The Stata Journal (yyyy) **vv**, Number ii, pp. 1–15

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 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 depending on the assumptions of the underlying 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

Quantile regressions allow researchers to identify the heterogenous effect covariates could have over the entire conditional distribution of the dependent variable. Let y_i be the dependent variable, x_i the vector of covariates excluding a constant, and $0 < \tau < 1$ is a

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.

parameter such that $q_\tau(y_i|X)$ identifies the τth quantile of the conditional distribution of $y_i|X$. Under the assumption that conditional quantiles are linear functions of the parameters, the quantile regression model can be written as:

$$q_\tau(y_i|X) = \beta_0(\tau) + x_i'\beta(\tau) \tag{1}$$

Where $\beta(\tau)$ is the vector of coefficients that may vary across τ and needs to be estimated, and x_i is a vector of exogenous covariates that may include nonlinear functions of underlying variables. This expression indicates that, conditional on x, the τ-th quantile of y can be approximated by a linear function of X.

Under the assumption that the conditional quantile function is linear and correctly specified, a useful way to think about the data generating process is to consider the following model:

$$y_i = \beta_0(U_i) + x_i'\beta(U_i) \tag{2}$$

where U_i is a random variable that follows a uniform distribution. It can be seen as the rank an individual belongs to among all individuals with the same characteristics. In addition, β_0 and $\beta(U_i)$ are smooth functions that depend on U_i.[2]

As explained in Wooldridge (2010), the coefficient of quantile regression models can be identified by minimizing the following loss function, with respect to $\beta(\tau)$:

$$\hat{\beta}_0(\tau), \hat{\beta}(\tau) = \min_{\beta(\tau)} \sum_{i=1}^{n} \rho_\tau\big(y_i - \beta_0(\tau) - x_i'\beta(\tau)\big) \tag{3}$$

Where $\rho_\tau(u) = u\big(\tau - I(u < 0)\big)$ is the check function, and $I(\cdot)$ is the indicator function. In essence, quantile regressions are estimating the parameters locally around the τ-th quantile, although other approaches are possible [3].

Most commands for estimating quantile regression models focus on estimating the above loss function, using linear programming techniques, while others like Kaplan and Sun (2017) (`sivqr`) and Chernozhukov et al. (2022) (`qrprocess`) use other optimization techniques.

When no unobserved heterogeneity is present, quantile regression model can be easily implemented in a panel setting (see Wooldridge (2010)), using a pooled version of the model. However, when unobserved heterogeneity is present explicitly, the estimation

2. This way of thinking about quantile regression coefficients is similar to the use of Smooth varying coefficient models, except that the running variable is not observed.
3. Kaplan and Sun (2017) for example uses a nonparametric approach that produces smooth set of beta coefficients. And Bottai and Orsini (2019), proposes methods for estimating parametric quantile regression models, impossing parametric restrictions on the quantile coefficients across the distribution.

of quantile regressions is more challenging. Consider the case of panel data and the following data generating process:

$$y_{it} = \beta_0(U_{it}) + x'_{it}\beta(U_{it}) + \alpha_i(U_{it}) \tag{4}$$

Where U_{it} is a random variable that follows a uniform distribution, and $\alpha_i(U_{it})$ is the unobserved effect that varies across individuals. In this case, the conditional quantile regression model can be written as:

$$q_\tau(y_{it}|x_{it}, \alpha_i(\tau)) = \beta_0(\tau) + x'_{it}\beta(\tau) + \alpha_i(\tau) \tag{5}$$

This specification is explicitly considering that the unobserved effect is identified for each i_{th} observation, and that it varies across quantiles ($\alpha_i(\tau)$). A common approach used, yet incorrect due to the incidental parameter problem, is to estimate this model by adding dummy variables for each individual in the quantile regression model (as in Budig and England (2001)), or by demeaning the explanatory variables (as in Budig and Hodges (2010)). In contrast with standard linear models, there is no transformation of the data that can eliminate the individual fixed effects for non-linear models like quantile regressions.

In this framework, the problem of the incidental parameter problem occurs because the unobserved factors cannot be differenced out of. In other words, unless specific assumptions are made, the estimation requires the explicit estimation of the unobserved fixed effect. Unfortunately, because the number of available observations per individual fixed effect is limited, they cannot be estimated with precision. In turn, the cumulative errors in the estimation of the fixed effects will also affect the conditional distribution of the outcome, which quantile regressions leverage on, leading to inconsistent estimates of all parameters. [4]

In the next section, we present a few solutions and implementations for the estimation of quantile regression models with multiple fixed effects.

2.1 Correlated Random Effects: CRE

The first approach we discuss is the use of Correlated Random Effects (CRE) models for the estimation of quantile regression models. The CRE model is an alternative methodology for the estimation of fixed effects models that was proposed by Mundlak (1978) and generalized by Chamberlain (1982). In contrast with standard fixed effects, the approach allows users to control for time fixed covariates in addition to time-varying covariates. And, in contrast with the random effects model, it does not make the assumption that the unobserved effect is uncorrelated with the observed covariates. Interestingly, in the context of linear models, the CRE model is equivalent to the fixed effects model (Wooldridge 2010).

4. This is similar to the measuring error problem of dependent variables in quantile regression models discussed in Hausman et al. (2021).

Consider the following model:

$$y_{it} = \beta_0 + x_{it}\beta + \alpha_i + u_{it} \tag{6}$$

It is well known that if α_i is correlated with x_{it}, the Random Effects (RE) estimator will be inconsistent, due to the ommited variable bias. The solution proposed by Mundlak (1978) and Chamberlain (1982) was to explicitly account for that correlation in the model, by assuming the unobserved effect α_i is a linear projection of the observed time-varying variables plus an uncorrelated disturbance. Specifically:

$$\begin{aligned} Mundlack: \quad & \alpha_i = \gamma_0 + \bar{x}_i\gamma + v_i \\ Chamberlain: \quad & \alpha_i = \gamma_0 + x_{i1}\gamma_1 + x_{i2}\gamma_2 + \cdots + x_{iT}\gamma_T + v_i \end{aligned} \tag{7}$$

The main difference between both approaches was that Chamberlain (1982) proposes a more flexible specification allowing all realizations of the time-varying variables to explain the unobserved effect. In contrast, Mundlak's approach only considers the average of the time-varying variables, which is a more restrictive specification. Using either model specification, if we substitute Equation 7 into Equation 6, the final model can be written as:

$$y_{it} = \beta_0 + x_{it}\beta + \gamma_0 + f(x_{it})\Gamma + v_i + u_{it} \tag{8}$$

where $f(x_{it})$ can be the full set of time-varying variables or just the average of them. Notice that in this specification, β_0 and γ_0 cannot be independently identified, and that the new model now has a compound error $v_i + u_{it} = \mu_i t$, which is uncorrelated with x_{it}. To account for the within indivual correlation driven by v_i, the CRE model should be estimated using either random effects, or clustering standard errors at the individual level (see Wooldridge (2010) for a discussion). Interestingly, either method provides the same results if the panel data is balanced, and all covariates are strictly exogenous. However, this identity breaks down in other cases (see Abrevaya (2013)).

The strategy proposed by Abrevaya and Dahl (2008) was to extend the CRE model (Chamberlain (1982) style) for the estimation of quantile regression models. This, however, has some limitations. First, when the number of periods is large, the number of additional regressors grows quickly, which can lead to other problems during estimation. Second, while the application of Chamberlain (1982) projection approach for unbalance data is possible (see Abrevaya (2013)), it is not straightforward to implement in practice, specially for the framework of quantile regressions. Instead, we follow Wooldridge (2010) and Wooldridge (2019), and use the Mundlak representation of the CRE model for the estimation of quantile regression models. Wooldridge (2019) has shown that this can be easily applied for cases with unbalanced panels, and the estimation of non-linear models.

Specifically, Wooldridge (2010) suggests that we could use a local projection of the quantile specific unobserved effect. If we concentrate on $\alpha(U_{it})$, where U_{it} is a random variable that follows a uniform distribution, we could write the unobserved effect as:

$$\alpha_i(U_{it}) = \gamma_0(U_{it}) + \bar{x}_i'\gamma(U_{it}) + v_i^{U_{it}} \tag{9}$$

Then, we can use Equation 9 to write the new Data Generating Process (DGP):

$$y_{it} = \beta_0(U_{it}) + x_{it}\beta(U_{it}) + \gamma_0(U_{it}) + \bar{x}_i'\gamma(U_{it}) + v_i^{U_{it}} \tag{10}$$

Two important points to note here. First, as before, $\beta_0(\cdot)$ and $\gamma_0(\cdot)$ cannot be independently identified, which makes the interpretation of the constant term difficult. Second, $v_i^{U_{it}}$ is not a smooth function of U_{it}, but rather an unrelated disturbance that is left after modeling the unobserved effect, and remains unobserved. If we assume that $v_i^{U_{it}}$ is small enough compared to the overall variation driven by U_{it}, we could identify the quantile regression coefficients as follows:

$$q_\tau(y_{it}|x_{it}, \bar{x}_i) = b_0(\tau) + x_{it}\beta(\tau) + \bar{x}_i'\gamma(\tau) \tag{11}$$

Which can be estimated using any standard quantile regression method. However, if $v_i^{U_{it}}$ is large, standard estimators will leverage the distribution of the compound error $v_i^{U_{it}}$ and U_{it}, which may lead to inconsistent estimates of the quantile coefficients. [5]

Nevertheless, assuming that the residual $v_i^{U_{it}}$ is small, the CRE-quantile regression approach has few other benefits that may be of interest. First, as discussed in Wooldridge (2019), it can be easily use in the presence of unbalanced panels. Second, it may also provide an approach to control for multiple fixed effects. [6]. For example, let us expand on Equation 4, and consider the case of a two-way fixed effects model:

$$y_{it} = \beta_0(U_{it}) + x_{it}'\beta(U_{it}) + \alpha_i(U_{it}) + \alpha_t(U_{it})$$

To apply the two-way CRE model, we could use the following representation of the unobserved effects:

$$\alpha_i(U_{it}) + \alpha_t(U_{it}) = \gamma_0(U_{it}) + \lambda_i^x\gamma_i(U_{it}) + \lambda_t^x\gamma_t(U_{it}) + v_{it}^{U_{it}} \tag{12}$$

where λ_i^x and λ_t^x are obtained by estimating the following model for each explanatory variable x_{it}:

$$x_{it} - \bar{x} = \lambda_i^x + \lambda_t^x + \epsilon_{it} \tag{13}$$

5. This is the main critique raised by Canay (2011) to the estimator proposed by Abrevaya and Dahl (2008). In fact the more dominant becomes $v_i^{U_{it}}$, the more the estimates will resemble the OLS estimates.
6. Baltagi (2023) and Wooldridge (2021) discusses this for the two-way Mundlack estimator

We use the centered transformation of the explanatory variable, that is $x_{it} - \bar{x}$, so that all $\lambda's$ have an expected value of zero. In contrast with Baltagi (2023), we suggest that rather than modeling each individual component separately, it is easier to think of the problem of modeling the combination of the two (or many) unobserved components as a function of λ_i^x and λ_t^x, which are the equivalent to $\bar{x}_i$ in the Mundlack one-way fixed effect model. Additionally, different from Wooldridge (2021) and Baltagi (2023), we emphasize that the estimation of λ_i^x and λ_t^x should be done simultaneously (Equation 13), rather than estimating the conditional means separately. This is more general and applicable to any number of fixed effects. This can be done using an iterative process similar to Rios-Avila (2015) or Correia (2016).[7]

With this considerations, the conditional quantile regression can be written as:

$$q_\tau(y_{it}|x_{it}, \lambda_i^x, \lambda_t^x) = x_{it}'\beta(\tau) + \lambda_i^{x\prime}\gamma_i(\tau) + \lambda_t^{x\prime}\gamma_t(\tau)$$

Which could be extended to any number of fixed effects. As before, this approach is valid if the residual $v_{it}^{U_{it}}$ from the time and individual fixed effects (or all fixed effects considered) are small enough compared to the variation driven by the latent rank variable U_{it}.

In terms of the standard errors, for the linear CRE model, it is suggested to use the random effects estimator, or clustering standard errors at the individual level. For the quantile regression model, clustering standard errors at the individual level is also suggested by Wooldridge (2010), and some rutines already implement this feature. When multiple fixed effects are considered, it is suggested to use the bootstrap methods for the estimation of the standard errors.

2.2 Canay (2011) Estimator

The second approach under consideration is the estimator proposed by Canay (2011). As mentioned before, this paper argues that the estimator proposed by Abrevaya and Dahl (2008), and thus the implementation described above, may not provide consistent estimates of the quantile regression coefficients, as long as there is a disturbance $v_i^{U_{it}}$ left after modeling Equation 9. Instead, under the assumption that the unobserved effect is a pure location shift, they propose an alternative estimator can be used to consistently estimate the quantile regression coefficients.

Before presenting the estimator, it is convinient to review a second approach that has been used to understand quantile regression models: The location-scale model. Under this specification, consider the following data generating process:

7. If the panel is perfectly balanced, estimating λ_i^x and λ_t^x separately will provide the same results as estimating them simultaneously. Internally, we use Correia (2016) `reghdfe` to obtain the predicted fixed effects

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \gamma_0(U_i) + \gamma_1(U_i)x_i \text{ or} \\ y_i &= \beta_0 + \beta_1 x_i + \mu_i \end{aligned} \tag{14}$$

In this specification, we assume that β_0 and β_1 are the location parameters that capture how the whole distribution of y_i is affected by x_i. In contrast, $\gamma_0(U_i)$ and $\gamma_1(U_i)$ are the scale parameters that capture the heterogenous effect of x_i on y_i that deviates from the location effect. When using a simple linear regression model, estimated via OLS, we could assume that the compound component $\gamma_0(U_i)+\gamma_1(U_i)x_i$ is fully captured by the error term μ_i.

This provides three insights. First, that OLS could be used to identify the location effect of x_i on y_i, which we know as the average or conditional mean effect. Second, if $\gamma_1(U_i)$ is different from zero, the model is heteroskedastic, and the quantile regression could be used to identify this type of unobserved heterogenity. Lastly, if a covariate has no scale effect, all quantile coefficients will be the same as the OLS coefficient, except for the constant.

Although of little use, this location-scale model can be easily estimated using a two-step approach. First, estimate the location effect of x_i on y_i using OLS:

$$y_i = \beta_0 + x_i'\beta + \mu_i \tag{15}$$

Then, using the predicted residuals $\hat{\mu}_i$, estimate the quantile regression model:

$$q_\tau(\hat{\mu}_i|x_i) = \gamma_0(\tau) + x_i'\gamma(\tau) \tag{16}$$

It is a simple excersise to show that adding $\beta+\gamma(\tau)$ provides the same point estimates as estimating the full quantile regression model. However, it does provide a simple connection between OLS and quantile regression models.

Now, lets reconsider the data generating process in Equation 4. Canay (2011) imposes the assumption that α_i is a pure location shift that should be constant across quantiles. More explicitly, the data generating process in a panel data setting, can be written as follows:

$$y_{it} = \beta_0(U_{it}) + x_{it}\beta(U_{it}) + \alpha_i \tag{17}$$

which imply that the conditional quantile regression model can be written as:

$$q_\tau(y_{it}|x_it, \alpha_i) = \beta_0(\tau) + x_{it}\beta(\tau) + \alpha_i \tag{18}$$

This assumption has important implications for the identification of the quantile regression coefficients. First, by assuming that α_i is a pure location shift, it reduces the number of parameters that need to be estimated in the model, because α_i an now be estimated globally, while the quantile regression coefficients can be estimated locally. Second, based on the previous insights of the location scale model, it suggests that we can estimate the quantile regression model using a two step approach. First, estimate all location effects using OLS, and then estimate the scale effects using the predicted residuals, but excluding variables we assume have a pure location shifts effect. More formally, the estimator proposed by Canay (2011) can be described as follows:

1. Estimate the location effect of x_{it} and the unobserved heterogeneity α_i on y_{it} using OLS:

$$y_{it} = \beta_0 + x'_{it}\beta_1 + \alpha_i + \varepsilon_{it}$$

2. Use the predicted fixed effects $\hat{\alpha}_i$[8] to transform the dependent variable as $\tilde{y}_{it} = y_{it} - \hat{\alpha}_i$, and estimate the quantile regression model:

$$q_\tau(\tilde{y}_{it}|x_{it}) = \beta_0(\tau) + x'_{it}\beta_1(\tau)$$

This simple approach allows for the identification of the quantile coefficients, by imposing the assumption that the unobserved characteristics only has a location shift effect on the outcome. In addition, like the CRE model, it can be extended to multiple fixed effects, as long as one is willing to assume that the unobserved effects are pure location shifts. For example, consider a case with two fixed effects dimensions (individual and time), under the assumption that the unobserved effects are pure location shifts, the data generating process can be written as:

$$y_{it} = \beta_0(U_{it}) + x'_{it}\beta(U_{it}) + \alpha_i + \alpha_t \tag{19}$$

As before, if we assume that α_i and α_t are constant across quantiles, we could use the same two-step approach to estimate the quantile regression coefficients. First, estimate the location effects using OLS, and then estimate the quantile regression model using the transformed dependent variable, after absorbing the predicted fixed effects.

$$q_\tau(y_{it} - \hat{\alpha}_i - \hat{\alpha}_t|x_{it}) = q_\tau(\tilde{y}_{it}|x_{it}) = x_{it}\beta(\tau) \tag{20}$$

Which again can be easily estimated using standard quantile regression methods, and extended to any number of fixed effects.

8. Empirically, this can be done using `reghdfe` command, as part of the `abs()` suboptions.

2.3 Modified Canay(2011) Estimator

Perhaps one of the main limitations of Canay (2011) is that it assumes that the unobserved effect is a pure location shift. In fact, this is one of the criticisms raised by Machado and Santos Silva (2019) to the estimator. While this makes sense intuitively, because an individual will only be assigned to a single rank at a given point in time, it is not consistent with the idea that the unobserved effect is in fact a proxy for an unobserved characteristic of the individual, and that characeristic could have a different impact on the dependent variable across quantiles. In this case, if this assumption is violated, it may lead to inconsistent estimates of the quantile coefficients. To address this limitation, we propose a small modification to the Canay estimator.

We start by assuming that the unobserved effect represents some characteristics of the individual that are constant across quantiles, and that can be compared across individuals. Under this consideration, the data generating process can be written as:

$$y_{it} = \beta_0 + x'_{it}\beta + \beta_\alpha \alpha_i + \gamma_0(U_{it}) + x'_{it}\gamma(U_{it}) + \gamma_\alpha(U_{it})\alpha_i \tag{21}$$

Where α_i is the unobserved effect, β_α is the location coefficient of the unobserved heterogeneity, and $\gamma(U_{it})$ is a smooth function that varies across quantiles. For the identification of α_i, we start with the same approach as Canay (2011), impossing the asumption that $\beta_\alpha = 1$. In other words, the first step from Canay (2011) is the same as the first step presented before.

The second step, however, suggest that rather than transforming the dependent variable using the predicted fixed effects, we should estimate the quantile regression model using the predicted unobserved effects as an additional explanatory variable. This can be done by estimating the following model:

$$q_\tau(y_{it}|x_{it}, \hat{\alpha}_i) = x_{it}\beta(\tau) + \beta(\tau)\hat{\alpha}_i$$

As before, this model can be extended to multiple fixed effects, by simply estimating the unobserved effects using OLS, and then estimating the quantile regression model using the predicted unobserved effects. The main advantage over Canay (2011) is that this estimator allows for the unobserved effect to have a different impact on the dependent variable across quantiles, which may be more realistic in many applications. However, it assumes the OLS estimator does allow for the consistent estimation of an unobserved effect that is comprarable across individuals, which may not always be the case.

In terms of Standard errors, Canay (2011) provides some guidance for the derivation of analytical standard errors for their estimator. Recently, however, Besstremyannaya and Golovan (2019) has shown that the analytical standard errors derivations are incorrect. Instead, based on their recommendations, we suggest that the bootstrap method should be used for the estimation of the standard errors for both the Canay and Modified Canay estimators.

2.4 Method of Moments Quantile Regression Machado and Santos Silva (2019)

The last methodology we consider is the Method of Moments Quantile Regression (MMQREG) estimator proposed by Machado and Santos Silva (2019), and extended by Rios-Avila et al. (2024). The methodology was proposed as a feasible approach to incorporate fixed effects in a quantile regression model, allowing for unobserved effects to have a different impact on the dependent variable across quantiles. This is done by separating the identification of quantile coefficients into a location, scale and quantile effect, using a method of moments approach.

To understand this approach, lets consider the data generating process from Equation 14. As mentioned earlier, this approach suggests that a quantile regression model can be identified using a location-scale model, where the location effect shows the average effect of the covariates on the dependent variable, and the scale effect shows the heterogenous effect of the covariates, as diversion from the location effect. Machado and Santos Silva (2019) extends this idea, by suggesting that the scale component can be further decomposed into a pure scale effect, and a mediating factor. Specifically, the author considers the case where the data generating process can be written as:

$$\begin{aligned} Y_i &= \beta_0 + x_i'\beta + (\delta_0 + X_i'\delta) * \mu_i) \text{ or} \\ Y_i &= \beta_0 + x_i'\beta + (\delta_0 + X_i'\delta) * F^{-1}(U_i)) \text{ or} \end{aligned} \tag{22}$$

This specification assumes that μ_i is an identically and independently distributed random variable with any arbitrary distribution. $\delta_0 + X_i'\delta$ denotes the multiplicative scale component (heteroskedasticty generating component), and $\beta_0 + x_i'\beta$ denotes the location coefficients. The second line in Equation 22, represents the same model, but using the inverse of the distribution function of U_i as the mediating factor. As before, U_i is a random variable that follows a uniform distribution, and captures the rank of the individual among all individuals with the same characteristics.

What is interesting about this specification is that it simplifies the identification of the quantile coefficients by impossing a strict parametric relationship across the quantile specific coefficients. Specifically, that the location and scale coefficients can be identified globally, requiring only the local identification of the distribution of μ_i being necessary to identify all quantile coefficients.

For this simple case, the MMQREG estimator can be decribed as follows:[9]

1. Estimate the location effect of x_i on y_i using OLS:

$$y_i = \beta_0 + x_i'\beta + R_i$$

9. Further details on the implementation of the methodology can be found in Machado and Santos Silva (2019), with additional extensions in Rios-Avila et al. (2024).

2. Use the predicted residuals $\hat{R}_i$ to estimate the scale effect using OLS, where the dependent variable is defined as $|\hat{R}_i|$:

$$|\hat{R}_i| = \delta_0 + x_i'\delta$$

3. Obtain a standardized residual by dividing the residuals of (1) by the predicted scale effect from (2), and estimate the τth unconditional quantile of this distribution:

$$\begin{aligned} \hat{\mu}_i &= \frac{\hat{R}_i}{\delta_0 + x_i'\delta} \\ q_\tau(\hat{\mu}_i) &= q_0(\tau) \end{aligned}$$

4. Finally, the quantile regression coefficients can be estimated by the following equation:

$$\beta_\tau = \beta + q_0(\tau)\delta$$

It can be seen that because Steps 1 and 2 are estimated globally using OLS, they can easily be extended to multiple fixed effects, without mayor difficulties. One can simply assume, for example, that the fixed effects are estimated using a dummy variable approach, and all the steps described above follow. [10] The derivation of standard errors follows from the use of the method of moments approach, and the identification of the empirical influence functions, based on the following moment conditions:

$$\begin{aligned} E[y_i - \beta_0 - x_i'\beta] &= 0 \\ E[|y_i - \beta_0 - x_i'\beta| - \delta_0 + x_i'\delta] &= 0 \\ E\left[I\left(\frac{y_i - \beta_0 - x_i'\beta}{\delta_0 + x_i'\delta} < q_0(\tau)\right)\right] &= \tau \end{aligned} \tag{23}$$

Despite the flexibility of this approach, it is important to note that the MMQR estimator is based on the assumption that all quantile coefficients are constructed as a simple combination of the location and scale coefficients. They only differ in the value of the mediating factor, the unconditional quantile of the standardized residual. This may be a strong assumption, and may not be applicable in all cases. However, it does provide few advantages over the other approaches. First, as suggested by Machado and Santos Silva (2019), with a sufficiently flexible specification of the scale component, the MMQR estimator can capture most of the important features related to the

10. In Rios-Avila et al. (2024), additional steps are presented because of the use of partial-out covariates. However, the same results can be obtained using the simpler approach described here.

heterogenous effect of the covariates on the dependent variable. Second, because of the identification assumptions, this estimator avoids the problem of quantile crossing[11], which is common in other quantile regression models. Nevertheless, it is important to note that the MMQR assumptions may not be applicable in all cases.

3 Implementation: `qregfe`

While the estimation of the methods described above can be easily implemented using standard Stata commands, the process can be cumbersome and promp to errors. To address this, we introduce the Stata command `qregfe`. This command allows users to estimate quantile regression models with fixed effects using the CRE, Canay, and Modified Canay estimators, as well as the MMQR estimator.[12] The command is designed to be user-friendly, allowing for the estimation of quantile regression models with fixed effects in a single line of code, with only minor changes for model estimation.

The command uses the standard Stata syntax:

```
qregfe depvar [indepvars] [if] [in] [pw], quantile(#) [options]
```

Where `depvar` is the dependent variable, `indepvars` are the independent variables, `if` and `in` are the standard Stata options. Because most of the estimators require the use of bootstrapping, canay and cre estimators do not allow for the use of weights. Only **?** estimator allows for it.

The main options for the command are:

- Fixed effect estimator:
  - `cre`: Correlated Random Effects estimator
  - `canay`: Canay (2011) estimator
  - `canay(modified)`: Modified Canay (2011) estimator
  - `mmqreg`: Machado and Santos Silva (2019) estimator. This is the only estimator also has its on command `mmqreg`.
- `qmethod(,qmethod_options)`: Specifies the method used to estimate the quantile regression component. The default is `qreg`, but any other quantile regression command can be used. This will affect how `quantile(#)` is specified.
  - If necessary, one can also request `qmethod_options` that are specific to the quantile regression command used.
- `q(#)`: specifies the specific quantile for which coefficients will be obtained. The default is the median `q(50)`. The values this could take depend on the `qmethod()`

11. This requires that the linear model used to model heteroskedasticity ($X\delta$) produces a strictly possitive value, at least within the sample range.
12. The MMQR estimator is also implemented by the `xtqreg` command, which was provide by the authors. There is also `mmqreg`, which was originally implemented as a stand alone command, but it is now integrated within this command.

used. For example, if using `qmethod(sqreg)`, one could specify `q(10 25 90)` to obtain the 10th, 25th, and 90th quantiles. However, if one uses `qmethod(qrprocess)` [13], one could specify `q(0.1 0.25 0.9)` to obtain the same quantiles. If using `mmqreg`, multiple quantiles are always allowed.

- `abs(varlist)`: specifies the variables that would be used to absorb the fixed effects. This is necessary for CRE and Canay estimators. For the `mmqreg` estimator, this is not necessary, as the estimator still works for the case of no fixed effects.
- `boot[(bootstrap_options)]`: request bootstrap standard errors to be computed.
  - if not specified, standard errors correspond to the default from `qmethod()`, except for `mmqreg`, which uses the gls standard errors, as proposed by Machado and Santos Silva (2019) and Rios-Avila et al. (2024).
  - `bootstrap options` are the standard options for the `bootstrap` in Stata, using the same results.
- `parallel`: request the estimation of bootstrap standard errors to be performed using `parallel` package (Vega Yon and Quistorff (2019))
  - `parallel_cluster(#)` specifies the number of clusters to be used for the parallelization. The default is 2.
- `seed(#)`: specifies the seed for the random number generator. The default is to use none.
- `robust` and `cluster(varname)`: These options are available when using `mmqreg` estimator. They follow the derivations from Rios-Avila et al. (2024), based on robust and cluster standard errors for gmm estimators..
- `other_options`: Other options are also specific to the `qmethod()` used can also be passed to the command directly.

4

13. `qrprocess` is a user-written command that uses a different algorithm for the estimation of quantile regression models.

5 References

Abrevaya, J. 2013. The projection approach for unbalanced panel data. *The Econometrics Journal* 16(2): 161–178. http://dx.doi.org/10.1111/j.1368-423X.2012.00389.x.

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

Baltagi, B. H. 2023. The two-way Mundlak estimator. *Econometric Reviews* 42(2): 240–246. https://www.tandfonline.com/doi/full/10.1080/07474938.2023.2178139.

Besstremyannaya, G., and S. Golovan. 2019. Reconsideration of a simple approach to quantile regression for panel data. *The Econometrics Journal* 22(3): 292–308. http://dx.doi.org/10.1093/ectj/utz012.

Bottai, M., and N. Orsini. 2019. qmodel: A command for fitting parametric quantile models. *The Stata Journal: Promoting communications on statistics and Stata* 19(2): 261–293. http://dx.doi.org/10.1177/1536867X19854002.

Budig, M. J., and P. England. 2001. The Wage Penalty for Motherhood. *American Sociological Review* 66(2): 204–225. http://dx.doi.org/10.1177/000312240106600203.

Budig, M. J., and M. J. Hodges. 2010. Differences in Disadvantage. *American Sociological Review* 75(5): 705–728. http://dx.doi.org/10.1177/0003122410381593.

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.

Chernozhukov, V., I. Fernández-Val, and B. Melly. 2022. Fast algorithms for the quantile regression process. *Empirical Economics* 62(1): 7–33. https://doi.org/10.1007/s00181-020-01898-0.

Correia, S. 2016. A Feasible Estimator for Linear Models with Multi-Way Fixed Effects. *Unpublished Manuscript* .

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

Hausman, J., H. Liu, Y. Luo, and C. Palmer. 2021. Errors in the Dependent Variable of Quantile Regression Models. *Econometrica* 89(2): 849–873. http://dx.doi.org/10.3982/ECTA14667.

Kaplan, D. M., and Y. Sun. 2017. Smoothed estimating equations for instrumental variables quantile regression. *Econometric Theory* 33(1): 105–157. Publisher: Cambridge University Press.

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.

Mundlak, Y. 1978. On the Pooling of Time Series and Cross Section Data. *Econometrica* 46(1): 69. http://dx.doi.org/10.2307/1913646.

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

Rios-Avila, F. 2015. Feasible Fitting of Linear Models with N Fixed Effects. *The Stata Journal* 15(3): 881–898. Publisher: SAGE Publications.

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

Vega Yon, G. G., and B. Quistorff. 2019. parallel: A command for parallel computing. *The Stata Journal* 19(3): 667–684. https://doi.org/10.1177/1536867X19874242.

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.

———. 2021. Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators. Working paper. Available at SSRN: https://ssrn.com/abstract=3906345.

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.