i}{}'_1, ..., X_{iT}{}'\ )'$ and unobserved heterogeneity η_i, as follows:

Note that η_i does not depend on the percentile value τ. Were data on η_i available, one could use a standard quantile regression package to recover the parameters β (τ) and γ (τ).

Model (1) specifies the conditional distribution of Y_{it} given X_{it} and η_i. In order to complete the model, we also specify the conditional distribution of η_i given the sequence of covariates X_i. For this purpose, we introduce an additional layer of quantile regression and specify the τ -th conditional quantile of η_i given covariates as follows:

This modelling allows for a flexible conditioning on strictly exogenous regressors—and on initial conditions in dynamic settings—that may also be of interest in other panel data models. Together, equations (1)-(2) provide a fully specified semiparametric model for the joint distribution of outcomes given the sequence of strictly exogenous covariates. The aim is then to recover the model's parameters: β (τ), γ (τ), and δ (τ), for all τ

Our identification result for the REQR model is nonparametric. In particular, identification holds even if the conditional distribution of individual effects is left unrestricted. Recent research has emphasized the identification content of nonlinear panel data models with continuous outcomes (Bonhomme, 2012), as opposed to discrete outcomes models where parameters of interest are typically set-identified (Honor´e and Tamer, 2006, Chernozhukov, Fern´andez-Val, Hahn and Newey, 2011). Pursuing this line of research, our analysis provides conditions for nonparametric identification of REQR in panels where the number of time periods T is fixed, possibly very short (e.g., T = 3). One of the required conditions to apply Hu and Schennach (2008)'s result is a completeness assumption. Although completeness is a high-level assumption, recent papers have provided primitive conditions in specific models, including a special case of model (1).[3]

$$Q\ (\eta_i \mid X_i, \tau\ ) = X_i{}'\delta(\tau\ ), \quad \text{for all } \tau \in (0, 1).$$

Our analysis is most closely related to Wei and Carroll (2009), who proposed a con-sistent estimation method for cross-sectional linear quantile regression subject to covariate measurement error. In particular, we rely on the approach in Wei and Carroll to deal with the continuum of model parameters indexed by τ ∈ (0, 1). As keeping track of all parameters in the algorithm is not feasible, we build on their insight and use interpolating splines to combine the quantile-specific parameters in (1)-(2) into a complete likelihood function that depends on a finite number of parameters. Our proof of consistency—in a panel data asymp-totics where N tends to infinity and T is kept fixed—also builds on theirs. As the sample size increases, the number of knots, and hence the accuracy of the spline approximation, increase as well. A key difference with Wei and Carroll is that, in our setup, the conditional distribution of individual effects is unknown, and needs to be estimated along with the other parameters of the model.

2. Model and identification

In this section and the next we focus on the static version of the random-effects quantile regression (REQR) model. Section 6 will consider various extensions to dynamic models. We start by presenting the model along with several examples, and then provide conditions for nonparametric identification.

2.1.**Model**

Let $Y_i = (Y_{i1}, ..., Y_{iT}\ )'$ denote a sequence of T scalar outcomes for individual i, and let $X_i = (X_{i}{}'_1, ..., X_{iT}{}'\ )'$ denote a sequence of strictly exogenous regressors, which may contain a constant. In addition, let η_i denote a q-dimensional vector of individual-specific effects, and let U_{it} denote a scalar error term. The model specifies the conditional quantile response function of Y_{it} given X_{it} and η_i as follows:

$Y_{it} = Q_Y (X_{it}, \eta_i, U_{it})$

$i = 1, ..., N,$

$t = 1, ..., T$..............................(3)

We make the following assumptions.

Assumption 1 (outcomes)

(i) U_{it} follows a standard uniform distribution conditional on X_i and η_i.

(ii) τ 7→Q (x, η, τ) is strictly

increasing on (0, 1), almost surely in (x, η).

(iii) U_{it} is independent of U_{is} for each t =6 s conditional on X_i and η_i.

Assumption 1 (i) contains two parts. First, U_{it} is assumed independent of the full se-quence X_{i1}, ..., X_{IT} and independent of individual effects. This assumption of strict exo-geneity rules out predetermined or endogenous covariates. Second, the marginal distribution of U_{it} is normalized to be uniform on the unit interval. Part (ii) guarantees that outcomes.

2.2 Identification

In this section we study nonparametric identification in model (3)-(4). We start with the case where there is a single scalar individual effect (i.e., q = dim η_i = 1), and we set T = 3.

Under conditional independence over time—Assumption 1 (iii)—we have, for all y_1, y_2, y_3,

x = $(x'_1, x'_2, x'_3)'$, and η:

$fY_1,Y_2,Y_3|\eta,X$ (y1, y2, y3 | η, x) = $fY_1|\eta,X$ (y1 | η, x) $fY_2|\eta,X$ (y2 | η, x) $fY_3|\eta,X$ (y3 | η, x) ....(4)

Hence the data distribution function relates to the densities of interest as follows: **Z**

$fY_1,Y_2,Y_3|X$
(y1, y2, y3 | x)$fY_1|\eta,X$ (y1 | η, x) $fY_2|\eta,X$ (y2 | η,
= x) $fY_3|\eta,X$ (y3 | η, x)

$\times f_{\eta|X}$ (η | x) dη...... (5)

The goal is the identification of $f_{Y1|\eta,X}$, $f_{Y2|\eta,X}$, $f_{Y3|\eta,X}$ and $f_{\eta|X}$ given knowledge of $f_{Y1,Y2,Y3|X}$. The setting of equation (5) is formally equivalent (conditional on x) to the instrumental variables setup of Hu and Schennach (2008), for nonclassical nonlinear errors-in-variables models. Specifically, according to Hu and Schennach's terminology Y_3 would be the outcome variable, Y_2 would be the mismeasured regressor, Y_1 would be the instrumental variable, and

η would be the latent, error-free regressor. We closely rely on their analysis and make the following additional assumptions.

3. REQR estimation

This section considers estimation in the static model (6)-(7). We start by describing the moment restrictions that our estimator exploits, and then present the sequential estimator. In the next two sections we will study the asymptotic properties of the estimator and discuss implementation issues in turn.

3.1.Moment restrictions

The check function ρ_τ , which is familiar from the quantile regression literature (Koenker and Basset, 1978): ρ_τ (u) = (τ − 1 {u < 0}) u, and ψ_τ (u) = ∇ρ(u). Let also W_{it} (η) = $(X_{it}', \eta)'$.

In order to derive the main moment restrictions, we start by noting that, for all τ ∈ (0, 1), the following infeasible moment restrictions hold, as a direct implication of Assumptions 1

Indeed, (6) is the first-order condition associated with the infeasible population quantile regression of Y_{it} on X_{it} and η_i. Similarly, (5) corresponds to the infeasible quantile regression of η_i on X_i.

4. CONCLUSION

Random-effects quantile regression (REQR) provides a flexible approach to model nonlinear panel data models. In our approach, quantile regression is used as a versatile tool to model the dependence between individual effects and exogenous regressors or initial conditions, and to model feedback processes in models with

and 2: " $_{t=1}$ Wit (ηi) ψτ
Yit − Wit $(\eta i)'$ θ
(τ) # = 0, (6)

$$E\,[X_i\psi_\tau\ (\eta_i - X_i'\delta\,(\tau\,))] = 0. \qquad (7)$$

Predetermined covariates. The empirical application illustrates the benefits of having a flexible approach to allow for heterogeneity and nonlinearity within the same model in a panel data context.

The analysis of the asymptotic properties of the REQR estimator requires an approxima-tion argument. However, while our consistency proof allows the quality of the approximation to increase with the sample size, at this stage in our characterization of the asymptotic distribution we keep the number of knots L fixed as the number of observations N increases. Assessing the asymptotic behavior of the quantile estimates as both L and N tend to infinity is an important task for future work.

Lastly, note that our quantile-based modelling of the distribution of individual effects could be of interest in other models as well. For example, one could consider semiparametric likelihood panel data models, where the conditional likelihood of the outcome Y_i given X_i and η_i depends on a finite-dimensional parameter vector α, and the conditional distribution of η_i given X_i is left unrestricted. The approach of this paper is easily adapted to this case, and delivers a semiparametric likelihood of the form: $\mathbf{Z}\ f\,(y_i|x_i;\ \alpha,\ \delta(\cdot)) = f\,(y_i|x_i,\ \eta_i;\ \alpha) f\,(\eta_i|x_i;\ \delta(\cdot)) d\eta_i$,

where δ(·) is a process of quantile coefficients.

As another example, our framework naturally extends to models with time-varying un-observables, such as:

Yit = QY (Xit, ηit, Uit) ,

it = Qη ηi,t−1, Vit ,

Where U_{it} and V_{it} are i.i.d. and uniformly distributed. It seems worth assessing the usefulness of our approach in these other contexts.

REFERENCES

1. Abrevaya, J. (2006): "Estimating the Effect of Smoking on Birth Outcomes Using a Matched Panel Data Approach," Journal of Applied Econometrics, vol. 21(4), 489–519.

2. Abrevaya, J., and C. M. Dahl (2008): "The Effects of Birth Inputs on Birthweight," Journal of Business & Economic Statistics, 26, 379–397.

3. Ai, C. and Chen, X. (2003): "Efficient estimation of models with conditional moment restrictions containing unknown functions," Econometrica, 71, 1795–1843.

4. Andrews, D. (2011): "Examples of L^2-Complete and Boundedly-Complete Distribu-tions," unpublished manuscript.

5. Arcidiacono, P. and J. B. Jones (2003): "Finite Mixture Distributions, Sequential Like-lihood and the EM Algorithm," Econometrica, 71(3), 933–946.

6. Arcidiacono, P. and R. Miller (2011): "Conditional Choice Probability Estimation of Dy-namic Discrete Choice Models with Unobserved Heterogeneity," Econometrica, 7, 1823– 1868.

7. Arellano, M. and S. Bonhomme (2011): "Nonlinear Panel Data Analysis," Annual Review of Economics, 3, 2011, 395-424.

8. Arellano, M. and S. Bonhomme (2012): "Identifying Distributional Characteristics in Random Coefficients Panel Data Models," Review of Economic Studies, 79, 987–1020.