Semi & non-parametric econometrics

Dissertation

Nonlinear Panel Data Estimation via Quantile Regressions

Arellano M. & Bonhomme S.

Samuel Ritchie, Antoine Roy, Julien Sauvan - 3A

1 Introduction

Since the 1980s, a new literature in econometrics has focused on heterogeneous effects, while previous literature had so far only focused on average effects within a given population. Koenker and Bassett (1978) were the first to introduce the so-called quantile regression, which has now become a common tool in econometrics. Compared to the simple linear regression, this model has the advantage of accounting for heterogenous effects. Indeed, the effect of a variable may not be the same for all the individuals, and simple linear regression parameter estimates therefore not be fully satisfying to draw conclusions. Moreover, this method is also robust to outliers and heavy tails.

This model was first limited to the analysis of cross sectional data. Koenker himself (2004) has then extended the theory to adapt quantile regression to panel data. This method provides two solutions to some common issues in econometrics. First, as said before, it allows to study the heterogeneity of the effect over the outcome. It also includes a fixed effect, solving the possible endogeneity problem due to some unobserved heterogeneity. Following these papers, large literature has addressed several dimensions of quantile regression.

Even if basic panel data quantile regression was of a huge help to the scientific community, this model remains unperfect as several papers have addressed its failures. Galvao (2011) shows that the model is biased in the presence of lagged dependent variables as regressors. Nonetheless, he proposes the use of an instrumental variable with lagged regressors to solve the problem. Arellano and Weidner (2015) also identified a problem of endogeneity in short panels. In this paper, the authors conduct their analysis for a fixed period but with the number of individuals tending to infinity. In that precise case, they are not concerned by these asymptotic issues.
Several researchers have since revisited this quantile panel data model, adding new elements, considering specific cases like separable panel models (Chernozhukov et al.), including multiple random effects (Geraci and Bottai, 2007). However, the main drawback of these revisited models remains that they are systematically defined in a linear framework.

The paper we study in the following dissertation, written by Arellano M. & Bonhomme S. in 2016, aims at contributing to this literature, developing an estimation strategy for nonlinear panel models. It can thus be seen an extension of the standard linear panel data methods (Koenker, 2004) to nonlinear settings.
This interest in quantile regression with panel data is motivated by the possibility to deal with complex interactions between covariates and latent heterogeneity. This model allows the authors to build flexible models for the dependence of unobserved heterogeneity on exogenous covariates or initial conditions.
The objective of the authors is thus to build a framework that can deal with general nonlinear and dynamic relationships, providing an extension of standard linear panal data methods to nonlinear settings. They also extend this methodology to a dynamic setting.

In the following dissertation, we will first present the general and theoretical framework of Arellano M. & Bonhomme S. 's paper. While they develop both a static and a temporal dynamic model, we will exclusively concentrate our study on the static restriction. As an application, we will then apply their model to estimate the impact of union membership on the wages in the US.

2 Theoretical method

2.1 Quantile regressions : a brief reminder

The most widely applied econometric model in literature is undoubtedly linear ordinary least square regressions (OLS), as it the most simple on terms of specification and parameter interpretation. However, one of the main drawbacks of simple OLS approach is the fact it focuses on average effects, while the effect of a variable may not be the same for all individuals. This heterogeneity regarding the value of parameters may be very important for public policy, for example when studying the impact of an increase of the minimum wage on wages. The effect of such an increase is likely to be much larger for low wages than for high ones, which is not taken into account when evaluating the average effect by standard OLS regression.
As a consequence, quantile models specify the error term as a function of the τ-th quantile of the explained variable, which implies that the estimated parameter will also depend on these quantiles :

$$Y = X^{'}\beta_\tau + \epsilon_\tau \quad , \quad q_\tau(\epsilon_\tau|X) = 0$$

where the τ-th quantile of a random variable U is defined by $q_\tau(U) = \inf\{x/F_U(X) \leq \tau\}$.

Equivalently, the above formula can be re-written in terms of quantiles of Y given X, i.e. $q_\tau(Y|X) = X^{'}\beta_\tau$. Using this definition, we obtain that similarly as in linear regressions, β_τ satisfies

$$\beta_\tau = \frac{\partial q_\tau(Y|X=x)}{\partial x} = \mathbb{E}\left(\frac{\partial q_\tau(Y|X)}{\partial x}\right)$$

which is the marginal effect of X on the conditional quantile of Y. Even if it is tempting to interpret β_τ as the effect of a small variation in X for individuals located at the τ-th quantile of $Y|X = x$, this is only the case under rank invariance condition, which happens if individuals have the same ranking of the distribution of $Y(x)$ whatever x.

In order to make an easier parallel with the model developed in our article, we use another alternative representation of linear quantile models, which is the more general random coefficient model :

$$Y = X^{'}\beta_U \quad , \quad U|X \sim \mathcal{U}([0,1]) \tag{1}$$

where $\tau \mapsto x^{'}\beta_\tau$ is supposed to be strictly increasing. This specification considers a unique underlying variable U which determines the ranking of each individual in terms of Y, within his subpopulation X. This vision of linear quantile models fits with the traditional one defined above insofar as

$$P(Y \leq X^{'}\beta_\tau|X) = P(X^{'}\beta_U \leq X^{'}\beta_\tau|X) = P(U \leq \tau|X) = \tau$$

2.2 Generic formalization of the problem

In their article, Arellano M. & Bonhomme S. aim to take into account non-linearity regarding a panel data structure. Moreover, authors aim to specify individual heterogeneity in order to avoid possible endogeneity issues.
To do so, the first step is to specify outcome variable Y_{it} as a non-linear function of both covariates X_{it} and latent heterogeneity η_i :

$$Y_{it} = \sum_{k=1}^{K1} \theta_k(U_{it}) g_k(X_{it}, \eta_i) \tag{2}$$

where $U_{i1}, ...U_{iT}$ are independant uniform random variables.
Regarding these notations, non-linearity is taken into account through the function g and individual effects through η_i, enabling us to avoid endogeneity of the error term. As it is written, the quantile aspect of the model cannot be seen. It will depend on the term $\theta_k(U_{it})$ and assumptions made regarding it, as we will see in the following subsection[1].

To fully define the model, it is also necessary to specify the dependence of η_i on covariates $X_i = (X'_{i1}..., X'_{iT})$:

[1] An analogy can be made between $\theta(U)$ and the β_U term of the general random coefficient representation of quantile models

$$\eta_i = \sum_{k=1}^{K2} \delta_k(V_i) h_k(X_i) \tag{3}$$

with V_i is a uniform random variable and the function h, similarly to g, accounts for the non-linear aspect of our model.

The approach of the article contains two major challenges :

- The first challenge is econometric : the individual heterogeneity is unknown. Indeed, if it was known, a simple ordinary (non-linear) quantile regression would be sufficient to estimate model parameters. As a consequence, one of the algorithm steps will consist in drawing values $\eta_i^{(m)}$ conditionnaly to the data. The conditional aspect (relatively to θ and δ parameters) implies we will go back and forth between updating parameters and drawing new values of η
- The second challenge is a computational / mathematical one. Insofar as θ and δ parameters are functions, it will be necessary to approach them by finite-dimensional approximations, such as splines.

Y and η being fully defined by the previous equations, the algorithm developed by the authors can then simply be seen as a variant of the Expectation-Maximisation (EM) algorithm. EM algorithm is a widely used iterative method to find maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables. The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step.
In our case, we do not update parameters using the log-likelihood function, but rather using another objective function : the quantile regression one, which has the good properties of being convex. The modified EM algorithm can thus be seen as an alternation of the following steps for a grid of quantiles τ and initial values of θ and δ parameters:

1. (E) Step : Given θ and δ, compute distribution of individual effect η and a draw sequence $\{\eta_i^{(m)}\}$ according to this distribution
2. (M) Step : Update θ and δ using quantile regression objective functions, respectively using equations (2) and (3)

The detailed process of how we draw the η sequence, as well as the way the (M) step is performed will be detailed in a further subsection, after having defined the exact static panel data model we restricted our interest to and that we applied to our data.

2.3 Nonlinear quantile models for panel data (static framework)

In their article, the authors introduce two main non-linear quantile models : a static and a dynamic one. Given that our simulation will only use the first one, we will not give a full description of the dynamic model. Therein, the following subsection exclusively focuses on the description of the static model.
As evoked in the previous part, the generic formalization introduced did not exhibit properly the quantile dependence of the outcome variable Y_{it}. To do so, using the same notations as previously, we define the conditional quantile response of Y_{it} conditionally to X_{it} and η_i as :

$$Y_{it} = \mathcal{Q}_Y(X_{it}, \eta_i, U_{it}) = W_{it}(\eta_i)' \theta(U_{it}) \tag{4}$$

This model allows covariate effects to differ both across individuals (through η_i) and naturally across the distribution of the outcome variable (through U_{it}), which is the study framework of the authors. Besides, if we use a series specification for W, i.e. if we set $W_{it}(\eta_i) = (g_1(X_{it}, \eta_i), ..., g_{K_1}(X_{it}, \eta_i))$, we obtain exactly the equation (2).

Similarly as in the random coefficient model (1), we make two assumptions on U_{it} in order to be able to properly define quantiles regarding these notations :

- U_{it} follows a standard uniform distribution, and is independent of (X_i, η_i) (independence of regressors and individual effects)
- $\tau \mapsto \mathcal{Q}_Y(x, \eta, \tau)$ is strictly increasing on $(0, 1)$ for almost all (x, η) in the support of (X_{it}, η_i)

Following these two assumptions, it comes that :

$$P(Y_{it} \leq \mathcal{Q}_Y(X_{it}, \eta_i, \tau)|X_i, \eta_i) = P(\mathcal{Q}_Y(X_{it}, \eta_i, U_{it}) \leq \mathcal{Q}_Y(X_{it}, \eta_i, \tau)|X_i, \eta_i) = P(U_{it} \leq \tau|X_i, \eta_i) = \tau$$

which, in other terms, shows that $\mathcal{Q}_Y(X_{it}, \eta_i, \tau)$ is indeed the conditional quantile of Y_{it} conditionally to X_i and η_i.
A last assumption is made regarding the error term U which is that for different times t, U_{it} are independent one another. This restriction is necessary to separate time-varying unobserved errors U_{it} from the time-invariant unobserved individual effects η_i, and has a important consequence on the choice of our explanatory variables for the application. This assumption typically disappears when considering dynamic models, which the authors do in a second part of their presentation.

The same specification as the one just detailed for the outcome variable is made for the unobserved heterogeneity term η_i, i.e. :

$$\eta_i = \mathcal{Q}_Y(X_i, V_i) = Z_i^{'}\delta(U_{it})$$

Same assumptions than the two first ones for U_{it} are made regarding V_i, as in the random coefficient model. Furthermore, taking $Z_i = (h_1(X_i), ..., h_{K_2}(X_i))$ leads to equation (3).

2.4 Estimation Strategy

In this subsection, we give detailed regarding the strategy used for parameters estimation, which leads to the modified stochastic EM algorithm approach briefly presented in 2.2.

Deriving first order conditions
In order to explain the parameter estimation strategy developed by the authors, it is necessary to introduce the concept of check functions. Check functions are widely used in quantile models, and are defined as

$$\rho_\tau : u \mapsto (\tau - \mathbb{1}\{u < 0\})u$$

For ordinary quantile regressions, it has been shown that $q_\tau(Y) \in \arg\min_a \quad \mathbb{E}\left(\rho_\tau(Y - a)\right)$. Similarly to conditional expectation, we can extend the reasoning to conditional quantiles. As a consequence,

$$q_\tau(Y|X = x) \in \arg\min_a \quad \mathbb{E}\left(\rho_\tau(Y - a)|X = x\right)$$

Thus, integrating over P^X :

$$(x \mapsto q_\tau(Y|X = x)) \in \arg\min_{h(.)} \quad \mathbb{E}\left(\rho_\tau(Y - h(X))|X = x\right)$$

Using the fact that in our precise case, we have $\mathcal{Q}_Y(X_{it}, \eta_i, U_{it}) = W_{it}(\eta_i)^{'}\theta(U_{it})$ (which we take as the function h in the above equation) and defining $\psi_\tau(u) = \frac{d\rho_\tau(u)}{du}$, first order conditions come straight-forward as :

$$\mathbb{E}\left[\sum_{t=1}^{T} W_{it}(\eta_i)\psi_\tau(Y_{it} - W_{it}(\eta_i)^{'}\theta(\tau))\right] = 0 \quad , \quad \forall\tau \in (0, 1) \tag{5}$$

Applying exactly the same logic for the unobserved heterogeneity term leads to :

$$\mathbb{E}\left[Z_i\psi_\tau(\eta_i - Z_i^{'}\delta(\tau))\right] = 0 \quad , \quad \forall\tau \in (0, 1) \tag{6}$$

In order to introduce the density of unobserved heterogeneity, we apply the law of iterated expectations $\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|Y))$ to the two above equations. Expressing the expectation of $f(\eta)$ as an integral, it comes that :

$$\mathbb{E}\left[\int \left(\sum_{t=1}^{T} W_{it}(\eta)\psi_\tau(Y_{it} - W_{it}(\eta)^{'}\theta(\tau))\right) f(\eta|Y_i, X_i; \theta(.), \delta(.))d\eta\right] = 0 \quad , \quad \forall\tau \in (0, 1)$$

and

$$\mathbb{E}\left[\int \left(Z_i\psi_\tau(\eta - Z_i^{'}\delta(\tau))\right) f(\eta|Y_i, X_i; \theta(.), \delta(.))d\eta\right] = 0 \quad , \quad \forall\tau \in (0, 1)$$

where the posterior density of η can be expressed thanks to the Bayes formula and law of total expectations, and using independence assumption :

$$f_{\eta|Y,X}(\eta\ y, x; \theta(.), \delta(.)) = \frac{\prod_{t=1}^{T} f_{Y_t|X_t,\eta}(y_t|x_t, \eta; \theta(.)) f_{\eta|X}(\eta|x, \delta(.))}{\int \prod_{t=1}^{T} f_{Y_t|X_t,\eta}(y_t|x_t, \tilde{\eta}; \theta(.)) f_{\eta|X}(\tilde{\eta}|x, \delta(.)) d\tilde{\eta}}$$

Identification of the above quantity has been shown in related work from Hu & Schennach (2008), regarding a different context of instrumental variables, but a parallel is made by the authors of our article in our context.

Functional approximation

The main difficulty in the first order conditions specified above is that the posterior density of η depends on a continuum of parameters through the undefined functions θ and δ. It is thus necessary to approximate these functions by finite-dimensional ones. To do so, authors use piecewise-linear splines as suggested by Wei & Carroll (2009). The idea is simple : for a given number of knots L (complexity and precision increase with L), we approximate θ and δ functions respectively by $\zeta_A = (\theta(\tau_1)^{'}, ..., \theta(\tau_L)^{'})^{'}$ and $\zeta_B = (\delta(\tau_1)^{'}, ..., \delta(\tau_L)^{'})^{'}$, where $\tau_l \in (0, 1)$.

Eventually, with this approximation, we obtain the final integrated moment restrictions of the approximating model : $\forall l \in 1, .., L$:

$$\mathbb{E}\left[\int \left(\sum_{t=1}^{T} W_{it}(\eta)\psi_\tau(Y_{it} - W_{it}(\eta)^{'}\theta(\tau_l))\right) f(\eta|Y_i, X_i; \zeta) d\eta\right] = 0 \quad , \quad \forall \tau \in (0, 1)$$

and

$$\mathbb{E}\left[\int \left(Z_i \psi_\tau(\eta - Z_i^{'}\delta(\tau_l))\right) f(\eta|Y_i, X_i; \zeta) d\eta\right] = 0 \quad , \quad \forall \tau \in (0, 1)$$

In order to model quantile functions in the intervals $(0, \tau_1)$ and $(\tau_L, 1)$, the simplest way is to assume that θ and δ are constant on these intervals. However, the authors also implement in alternative more precise exponential-based modelling. This enables to avoid the fact that the support of the likelihood function depends on the parameter value.

Final estimation algorithm

We now have all the necessary quantities to compute the final version of the estimation algorithm, which is based on a modified version of stochastic EM algorithm as explained before. Noting $(Y_i, X_i^{'})$ an i.i.d. sample of our data, the estimation algorithm is composed of the following steps (2-4), that we iterate until convergence to a stationary distribution.

1. Initialize $\widehat{\zeta^{(0)}}$
2. (E Step) $\forall i \in [1, N]$, compute the posterior density $f_i^{\hat{(s)}}(\eta) = f(\eta|Y_i, X_i, \zeta^{\hat{(s)}})$
3. Draw M values $\eta_i^{(1)}, ..., \eta_i^{(M)}$ from $f_i^{\hat{(s)}}$ using a Metropolis-Hastings algorithm[2]
4. (M Step) Solve[3], $\forall l \in [1, L]$

$$\hat{\theta}(\tau_l)^{(s+1)} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{m=1}^{M} \sum_{t=1}^{T} \rho_{\tau_l}(Y_{it} - W_{it}(\eta_i^m)'\theta)$$

$$\hat{\delta}(\tau_l)^{(s+1)} = \underset{\delta}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{m=1}^{M} \rho_{\tau_l}(\eta_i^m - Z_i'\delta)$$

[2]MH algorithm, that we will not detail here, is a standard Markov Chain Monte Carlo method for obtaining a sequence of random samples from a probability distribution for which direct sampling is difficult

[3]In order to compute the quantities, we replaced integrals by approximate sums

2.5 Interprating quantile marginal effects

In nonlinear panel data models, it is often of interest to compute the effect of marginal changes in covariates on the entire distribution of outcome variables.
In the case of a continuous variable X_{it}, the average quantile marginal effect (QME) is defined as :

$$M(\tau) = \mathbb{E}\left[\frac{\partial \mathcal{Q}_Y(X_{it}, \eta_i, \tau)}{\partial x}\right]$$

where $\partial \mathcal{Q}_Y / \partial x$ is the vector of partial derivatives with respect to its first $dim(X_{it})$ components.

In the case of a binary variable $D_{it} \in \{0, 1\}$, if we note X_{it} all the other time-varying covariates, the average quantile marginal effect is defined as :

$$M(\tau) = \mathbb{E}\left[\mathcal{Q}_Y(1, X_{it}, \eta_i, \tau) - \mathcal{Q}_Y(0, X_{it}, \eta_i, \tau)\right]$$

where $\mathcal{Q}_Y(0, X_{it}, \eta_i, U_{it}) = Y_{it}(d)$, d being 0 or 1.

3 Application

In this section we study the impact of union membership on the wages using the method presented above. We also present basic econometric methods to compare the estimated effects of union membership. To do so, we revisit Vella and Verbeek 1998's paper, using similar data on wages, union membership and other characteristics.

3.1 Research question

A primary goal of trade unions is to maintain and improve worker's term conditions. In the US, unions can influence wage levels through bargaining when wage levels are set in collective agreement. Large literature has tried to identify the impact of union membership on wages. This literature tries to identify how wage differs in union and non-union employment through what we will call the *union effect*. As Robinson (1989) has noticed, very different results can be found depending on the methods that are used. Nonetheless, Graham et al. (2015) have revisited the union effect using quantile regression with panel data but still with a linear relationship.

Indeed, one of the main challenge in the identification strategy is that the unobserved factors that influence union membership can also affect wages. Simple OLS regression often fails to identify the true effect given that it is not appropriate to capture the unobserved individual heterogeneity which can be responsible for union membership. Thus, panel data appears as a solution to control for this endogeneity through individual fixed effects. This is what Robinson or Vella and Verbeek have implemented. They decompose the endogeneity underlying union status and individual-specific component and an invidual/time-specific effect.

However, in this literature, authors always identify an average effect through OLS regression or panel data method. In this study, we used quantile regression with individual specific fixed effect effects to allow for both unobserved heterogeneity and nonlinearity in the relationship between wages and union membership.

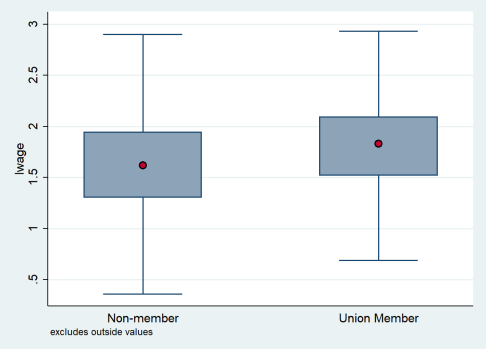
3.2 Data

Data is taken from the National Longitudinal Survey (Youth Sample) for the period 1980-1987 and is restricted to young men. We used exactly the same panel as Vella and Verbeek, it comprises a sample of about 600 full-time working males who have completed their schooling by 1890 and then followed over the period 1980 to 1987. Only individuals providing full information on their job situation were kept. The study is restricted to males in order to examine a relatively homogeneous group.

We only kept the variables that were time-variating, given that the η_i in our model allows us to capture the individual fixed effect. Our dependant variable is the log of hourly wage. We regressed it on three different variables. The experience squared and the marrital status which are the only time-variating variables in the dataset were used as control. The main variable of interest, union membership, was based on the answer to the question reflecting whether or not the individual had his wage set in a collective bargaining agreement. Thus, when we talk about an union effect, we in fact talk about a "collective bargaining" effect.

In the dataset, about 25% of the workers are members of an union. As we can see in the following figure, basic descriptive statistics show that wages are a bit higher for union member. This suggests that we should expect a positive union effect. However controling for other variables and including an individual fixed effect should bring more convincing results.

Figure 1: Boxplot of log wage depending on union membership



.

3.3 Method

We introduce a model of wage determination and union status. Using previous notations:

- Y_{it}= log wage of individual i at time t
- X_{it}= (experience squared at time t for individual i, marital status at time t for individual i, union membership status at time t for individual i)
- η_i= fixed effect of individual i capturing unobservable characteristics

In our model, the log wage is determined by some individual characteristics, the experience, the marital status and the union effect. We use three different methods to estimate this model of wage determination and compare union effect estimated each time.

3.4 Results

3.4.1 Linear Regressions

We first ran an OLS regression to get an insight of the impact of union membership on wages. Results are provided in the table below. We can see that union membership seems to have a positive effect on the log of wages with a significant coefficient of 0.170 (what is consistent with Graham et al. (2015)). As expected, the *union effect* is positive. However, we can wonder whether the magnitude is exact. As said before, the literature suggests that union membership is endogenous, which biases our results.

VARIABLES	OLS Results log(wage)
Union	0.170***
	(0.0162)
Married	0.190***
	(0.0182)
Experience squared	0.00118***
	(0.000197)
Constant	1.465***
	(0.0139)
Observations	4,360
R^2	0.068

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

To control for the endogeneity of the union membership variable, we therefore also ran a panel regression with individual fixed-effects, leading to the following results:

VARIABLES	Panel regression with fixed-effects log(wage)
Union	0.0828***
	(0.0198)
Married	0.107***
	(0.0182)
Experience Squared	0.00370***
	(0.000189)
Constant	1.395***
	(0.0123)
Observations	4,360
Number of individuals	545
R-squared	0.137

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

We can here observe that the effect of union membership is twice lower than in OLS (0.08 against 0.17, still consistent with Graham et al., 2015). This result confirms our assumption. If union membership was not correlated with some individual unobserved characteristics, the *union effect* would have been the same with both estimation methods. Thus, OLS estimate leads to an overestimation of the *union effect*, due to a correlation with unobserved characteristics that are captured by the individual fixed effect in the panel data regression. A Hausmann test significantly reject (at the 0.1 level) random effect.

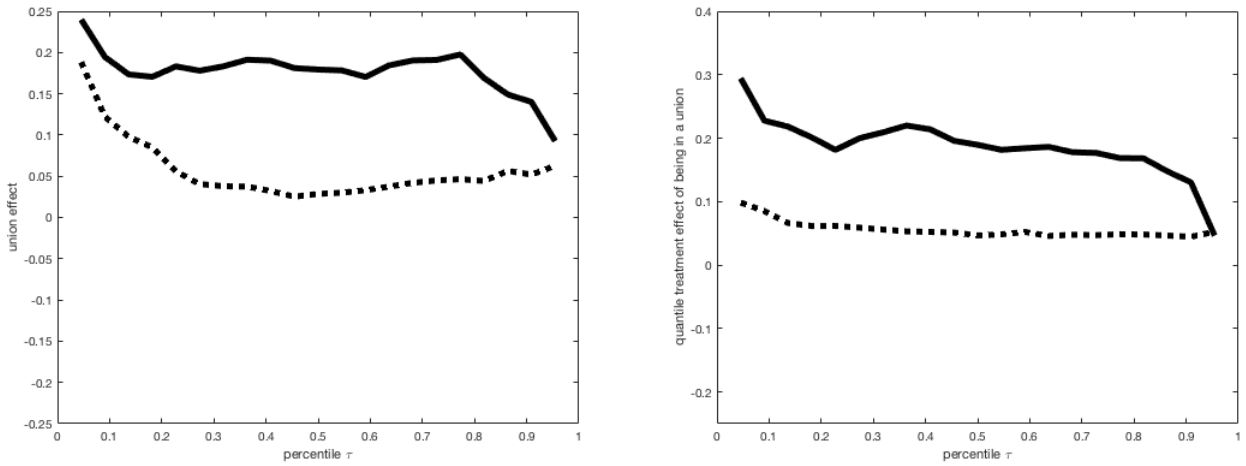
These basic results justify the use of an individual fixed effect in the following method. It also gives us an idea of the magnitude of the *union effect* parameter to be compared with when applying Arellano M. & Bonhomme S. method.

3.4.2 Non linear panel data estimation via quantile regression

We now use the nonlinear panel data method that we have presented before. The main outputs of the authors' implementation of the algorithm are different figures allowing to easily see and interpret the quantile effects of a covariate.

The left graph of figure 2 allows to compare the pooled quantile regression estimations of union membership (the solid line) to the estimation done by the algorithm (the dashed line). According to the solid line, the effect of union membership appears to be more important at lower quantiles of wages: the curve is decreasing with percentiles.

Figure 2: Quantile effects of being in a union on log-wages (linear quantile specification)



Left graph: solid line is the pooled quantile regression union coefficient ; dashed line is the panel quantile regression union coefficient.
Right graph: solid line is the raw quantile treatment effect of union membership ; dashed line is the quantile treatment effect estimate based on panel quantile regression

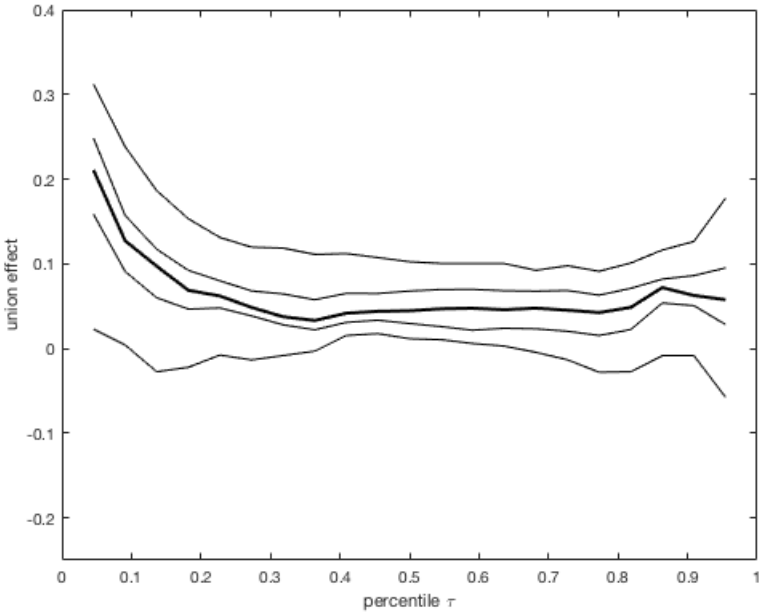
The method of panel quantile regression, done in the same framework as in the seminal paper *(linear quantile regression specification as in Example 2, augmented with a parametric exponential model in the tails intervals, with 21 knots, 100 iterations of the stochastic EM algorithm and 100 random walk Metropolis-Hastings draws within each iteration)* shows the union membership effect is less important when correcting for time-invariant endogeneity with an individual-specific fixed effects. This effect also seems to have a slightly lower impact on median percentiles, as the curve appears convex.

On the right graph, the solid line shows the empirical difference between unconditional quantiles, while the dashed line shows the quantile treatment effect that accounts for both observables and unobservables, using a semi-parametric model of the joint distribution of outcomes and unobservables. Results are similar to the left graph in the way that there is a positive overestimation of the union membership effect in pooled quantile regression, but here the effect measured seems to be more constant along the different percentiles of wages, even if slightly decreasing.

Both graphs allow to see the positive endogeneity bias as the difference between the solid line and the dashed line. So, our method allows us to correct this overestimation bias mostly thanks to the individual fixed-effect.

Figure 3 shows the difference in union effect for each estimation of the model (for each quantile). We see that the union effect is very heterogenous for the lower quantiles. That means that for people who earn the less, being member of an union can have a very high effect on wage as it can have no effect at all. For people at the median, the union effect is quite similar for everybody but it is, in average, much lower than for the people who earn the less. Surprisingly, the *union effect* is also greater at the top of the wage distribution, but it is also heterogeneous. It can even be negative for people at the higher quantile.

Figure 3: Quantile effects of being in a union on log-wages (interacted quantile specification)



Lines represent the percentiles 0.05, 0.25, 0.5, 0.75 and 0.95 of the heterogeneous union membership effect across individuals, at various percentiles.

3.4.3 Conclusion

Finally, this model provides new evidence on the *union effect* compared to the previous findings (from linear modeling). First, it allows to study the effect over the distribution of wages what could not be done when studying an average effect. We see that the *union effect* is larger for people who earn the less (and above the average effect computed with the panel data model with fixed effects) but not as much as we could have expected just looking at the raw quantile effect. This effect is also a decreasing function of the quantile considered. Thus, collective bargaining allows for a rent extraction for the workers mostly at the bottom of the wage distribution.

Second, the model brings new evidence of the heterogeneity of the *union effect*. Although it is larger for people who earn the less, the *union effect* is also very heterogeneous for them. Indeed, there is a huge difference between the percentiles 0.05 and 0.95 of the union effect.

The authors' model allowed to revisit Vella and Verbeek's results and provide new evidence on the *union effect*. So far, the literature had mostly focused on an average effect within the population computed with linear models, two restriction that disappear with this approach.

4 Bibliography

Arellano, M. and M. Weidner (2015). Instrumental variable quantile regressions in large panels with fixed effects. Working paper, CEMFI.

Galvao, A. F. (2011). Quantile regression for dynamic panel data with fixed effects. Journal of Econometrics 164, 142–57.

Geraci, M. and M. Bottai (2007). Quantile regression for longitudinal data using the asymmetric Laplace distribution. Biostatistics 8, 140–54.

Graham B, Hahn J., Poirier A. and Powell J. (2015), Quantile regression with panel data, WP

Koenker, R. and G. J. Bassett (1978). Regression quantiles. Econometrica 46, 33–50.

Koenker, R. (2004). Quantile regression for longitudinal data. Journal of Multivariate Analysis 91,

Robinson, C. (1989), 'The joint determination of union status and union wage effects: some tests of alternative models', Journal of Political Economy

Vella F. and Verbeek M. (1998), Whose wages do unions raise. A dynamic model of unionism and wage rate determination for young men, Journal of Applied Econometrics 13, 163-183