The Stata Journal (yyyy) **vv**, Number ii, pp. 1–10

Estimation of Quantile Regressions with Multiple Fixed Effects

Fernando Rios-Avila
Levy Economics Institute
Annandale-on-Hudson, NY
friosavi@levy.org

Leonardo Siles
Universidad de Chile
Santiago, Chile
lsiles@fen.uchile.cl

Gustavo Canavire-Bacarreza
The World Bank
Washington, DC
gcanavire@worldbank.org

Abstract.

This is an example of StataJ article made by me

Keywords: st0001, Stata, LaTeX, Quarto, StataJ

1 Introduction

Quantile regression, introduced by Koenker and Bassett (1978), has become an important tool in economic analysis, allowing to examine how the relationship between the dependent and independent variables varies across different points of the conditional distribution of the outcome. While ordinary least squares focuses on analyzing the conditional mean, quantile regression provides a more comprehensive view of how covariates impact the entire conditional distribution of the dependent variable. This can reveal heterogeneous effects that may be otherwise overlooked when analyzing the conditional mean.

A relatively recent development in the literature has focused on extending quantile regression analysis in a panel data setting to account for unobserved, but time fixed heterogeneity. This is particularly important in empirical research, where unobserved heterogeneity can bias estimates of the effects of interest. However, as it is common in the estimation of non-linear models with fixed effects, introducing fixed effects in quantile regression models poses several challenges. On the one hand, the simple inclusion of fixed effects can lead to an incidental parameter problem, which can bias estimates of the quantile coefficients (Neyman and Scott 1948; Lancaster 2000). On the other hand, the computational complexity of estimating quantile regression models with fixed effects can be prohibitive, particularly for large datasets with multiple high-dimensional fixed effects. While many strategies have been proposed for estimating this type of model (see Galvao and Kengo (2017) for a review), none has become standard due to restrictive assumptions regarding the inclusion of fixed effects and the computational complexity.

In spite of the growing interest in estimating quantile regression models with fixed effects in applied research, particularly in the fields of labor economics, health economics, and public policy, among others, there are few commands that allow the estimation of such models. In Stata, there are three main built-in commands available for estimating quantile regression `qreg`, `ivqregress`, and `bayes: qreg`, and none of them allow for

© yyyy StataCorp LLC

the inclusion of fixed effects, other than using the dummy variable approach. From the community-contributed commands, there is `xtqreg`, which implements a quantile regression model with fixed effects based on the method of moments proposed by Machado and Santos Silva (2019), and more recently `xtmdqr` which implements a minimum distance estimation of quantile regression models with fixed effects described in Melly and Pons (2023). In both cases, these command are constrained to a single set of fixed effects.[1]

To address this, in this paper we introduce two Stata commands for estimating quantile regressions with multiple fixed effects: `mmqreg` and `qregfe`. The first command `mmqreg` is an extension of the method of moments quantile regression estimator proposed by Machado and Santos Silva (2019). The second `qregfe`, implements three other approaches: an implementation of a correlated random effects estimator based on Abrevaya and Dahl (2008), Wooldridge (2019) and Wooldridge (2010, Ch12.10.3); the estimator proposed by Canay (2011), and a proposed modification of this approach. In addition, we also present an auxiliary command `qregplot` for the visualization of the quantile regression models.

Both commands offer the advantage of allowing for the estimation of conditional quantile regressions while controlling for multiple fixed effects. First, they leverage over existing Stata commands, as well as other community-contributed commands, to allow users to estimate quantile regression models and their standard errors under different assumptions. Second, they reduce the impact of the incidental parameters problem under different assumptions regarding the data generating process. In terms of standard errors, `mmqreg` allows for the estimation of analytical standard errors (see Machado and Santos Silva (2019) and Rios-Avila et al. (2024)), whereas `qregfe` emphasizes the use of bootstrap standard errors. Finally, both commands are designed to be user-friendly, allowing for the estimation of quantile regression models with fixed effects in a single line of code.

The remainder of the paper is organized as follows. Section 2 reviews the methodological framework for quantile regression. Section 3 describes the methods and formulas used by `mmqreg` and `qregfe` commands. Section 4 introduces the commands, along with a brief description of their syntax and options. Section 5 introduces an auxiliary command for the visualization of quantile regression models. Section 6 provides an empirical applications demonstrating their use. Section 7 concludes.

2 The Basics

While standard regression techniques allow the researcher to determine the effect of covariates X on the expected value of a response variable Y, regression quantiles are capable of characterizing how changes in X affect the entire distribution of Y. In applied

1. There are other community-contributed commands like `xtrifreg`, `rifhdfe`, `qregpd`, `rqr` among others that allow for the estimation of quantile regression models, but do not estimate conditional quantile regressions, but instead focus on unconditional quantile regressions, or quantile treatment effects.

research, this may be useful when one seeks to determine how does a given percentile of the distribution of interest respond to changing values of explanatory variables. For example, consider the effects of smoking on birth weight, as (Abrevaya and Dahl 2008) studied in the context of quantile regression. Instead of fixing an *unconditional threshold* so as to divide the sample in low birth weights and non-low birth weights, the authors followed a *conditional quantile* approach that enabled them to study the effects of birth inputs (e.g. smoking) in different parts of the birth weight distribution ((Abrevaya and Dahl 2008)).

In its more common setting, quantile regressions are assumed to follow a linear form:

$$Q_\tau(Y|X) = X\beta(\tau) \tag{1}$$

Where $Q_\tau(Y|X)$ is the τ-th quantile of Y conditional on

Such that, while the parameters of interest change across quantiles, the effects of covariates X on the τ-th quantile of Y are assumed to be linear.

In addition to heterogeneous effects of explanatory variables on the outcome of interest, the researcher might be interested as well in properly addressing the problem of unobserved heterogeneity. Returning to the example of birth weights, it may be the case that unobservable characteristics of the mother are correlated with both the birth outcome and birth inputs, such as her smoking status during pregnancy or whether she attended prenatal care visits ((Abrevaya and Dahl 2008)). In this case, a panel data approach may be useful for obtaining causal estimates of the effects of explanatory variables on quantiles of Y.

At the same time, panel data models with unobserved time-invariant effects depend upon demeaning techniques to eliminate such effects and yield consistent estimators of the parameters of interest. However, differentiating out individual effects is not a feasible approach within the context of regression quantiles, the reason being that quantiles are *not* linear operators. Many authors, for example (Abrevaya and Dahl 2008), (Canay 2011) or (Machado and Santos Silva 2019) have stated that this key limitation has prevented further progress in the estimation of regression quantiles with fixed effects.

In this paper, we review four estimators for regression quantiles in the presence of fixed effects. All of them are relatively simple to implement with our provided commands for Stata (see sections 3, 4 and 5 below). However, this easiness comes at the cost of making important assumptions regarding, among others, the data generating process (DGP) and the effect of unobserved heterogeneity α_i on the distribution of Y. The following sections describe each of the estimators, their key assumptions and provide examples for their implementation in Stata.

3 Correlated Random Effects

This estimator, described by (Wooldridge 2010), rests upon the Mundlak repsesentation of a Correlated Random Effects (CRE) model for the estimation of regression quantiles

with panel data. The underlying DGP for the outcome of interest Y_{it} is:

$$Y_{it} = X'_{it}\beta + \alpha_i + u_{it}$$

where α_i is the time-invariant individual effect and can be described in terms of $X_i \equiv (X_{i1}, ..., X_{iT})$ as follows:

$$\alpha_i = \psi + \bar{X_i}'\lambda + v_i$$

where $\bar{X_i}$ denotes the time average of characteristics across individuals. Then, the conditional τ-th quantile of Y_{it} is:

$$Q_\tau(Y_{it}|X_i) = \psi + X'_{it}\beta + \bar{X_i}'\lambda + Q_\tau(a_{it}|X_i)$$

where $a_{it} \equiv v_i + u_{it}$ is the composite error. The CRE estimator using the Mundlak representation imposes the following key assumption:

CRE1. a_{it} is independent of X_i.

Stringent as may be, this assumption allows to consistently estimate β and λ from a pooled quantile regression of Y_{it} on X_{it} and $\bar{X_i}$. As (Wooldridge 2010) remarks, CRE1 amounts to assuming that quantile functions are parallel or, in other words, that by imposing a relationship between X_i and α_i of the form displayed in equation [eq:mundlak_rep] we can only estimate the location shifts of quantile regressions. As it will be shown below, this is a feature the CRE estimator shares with Canay's method (see section 4) for estimating regression quantiles in a panel data context. Both estimators are constrained in the sense that only location effects on quantile functions are accounted for, whereas the MMQREG method (see section 5) has the ability to estimate quantile coefficients that take into account both location and scale effects.

(Abrevaya and Dahl 2008) develop an alternative CRE estimator for regression quantiles with panel data drawing on the model of (Chamberlain 1982). However, its implementation is severely restricted because Abrevaya and Dahl's estimator regresses Y_{it} on X_{it} and X_i for a given τ. Even if we have a balanced panel available for our application, we will have to deal with the large quantity of regressors included in the equation $[K(1+T)]$ and its corresponding loss of degrees of freedom for statistical inference.

Introduction of multiple fixed effects in this setting is straightforward. First, modify equation [eq:dgp_cre] so as to include two or more sets of fixed effects $(\alpha^1, \alpha^2, ..., \alpha^k)$ and proceed by computing the averages of explanatory variables X with respect to the variable identifying the fixed effect. Finally, run a regression of Y on the set of original control variables and the averages just calculated. In terms of the quantile function, equation [eq:qr_cre] requires to be modified accordingly:

$$Q_\tau(Y_{it}|X_i) = \Psi + X'_{it}\beta + \bar{X}_i^{1\prime}\lambda^1 + \bar{X}_i^{2\prime}\lambda^2 + Q_\tau(a_{it}|X_i)$$

For the case of twoway fixed effects and where $\Psi \equiv \psi^1 + \psi^2$.

Implementation of the CRE estimator for quantile regression in Stata can be done with the command `cre`, which is actually a prefix for the estimation of CRE models. Option `[abs(varlist)]` is used to provide the command with variables identifying groups of observations that will be subsequently used to estimate the fixed effects. By default, singleton groups are dropped from the estimation. Now, we present an application of the CRE estimator for regression quantiles with multiple fixed effects using the aforementioned command. Note that the same dataset will be used for all estimation methods surveyed in this paper, so it will be worth to outline the main features of the data before proceeding to display, comment and compare the estimation results.

We use the panel from (Persson et al. 2007) who study the effect of alternative types of government on its overall spending in the last year of the legislature. The authors consider both coalition and single-party as government types, leaving out from consideration minority governments. Here, we will only focus on the relationship a coalition government has with its overall spending and expand (Persson et al. 2007) results by estimating the .25, .5 and .75 quantiles of the government spending distribution.

Variables used for estimation include central government spending as percentage of GDP in the last year of legislature (*last_exp*), an indicator variable for coalition governments (*coalition*), population size in log scale (*lpop*), the percentage of population over 65 years of age (*prop65*), the log-deviation of output from the country trend (*ygap*) and the length (in years) of the legislature (*length*). Both country and period fixed effects are included in our estimations, in the same fashion as (Persson et al. 2007).

Table 1: Government spending and coalition government - CRE Estimator

Column 1	Column 2	Column 3
Data 1	Data 2	Data 3
Data 4	Data 5	Data 6

Estimation results from (Persson et al. 2007) dataset using the CRE estimator for quantile regression are displayed in table 1 and show heterogeneous effects of a permanent switch from a single-party to a coalition government on government spending. The largest effect is obtained at the median of the distribution, and the estimated coefficient is close to the one (Persson et al. 2007) computed using the within estimator (2.36). At the lower and upper quartile of the distribution we still find positive coefficients, which support the theoretical predictions of the model developed by the same authors. However, none of the coefficients for coalition across quantiles are statistically different from zero. In the context of CRE estimation, this fact may be partly due to the loss of degrees of freeedom implied by the addition of 12 extra regressors to the equation. In

general, CRE will add $W \times K$ explanatory variables, with W being the number of fixed effects we are accounting for.

4 Canay (2011) Estimator

Canay's estimator key assumption is that fixed effects are merely treated as location-shifters. That is, unobserved individual effects only change the distribution of the response variable in a parallel fashion. While this may be a rather strong assumption, it helps to consistently estimate the parameters of interest using a simple transformation of the data as detailed below (Canay 2011).

The underlying DGP of Y_{it} is:

$$Y_{it} = X'_{it}\beta(u_{it}) + \alpha_i$$

To consistently estimate the β_τ from the data, the following assumptions are in place:

CANAY1. Conditional on $X_{it} = x$ the random variables X_{it} and α are independent.

CANAY2. U_{it} is independent of X_i and α_i, with $U_{it} \sim U[0, 1]$

CANAY3. $\beta_\mu \equiv E[\beta(U_{it})]$. That is, β_μ exists.

Start defining the following equation, which is simply a restatement of equation [eq:dgp_canay], as follows:

$$Y_{it} = X'_{it}\beta(\tau) + \alpha_i + e_{it}(\tau)$$

Where $e_{it}(\tau) \equiv X'_{it}[\beta(U_{it} - \beta(\tau)]$. Then, due to assumption CANAY2, in the above equation only $\beta(\tau)$ and $e_{it}(\tau)$ depend on τ. Next, write a conditional mean equation for Y_{it} as the following:

$$Y_{it} = X'_{it}\beta_\mu + \alpha_i + u_{it}$$

Where $u_{it} \equiv X'_{it}[\beta(U_{it}) - \beta_\mu]$. Assumption CANAY3 allows us to write this kind of mean equation. Then, the two-step estimator due to (Canay 2011) is defined next:

Step 1. Let $\hat{\beta}_\mu$ be a $\sqrt{nT}$-consistent estimator of β_μ. Define $\hat{\alpha}_i \equiv \mathbb{E}_T[Y_{it} - X'_{it}\hat{\beta}_\mu]$.

Step 2. Let $\hat{Y}_{it} = Y_{it} - \hat{\alpha}_i$ and define the two-step estimator as:

$$\hat{\beta}(\tau) \equiv \min_{\beta \in \beta} \mathbb{E}_{nT}[\rho(\hat{Y}_{it} - X'_{it}\beta)]$$

Where $\rho(\cdot)$ is the check function (Koenker and Bassett 1978), $\mathbb{E}_T(\cdot) \equiv T^{-1}\sum_{t=1}^{T}(\cdot)$ and $\mathbb{E}_{nT}(\cdot) \equiv (nT)^{-1}\sum_{t=1}^{T}\sum_{i=1}^{n}(\cdot)$. (Canay 2011) defines this estimator as "consistent and asymptotically normal". This stated, Canay's estimator stands out for its simplicity: Step 1 may be conveniently restated as obtaining the known within estimator of β and then subtracting the time-average residuals from the dependent variable. Next, using this simple transformation of the data, Step 2 tells us to obtain regression quantiles of $\hat{Y}_{it}$ that are guaranteed to be consistent, as long as key assumptions hold.

It is important to remark that consistency of Canay's estimator is obtained when $T \to \infty$ (Canay 2011). Thus, the estimator at hand may perform poorly in contexts of panels with few time periods. In addition to this, the key assumption of the approach relies on the fact that the α_i's are pure location shifts: in other words, that the unobservable time-invariant characteristics grouped in α_i have coefficients that are constant across τ (Canay 2011). As (Machado and Santos Silva 2019) point out using simulated data, the estimator's performance deteriorates as the DGP features fixed effects that alter the entire distribution of interest, not only its location.

Regarding statistical inference, (Canay 2011) provides a simple algorithm to obtain bootstrapped standard errors. The key in this procedure is to compute the two step estimator for each bootstrap sample, avoiding the mistake of just sampling from the transformed data $\hat{Y}_{it}$. As (Canay 2011) remarks, the algorithm should be such that it computes the first step estimators $\hat{\beta}_\mu$ and $\hat{\alpha}_i$ for each repetition.

Finally, we propose a Modified Canay estimator in which, instead of subtracting out the estimated fixed effects $\hat{\alpha}_i$ from the dependent variable Y_{it}, they are included as regressors in equation [eq:qreg_canay]. In stark contrast to the CRE estimator (see section 3), the Modified Canay only adds W variables to the estimation, which makes it preferable in contexts of a small sample size when our goal is to estimate only location shift effects. However, we must be aware of the fact that the requirement of $T \to \infty$ for consistency may not be met, especially for short panels.

Describe the implementation in Stata here.

Table 2: as

Column 1	Column 2	Column 3
Data 1	Data 2	Data 3
Data 4	Data 5	Data 6

Tables 2 and 3 display the results of Canay and Modified Canay estimations using the (Persson et al. 2007) dataset of government expenditures. Once again, the median coefficients are close to the mean coefficient from the original paper, with the Canay estimator delivering the closest coefficient to 2.36. Even further, statistical significance is achieved for the median coefficient in table 2, strengthening our reasoning on degrees of freedom above. Otherwise, we still obtain positive coefficients of the coalition variable at the 25th and 75th percentile of government spending.

It is also worth stating the increasing monotonic pattern of quantile coefficients of coalition in table 3, conveying us that the effect of a permanent switch from single-party to coalition governments increases as we move toward greater government spending. Finally, note that even after explicitly accounting for country and period of legislature fixed effects in the Modified Canay estimator, the median coefficient remains stable, although it is no longer significant.

Column 1	Column 2	Column 3
Data 1	Data 2	Data 3
Data 4	Data 5	Data 6

5 Method of Moments Quantile Regression

This approach of estimating regression quantiles with multiple fixed effects distinguishes itself from the other estimators reviewed above in the sense that (Machado and Santos Silva 2019) introduce location-scale effects of fixed effects upon the distribution of interest. Compared to Canay's estimator, the Method of Moments Quantile Regression (MMQREG) not only allows the α_i's to affect Y_{it} through location shifts, but rather MMQREG is able to identify the scale shifts that alter different points of the distribution belonging to Y.

We begin by defining the DGP of the location scale model:

$$Y_{it} = \alpha_i + X'_{it}\beta + (\delta_i + X'_{it}\gamma)u_{it} \tag{2}$$

Where parameters α_i and δ_i capture the individual fixed effects. Note that, compared to equation [eq:dgp_canay], the fixed effects not only enter the model in an additive fashion, instead they also have a multiplicative effect upon the error term. In addition, the U_{it} are i.i.d. across i and t, statistically independent of X_{it} and satisfy $E(U) = 0$ and $E(|U|) = 1$ both of which normalize the random variable.

Our location scale model in equation [eq:dgp_mmqreg] implies that:

$$Q_\tau(Y_{it}|X_i) = [\alpha_i + \delta_i q(\tau)] + X'_{it}\beta + X'_{it}\gamma q(\tau)$$

Where the scalar coefficient $\alpha_i(\tau) \equiv \alpha_i + \delta_i q(\tau)$ is the quantile-τ fixed effect for individual i, which represent how time-invariant variables have *different impacts on different regions* of the conditional distribution of Y_{it}. However, our real interest is in the regression quantile coefficients:

$$\beta_\tau = \beta + q(\tau)\gamma$$

Which are simply a linear combination of the location coefficients (β) and the scale coefficients (γ), where the second vector of coefficients is weighted by the value of the τ-th quantile of the variable of interest Y_{it}. (Machado and Santos Silva 2019) develop the following algorithm for implementing the MMQREG estimator:

1. Obtain $\hat{\beta}_k$ by regressing time-demeaned Y_{it} on time-demeaned controls X_{it}, i.e. obtain $\hat{\beta}$ by the within estimator.
2. Estimate the $\hat{\alpha}_i$'s and calculate the residuals $\hat{R}_{it} = Y_{it} - \hat{\alpha}_i - X'_{it}\hat{beta}$.
3. Obtain $\hat{\gamma}_k$ by the within estimator using $|\hat{R}_{it}|$ as the dependent variable.
4. Estimate $\hat{\delta}_i$ by taking the time average of $|\hat{R}_{it}| - X'_{it}\hat{\gamma}$.
5. Estimate $q(\tau)$ by $\hat{q}$, which corresponds to the regression quantile of standardized residuals $[\hat{R}_{it}/(\hat{\gamma}_i + X'_{it}\hat{\gamma})]$ upon an intercept term.

Note that Steps 1 and 2 from the MMQREG algorithm are the same as those performed in Canay's estimator. However, the location-scale model used in (Machado and Santos Silva 2019) adds three additional steps that are required to estimate $\hat{\gamma}$ and $\hat{q}$, so that regression quantile coefficients β_τ are allowed to affect not only the location of the distribution, but also its shape.

Add notes on the assumptions of MMQREG here *if* necessary.

Statistical inference can be performed using the asymptotic distribution of the estimator derived in (Machado and Santos Silva 2019). Expanding on this literature, (Rios-Avila et al. 2024) propose methods for computing alternative standard errors using the empirical influence functions of the estimators. Robust and clustered standard errors can be estimated following this approach, and are readily available as options in the `mmqreg` command in Stata. Another extension of the MMQREG estimator due to (Rios-Avila et al. 2024) is to allow for the inclusion of multiple fixed effects using an application of the Frisch-Waugh-Lovell (FWL) theorem to partial out the effect of variables capturing unobserved heterogeneity from both dependent and explanatory variables. Likewise, the command `mmqreg` allows for multiple fixed effects, as will be shown next.

Table 4: Government spending and coalition government - MMQREG Estimator

Column 1	Column 2	Column 3
Data 1	Data 2	Data 3
Data 4	Data 5	Data 6

As we allow for both location and scale shifts in our estimation method, the monotonic pattern observed in the first row of table 3 is reversed. Now, the farther we move from the median of the distribution in the direction of greater government spending, the

lower is the effect of a permanent change from single-party to coalition upon spending. The median coefficient, once again, remains close to the mean coefficient (2.36) under MMQREG estimation. Results from table 4 support the conclusions from (Persson et al. 2007) regarding types of government and its spending during the last year of legislature, expanding the results to the response of the distribution of spending upon changes in explanatory variables.

We must also note that statistical inference —with robust standard errors displayed in table 4— for the (Persson et al. 2007) dataset show that many coefficients that with the previous estimators were not statistically different from zero, are now significant even at the 99% confidence level. This is due to the ability to compute robust standard errors for the MMQREG estimator, contrasting with the CRE and Canay estimators, for which we estimated the covariance matrix using the bootstrap. Although not displayed here, our findings are robust to clustering standard errors for country so that the coefficients of 25th and 50th quantiles corresponding to the indicator variable for coalition government remain statistically significant.

6 References

Abrevaya, J., and C. M. Dahl. 2008. The Effects of Birth Inputs on Birthweight. *Journal of Business & Economic Statistics* 26(4): 379–397.

Canay, I. A. 2011. A simple approach to quantile regression for panel data. *The Econometrics Journal* 14(3): 368–386.

Chamberlain, G. 1982. Multivariate regression models for panel data. *Journal of Econometrics* 18(1): 5–46.

Galvao, A. F., and K. Kengo. 2017. Quantile regression methods for longitudinal data. In *Handbook of quantile regression*, 363–380. Chapman and Hall/CRC.

Koenker, R., and G. Bassett. 1978. Regression Quantiles. *Econometrica* 46(1): 33–50.

Lancaster, T. 2000. The incidental parameter problem since 1948. *Journal of Econometrics* 95(2): 391–413.

Machado, J. A., and J. Santos Silva. 2019. Quantiles via moments. *Journal of Econometrics* 213(1): 145–173.

Melly, B., and M. Pons. 2023. Minimum Distance Estimation of Quantile Panel Data Models. Unpublished working paper.

Neyman, J., and E. L. Scott. 1948. Consistent Estimates Based on Partially Consistent Observations. *Econometrica* 16(1): 1–32.

Persson, T., G. Roland, G. Tabellini, et al.. 2007. Electoral rules and government spending in parliamentary democracies. *Quarterly Journal of Political Science* 2(2): 155–188.

Rios-Avila, F., L. Siles, and G. Canavire-Bacarreza. 2024. Estimating Quantile Regressions with Multiple Fixed Effects through Method of Moments. *Working Paper* .

Wooldridge, J. M. 2010. *Econometric analysis of cross section and panel data*. MIT press.

———. 2019. Correlated Random Effects Models with Unbalanced Panels. *Journal of Econometrics* 211(1): 137–150.

About the authors

Fernando Rios-Avila is a Research Scholar at the Levy Economics Institute of Bard College.

Gustavo Canavire-Bacarreza is a Senior Economist at the World Bank.