Stata commands to estimate quantile regression with panel and grouped data

Blaise Melly and Martina Pons

University of Bern

blaise.melly@unibe.ch

November 18, 2022

Summary

- We are interested in the effect of some treatment variables on the distribution of an outcome $\Longrightarrow$ quantile regression.
- We have panel data with $i = 1, ..., N$ and $t = 1, ...T$.
(This also covers grouped data!)
- We suggest quantile versions of the fixed effects, random effects, between, and Hausman and Taylor estimators.
- We use the minimum distance approach:
  - For each individual i regress with quantile regression the outcome on the time-varying regressors.
  - Regress the first stage fitted values on all the regressors with GMM using the appropriate instruments.
- We have implemented these estimators in Stata: `mdqr` for grouped data and `xtmdqr` for panel data.

Traditional panel data and grouped data

- Our results apply to traditional panel data models.
  Example: effect of union status on wages using the PSID.
- But our results also apply to grouped data, where we observe data at the individual level but the treatment varies at the group level. The i units are often called groups and the t units are individuals within these groups.
  - Effect of the food stamp program on the distribution of birth weights. Individual level data but the treatment vary at the county-time level. Almond, Hoynes and Schanzenbach (2011) for mean effects.
  - Effect of import competition on the within-industry wage distribution. Individual level data but the treatment varies at the level of the commuting zone. IV: a measure of import exposure. Autor, Dorn and Hanson (2013) for mean effects.

This presentation

- In the companion paper "Minimum Distance Estimation of Quantile Panel Data Models", we derive the statistical properties of our estimators and of our inference procedures. Link to the newest version
- Today I will describe our model and our estimators and provide an intuitive summary of the theoretical results.
- Then I will focus on the implementation in Stata and on the empirical application.
- The latest version of our codes can be installed by typing

```
net install mdqr,
from("https://raw.githubusercontent.com/bmelly/Stata/main/")
```

Dependencies: qrprocess, moremata, parallel, reghdfe

Model

We assume that the τth conditional quantile function of y_{it} of individual i can be represented by

$$Q(\tau, y_{it}|x_{1it}, x_{2i}, v_i) = x'_{1it}\beta(\tau) + x'_{2i}\gamma(\tau) + \alpha(\tau, v_i) \tag{1}$$

- x_{1it} is a K_1-dimensional vector of time-varying variables.
- x_{2i} is a K_2-dimensional vector of time-constant variables (includes a constant).
- v_i is an unobserved random vector.
- x_{1it} and x_{2i} are potentially correlated with $\alpha(\tau, v_i)$.
- Individual unobserved factors with $\mathbb{E}[\alpha(\tau, v_i)] = 0$.
- z_{it} is a L-dimensional vector of valid instruments, i.e. $\mathbb{E}[z_{it}\alpha(\tau, v_i)] = 0$.

Minimum Distance Quantile Estimator

1. **First stage:** For each individual i and quantile τ, regress y_{it} on the time-varying variables using quantile regression.

$$\hat{\beta}_i(\tau) \equiv \left(\hat{\beta}_{0,i}, \hat{\beta}'_{1,i}\right)' = \underset{(b_0,b_1)\in\mathbb{R}^{K_1+1}}{\arg\min} \frac{1}{T}\sum_{t=1}^{T} \rho_\tau(y_{it} - b_0 - x'_{1it} b_1) \quad (2)$$

where $\rho_\tau(x) = (\tau - 1\{x < 0\})x$ for $x \in \mathbb{R}$ is the check function.

Minimum Distance Quantile Estimator

1. **First stage:** For each individual i and quantile τ, regress y_{it} on the time-varying variables using quantile regression.

$$\hat{\beta}_i(\tau) \equiv \left(\hat{\beta}_{0,i}, \hat{\beta}'_{1,i}\right)' = \underset{(b_0,b_1)\in\mathbb{R}^{K_1+1}}{\arg\min} \frac{1}{T}\sum_{t=1}^{T} \rho_\tau(y_{it} - b_0 - x'_{1it}b_1) \quad (2)$$

where $\rho_\tau(x) = (\tau - 1\{x < 0\})x$ for $x \in \mathbb{R}$ is the check function.

2. **Second Stage**: Regress the fitted values from the first stage on all the variables using GMM with the moment condition $\mathbb{E}[g_i(\delta, \tau)] = 0$ where $g_i(\delta, \tau) = Z_i(\hat{Y}_i(\tau) - X_i\delta(\tau))$.

$$\hat{\delta}(\hat{W}, \tau) = \left(X'Z\hat{W}(\tau)Z'X\right)^{-1} X'Z\hat{W}(\tau)Z'\hat{Y}(\tau) \quad (3)$$

$\hat{W}(\tau)$ is a $L \times L$ symmetric weighting matrix and $\delta = (\beta', \gamma')'$.

Asymptotic results: summary

- We must assume that the number of time periods diverges to infinity otherwise the finite-sample bias of quantile regression dominates.
- The asymptotic theory is complex but the procedures we suggest are actually simple to use.
- The rate of convergence of the elements of $\hat{\delta}(\cdot)$ depends on (i) the presence of unobserved heterogeneity or not and on (ii) the type of variation that identifies the coefficient: time-varying or time-constant instruments.
- The first-order asymptotic distribution may give a poor approximation of the finite-sample behavior of the estimators.
- We show that clustered standard errors applied to the fitted values accounts automatically for the sampling variance arising from both stages of the estimation.
- Using this clustered variance estimates we obtain an efficient GMM estimator and an adaptive inference procedure.

Traditional linear panel data models

- Traditional panel data model:

$$y_{it} = x_{1it}\beta + x_{2i}\gamma + \alpha_i + \varepsilon_{it}$$

  Averaging over t: $\bar{y}_i = T^{-1}\sum_{t=1}^{T} y_{it}$ and $\bar{x}_i = T^{-1}\sum_{t=1}^{T} x_{it}$.

  Time demeaning: $\dot{y}_{it} = y_{it} - \bar{y}_i$ and $\dot{x}_{1it} = x_{1it} - \bar{x}_i$.
- The traditional fixed effects (FE) estimator of β can be computed by regressing y_{it} on x_{1it} for each i separately and construct the fitted values for each observation. Regress fitted values on x_{1it} using $\dot{x}_{1it}$ as an instrument. Minimum distance!
- Similarly, defining $x_{it} = (x'_{1it}, x'_{2i})$ and $z_{it} = (\bar{x}'_{1i}, x'_{2i})'$ we obtain the between estimator (BE).
- With $x_{it} = (x'_{1it}, x'_{2i})$ and $z_{it} = (\dot{x}_{1it}, \bar{x}'_{1i}, x'_{2i})$ we obtain the random effects estimator (RE).

Quantile FE, BE, RE, and pooled estimators

The quantile versions of traditional panel data estimators:

- FE: Regress $\hat{y}_{it}(\tau)$ on x_{1it} with instrument $\dot{x}_{1it}$.
- BE: Regress $\hat{y}_{it}(\tau)$ on x_{it} with instrument $\bar{x}_i$.
- Pooled: Regress $\hat{y}_{it}(\tau)$ on x_{it} with OLS.
- RE: Efficient GMM with instrument $(\dot{x}_{1it}, \bar{x}_i)$ or optimal instruments.

Alternatives:

- Time-demeaning or first differencing: not consistent for quantiles
- Quantile regression with indicator variables for each individual: much slower, does not work for random effects, Hausman-Taylor, grouped IV, etc.

Grouped IV Quantile Regression

Chetverikov et al. (2016) consider a grouped (IV) quantile regression model, which fits into our setup. They are only interested in $\gamma(\tau)$. They suggest a different two-stages estimator:

1. For each i and quantile τ, regress the y_{it} on x_{1it} using quantile regression.
2. Regress the **intercept** from the first stage on the x_{2i} variables with OLS or 2SLS, using one observation per group.

Comparison with our estimator

- Same first-order asymptotic distribution as our estimator (which is the same as if the first stage coefficients were known).
- It is not-invariant to linear reparametrization of x_{1it}.
- It is vulnerable to misspecification (the intercept is the fitted value for $x_{1it} = 0$, which may be outside of the support of x_{1it}).
- It does not impose equality of $\beta_i(\tau)$ and does not exploit the exogeneity of the between variation of x_{1it} $\implies$ less precise.
- If in reality $\beta_i(\tau)$ is not constant across groups i $\implies$ there is heterogeneity: $\gamma(\tau, x_{1it})$. Chetverikov et al. (2016) estimator converges to $\gamma(\tau, x_{1it} = 0)$ while we obtain the best linear approximation.
- In simulations, with their DGP, our estimator performs much better (like 10 times lower MSE).

Stata commands

- `mdqr` *depvar* [*varlist_exogenous* (*varlist_endogenous* = *varlist_instruments*)], group(*groupvar*) [*options*]
  - *varlist_ex* is the list of exogenous variables
  - *varlist_end* is the list of endogenous variables
  - *varlist_iv* is the list of excluded exogenous variables used as instruments for *varlist_end*
  - *groupvar* are the variable(s) defining the groups
  - `quantiles`(*numlist*): quantile(s), default is 0.5
- `xtmdqr` *depvar* [*indepvars*], [fe be re *options*]
  - `xtmdqr` requires that a panel variable has been set with `xtset`.
  - fe: request the fixed effects estimator
  - be: request the between estimator
  - re: request the random efffects estimator
  - `quantiles`(*numlist*): quantile(s), default is 0.5

Bells and whistles

- Automatic detection of the time-varying variables.
- Clustered robust standard errors or the clustered bootstrap can be used. Clustering is not an option: it takes the first stage variance into account.
- For the second stage estimator, the default commands are (1) regress, (2) ivregress 2sls, and (3) ivregress gmm. Another command (e.g. areg, ivreg2 or reghdfe) can optionally be provided.
- Most of the computing time arises in the first stage. It is possible to save the first stage results to avoid computing them again later.
- The first stage is embarrassingly parallelizable. We have included an argument to use parallel processing with the package parallel.

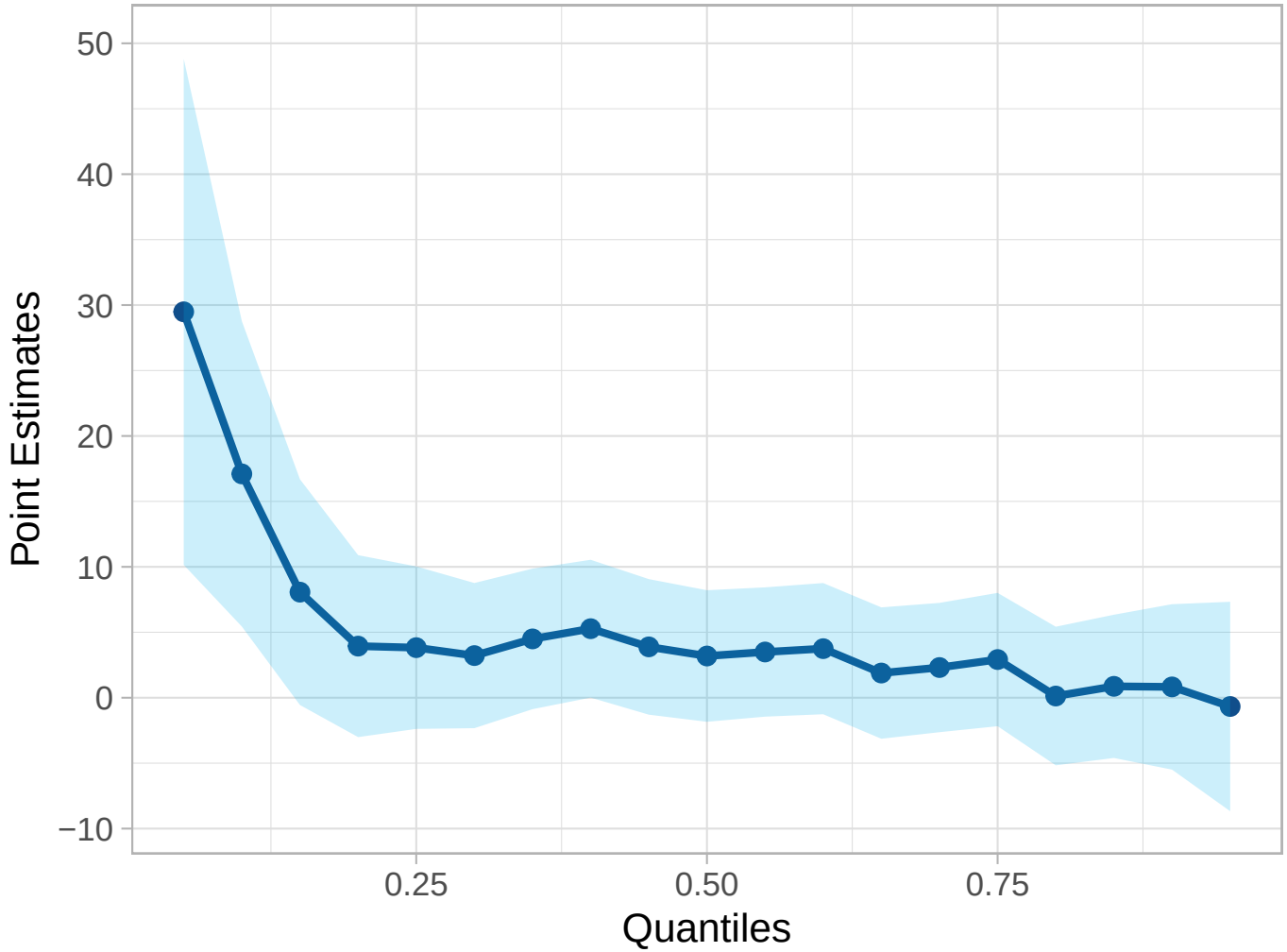
Effect of the food stamp program (FSP) on the distribution of birth weight

- We build on the work Almond et al. (2011) and estimate the distributional effects.
- 1964: Foot Stamp Act enabled counties to start their own (federally founded) FSP
- 1973: amendment to the FSA required all counties to establish a FSP by 1975.
- We use Natality data from 1968 to 1977 augmented with information on FSP rollout and county control variables.

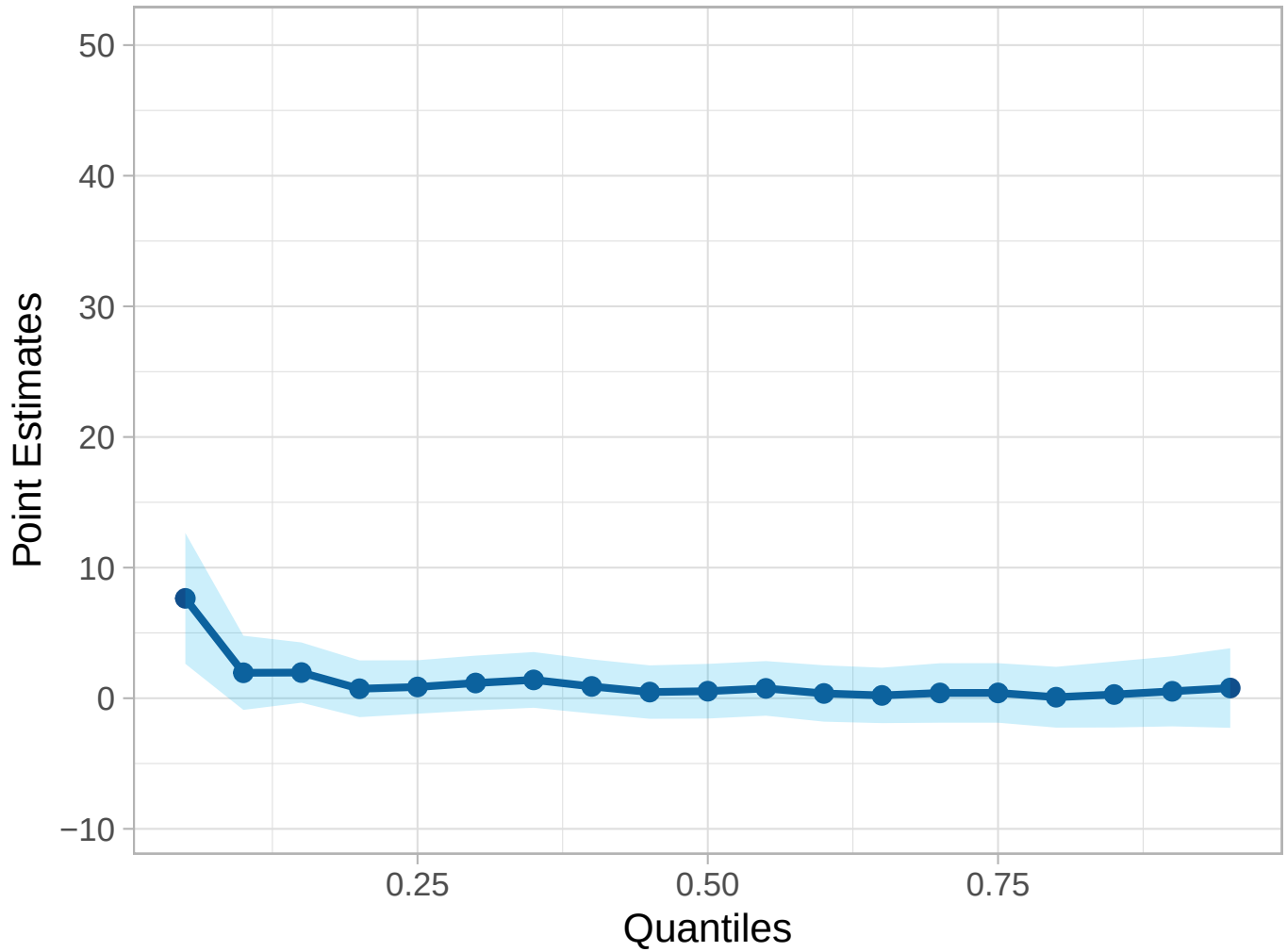
The two-steps procedure

1. For each quantile and county-quarter cell, regress birth weight on individual-level covariates x_{1it}. Save the fitted values.
   - Included variables: gender, mother's age, legitimacy dummies
   - Sample of black mothers: 22.023 groups, 19 different quantiles $\Rightarrow$ 418.437 first-stage quantile regressions (embarrassingly parallelizable)
2. For each quantile regress the fitted value $\hat{bw}_{jct}$ on all variables x_{1it} and x_{2i} including county and trimester and state-year fixed effects.
   - Included variables: gender, mother's age, legitimacy dummies, annual county-level controls, 1960 county-level characteristics interacted with linear time trends.
   - 19 second quantile regressions.

Results - Black Mothers



Results - White Mothers



Summary and limitations

- Summary
  - We suggest a general framework for quantile panel data models.
  - New random effects quantile estimator, new Hausman test, new Hausman-Taylor quantile estimator, new grouped (IV) quantile regression estimator.
  - We have implemented these estimators in Stata as well for traditional panel data as for grouped data.
  - The commands are straightforward to use and computationally fast also in large data sets.
- Limitations
  - Large T asymptotics (but simulations show good performance in finite T),
  - No time fixed effects (but linear, quadratic, etc. trends)
  - Conditional quantile effects (but it is possible to integrate over the individual effects, see Bargain, Etienne, and Melly (2018)).

Related Literature

- (IV) Quantile regression: Koenker and Bassett (1978), Chernozhukov and Hansen (2005). We consider different parameters (conditionally on the individual effects).
- Fixed effects quantile regression: Koenker (2004), Galvao and Wang (2015), Galvao et al. (2020). Special case of our framework.
- Random effects quantile regression: Galvao and Poirier (2019) use pooled quantile regression and estimate unconditional parameters. We suggest a new random effects estimator and a new Hausman test.
- Grouped (IV) quantile regression: Chetverikov et al. (2016). We provide a better estimator, relax the growth rate condition and also consider time-varying variables.
- Minimum distance QR: Chamberlain (1994). We generalize his results by allowing $N \to \infty$, time-varying regressors, and GMM.

References I

Almond, D., H. W. Hoynes, and D. W. Schanzenbach (2011): "Inside the war on poverty: The impact of food stamps on birth outcomes," *Review of Economics and Statistics*, 93, 387–403.

Bargain, O., A. Etienne, and B. Melly (2018): "Public Sector Wage Gaps Over the Long-Run: Evidence from Panel Administrative Data," .

Chamberlain, G. (1994): "Quantile Regression, Censoring, and the Structure of Wages," *Advances in econometrics*, 1, 171–209.

Chernozhukov, V. and C. Hansen (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, 73, 245–261.

Chetverikov, D., B. Larsen, and C. Palmer (2016): "IV Quantile Regression for Group-Level Treatments, With an Application to the Distributional Effects of Trade," *Econometrica*, 84, 809–833.

Galvao, A. and A. Poirier (2019): "Quantile Regression Random Effects," *Annals of Economics and Statistics*, 109–148.

Galvao, A. F., J. Gu, and S. Volgushev (2020): "On the unbiased asymptotic normality of quantile regression with fixed effects," *Journal of Econometrics*, 218, 178–215.

Galvao, A. F. and L. Wang (2015): "Efficient Minimum Distance Estimator for Quantile Regression Fixed Effects Panel Data," *Journal of Multivariate Analysis*, 133, 1–26.

Koenker, R. (2004): "Quantile Regression for Longitudinal Data," *Journal of Multivariate Analysis*, 91, 74–89.

Koenker, R. and G. Bassett (1978): "Regression Quantiles," *Econometrica*, 46, 33.