Minimum Distance Estimation of Quantile Panel Data Models*

Blaise Melly† and Martina Pons‡

First version: November 2020
This version: September 16, 2022
Preliminary version. Link to the newest version

Abstract

We propose a minimum distance estimation approach to quantile panel data models where the individual effects may be correlated with the covariates. The estimation method is computationally straightforward to implement and fast. We first compute a quantile regression within each individual and then apply GMM to the fitted values from the first stage. The suggested estimators apply (i) to classical panel data, where we follow the same units over time, and (ii) to grouped data, where we observe data at the individual level, but the treatment varies at the group level. Depending on the variables assumed to be exogenous, this approach provides quantile analogs of the classical least squares panel data estimators such as the fixed effects, random effects, between, and Hausman-Taylor estimators. For grouped (instrumental) quantile regression, we provide a more precise estimator than the existing estimators. We establish the asymptotic properties of our estimators when both the number of units and observations per unit jointly diverge to infinity. Monte Carlo simulations show that our estimator and inference procedure also perform well in finite samples when the time-series dimension is small.

1 Introduction

Quantile regression, as introduced by Koenker and Bassett (1978), is the method of choice when we are interested in the effect of a policy on the distribution of an outcome. The quantile treatment effect function provides more information than the average treatment effect; for instance, it allows evaluating the impact of the treatment on inequality. When panel data are available, new identification and estimation strategies become feasible. The researchers can alleviate endogeneity concerns, for instance, by allowing for correlated individual effects. They can obtain more precise estimates, for example, by using a random-effects estimator; or they can exploit time-varying variables to identify the impact of time-invariant variables, e.g., with the Hausman and Taylor (1981) estimator. In this paper, we propose a minimum distance

*We are grateful to Bo Honoré, Aleksey Tetenov and seminar participants at Bern University (4th of November 2020), the 2021 scientific meeting of the Italian statistical society (24th of June 2021), Geneva University (3rd of March 2022), the 27th International Panel Data Conference, and the COMPIE 2022 Conference, for useful comments.

†Department of Economics, University of Bern, Schanzeneckstrasse 1, 3001 Bern, Switzerland. blaise.melly@unibe.ch.

‡Department of Economics, University of Bern, Schanzeneckstrasse 1, 3001 Bern, Switzerland. martina.pons@unibe.ch.

estimation approach to quantile panel data models, which provides quantile analogs of the classical least-squares panel data estimators such as the fixed effects, random effects, between, and Hausman-Taylor estimators.

We use the notation (i and t subscripts) and terminology (individuals and time periods) commonly used in the panel data literature. However, our results also apply to grouped data, where we observe data at the individual level, but the treatment varies at the group level. In this part of the literature, the i units are often called groups, and the t units are individuals within these groups. For instance, in Autor et al. (2013), the groups are commuting zones in the United States while they are schools in Angrist and Lang (2004). In both cases, the treatment varies only between groups, but individual data are needed to estimate the conditional distribution of the outcome within each group. We discuss the application of our results to this framework in Section 5.

In all cases, we perform the estimation in two stages. The first stage consists of individual-level quantile regressions using time-varying covariates at each quantile of interest. In the second stage, the first-stage fitted values are regressed on time-invariant and time-varying variables. If these variables are potentially endogenous, an instrumental variable regression or, more generally, the generalized method of moment (GMM) estimator can be used. Thus, including external or internal instruments in the second stage is straightforward. This estimator is simple to implement, flexible, computationally fast and can be used in various applied fields. While this two-step procedure may sound unusual, Section 2.3 shows that it is numerically identical to the standard estimators (fixed effects, random effects, Hausman-Taylor) if we use least-squares in the first stage and the appropriate instruments.

As a nonlinear estimator, quantile regression with fixed effects is subject to the incidental parameter problem with a bias of order T^{-1}. Inference is justified using a 'large N, large T' asymptotic framework.[1] Recently, Galvao et al. (2020) have weakened the requirements on the relative rate of divergence of N and T for asymptotic normality of fixed effects quantile estimators. Using their results, we show that our estimator is asymptotically normal under the condition that $N(\log T)^2/T \to 0$. Under this condition and other assumptions, we show that our estimators are asymptotically normally distributed and centered at zero. The requirement on the growth rate of T relative to N can be weakened if only the coefficient vector on time-constant regressors is of interest. In this case, the milder condition that $\sqrt{N}(\log T)/T \to 0$ is sufficient for an unbiased asymptotic distribution.

The asymptotic distribution of the estimator is non-standard because the speed of convergence is not the same for all coefficients. The coefficients of the variables that are identified by time-varying instruments converge at the $\sqrt{NT}$ rate (the 'fast' coefficients) while the other converge only at the $\sqrt{N}$ rate (the 'slow' coefficients). This has several consequences on the first-

[1] Large T asymptotic has been widely used in the quantile and nonlinear panel data literature as well as in the nonstationary and dynamic panel data literature. For seminal contributions, see Phillips and Moon (1999), Hahn and Kuersteiner (2002), and Alvarez and Arellano (2003).

order asymptotic distribution: (i) The fast and slow coefficients are first-order asymptotically independent. (ii) If a coefficient is identified by time-varying and time-constant instruments, then the time-constant instruments are first-order asymptotically useless. For instance, the random-effect estimator is asymptotically equivalent to the fixed-effects estimator.[2] (iii) The first-order asymptotic distribution of the estimator of the slow coefficients is not affected by the first-stage estimation error. To improve inference in finite samples, we derive asymptotic results where we keep track of higher-order terms. Using this better approximation, we solve these three issues. This allows us to suggest a new quantile random effects estimator that is more precise than the fixed effects estimator in finite samples. We also improve the quality of the estimated standard errors by taking the first stage estimation error into account. Quite surprisingly, we find that clustering the standard errors in the second stage takes automatically into account the first-stage error. Thus, inference does not require estimating the density like in traditional quantile models.

This paper contributes to the literature on quantile panel data and IV models. Koenker (2004) introduced a penalized quantile fixed effects estimator that treated the individual heterogeneity as a pure location shift. A large share of the literature focused on fixed effects models (see, for example, Canay, 2011; Galvao and Kato, 2016; Gu and Volgushev, 2019). Kato et al. (2012) allow the individual effects to depend on the quantile of interest and contribute to the asymptotic theory of the estimator. Galvao and Wang (2015) suggest a two-step minimum distance (MD) estimator as a computationally fast way to estimate fixed effects quantile panel data model. Galvao and Poirier (2019) suggest using quantile regression as an estimator in the presence of random effects. Our random effects estimator is different because we focus on the conditional quantile function that also conditions on the individual effects. In other words, we estimate a different parameter, and quantile regression is not consistent for this parameter even if the random effects are not correlated with the covariates.

Our class of estimators nests the MD estimators of Chamberlain (1994) and Galvao and Wang (2015) as special cases. We generalize the results in Chamberlain (1994) by including time-varying regressors and allowing the number of individuals to go to infinity.[3] Galvao and Wang (2015) are only interested in the effect of time-varying covariates and do not exploit variation between individuals. In contrast, we aim to estimate the effect of both time-varying and time-invariant regressors. Furthermore, we allow for both internal and external instruments.

Chetverikov et al. (2016) consider a quantile extension of the Hausman-Taylor model. They focus on the effect of variables that vary only between individuals (between groups using their terminology) and allow for instrumental variables for identification. We can also apply our approach to this setup. The main difference is that, in the second stage, we regress the fitted

[2] See Ahn and Moon (2014) for similar results for least-squares estimators.

[3] Chamberlain (1994) uses a different terminology because he considers cross-sectional regressions. He analyses a quantile regression model with a finite number of combinations of values of the regressors. The number of cells is thus finite, and the regressors are constant within each cell.

values on all variables while they regress the estimated intercept on the time-invariant regressors. Since they use only the intercept in the second stage, their estimator is not invariant to reparametrizations of the time-varying regressors. In addition, by keeping all the variables in the second stage, we can easily impose equality of the coefficients on the time-varying regressors and increase precision at a minimal cost from a computational perspective. Simulations using the same data generating process as Chetverikov et al. (2016) show that our MD estimator has substantially lower variance and MSE across all sample sizes considered. From a technical point of view, we are able to weaken the growth rate of T relative to N necessary to obtain unbiased asymptotic normality of the estimator of the 'slow' coefficient. We also contribute to this literature by deriving the limiting distribution of the estimator of the 'fast' coefficients, which were not studied by Chetverikov et al. (2016).

The remainder of the paper is structured as follows. Section 2 presents the model and the estimator and briefly discusses equivalent methods to estimate average effects with panel data models to motivate our two-step approach. Section 3 presents the asymptotic theory. Section 4 focuses more in detail on the estimation of quantile panel data models, and we present a generalization of the Hausman test for the random effects assumption. Section 5 discusses the grouped quantile regression model and compares our estimator to the grouped IV quantile regression of Chetverikov et al. (2016). Monte Carlo simulations to analyze the finite sample performance are included in sections 4 and 5. In section 6, as an empirical application, we study the effect of the food stamp program on the distribution of birth weight. Section 7 concludes.

2 Model and Minimum Distance Estimator

2.1 Quantile Model

We want to learn the effects of the time-varying variables x_{1it} and the time-constant variables x_{2i} on the distribution of an outcome y_{it}. We observe these variables for the individuals $i = 1, \ldots, N$ and time periods $t = 1, \ldots, T$.[4] For some quantile index $0 < \tau < 1$, we assume that

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

where $Q(\tau, y_{it}|x_{1it}, x_{2i}, v_i)$ is the τth conditional quantile function of the response variable y_{it} for individual i in period t given the K_1-vector of time-varying regressors $x_{1,it}$, the K_2-vector of time invariant variables $x_{2,it} = x_{2,i}$, and an unobserved random vector v_i of unrestricted unknown dimension. In total, there are $K_1 + K_2 = K$ parameters to estimate. The parameters $\beta(\tau)$, $\gamma(\tau)$ and the unobserved individual heterogeneity $\alpha(\tau, v_i)$ can depend on the quantile index τ. Depending on the setting, $\beta(\tau)$ or $\gamma(\tau)$ (or both) might be the parameters of interest. We normalize $\mathbb{E}[\alpha(\tau, v_i)] = 0$, which is not restrictive because x_{2i} includes a constant.

[4]For notational simplicity, we assume a balanced panel. However, the results generalize to unbalanced panels.

Remark 1 (Conditional versus unconditional effects). In contrast to the average effect, the definition of a quantile treatment effect depends on the conditioning variables. In this paper, we model the distribution of y_{it} conditionally on the covariates and on the individual effect $\alpha(\tau, v_i)$. Thus, even if the individual effects are independent of the regressors, we identify different coefficients than those identified by quantile regression as introduced by Koenker and Bassett (1978) or by instrumental variable quantile regression as introduced by Chernozhukov and Hansen (2005). The following example illustrates the difference between these parameters. Consider an application where each unit i corresponds to a region and each unit t to an individual within this region. We do not have any x_{1it} variable. We are interested in the effect of a binary treatment x_{2i}, which has been randomized and is, therefore, independent from $\alpha(\tau, v_i)$. $\gamma(\tau)$ is the effect of this treatment for individuals that rank at the τ quantile of y_{it} in their region. On the other hand, the quantile regression of y_{it} on x_{2i} identifies the effect for individuals that rank at the τ quantile in the whole country (given the treatment status). These are different parameters except if $\alpha(\tau, v_i) = 0$ for all i or if the treatment effect is homogeneous such that $\gamma(\tau) = \gamma$. Suppose the unconditional treatment effect is of interest. In that case, one can naturally obtain the unconditional distribution functions by integrating out the individual effects (and possibly the other variables) and then inverting the resulting distribution functions to obtain the unconditional quantile functions, see Chernozhukov et al. 2013.[5]

When model (1) holds, the τ quantile regression of y_{it} on x_{1it} and a constant using only observations for individual i identifies the slope $\beta(\tau)$ and the intercept $x_{2i}'\gamma(\tau) + \alpha(\tau, v_i)$. To identify the coefficient on the time-constant variables, we need to consider variation across individuals. Note that

$$\mathbb{E}\left[Q(\tau, y_{it}|x_{1it}, x_{2i}, v_i)|x_{1it}, x_{2i}\right] = x_{1it}'\beta(\tau) + x_{2i}'\gamma(\tau) + \mathbb{E}\left[\alpha(\tau, v_i)|x_{1it}, x_{2i}\right].$$

If $\alpha(\tau, v_i)$ is exogenous with respect to x_{1it} and x_{2i} and the linear model is correctly specified, then $\mathbb{E}[\alpha(\tau, v_i)|x_{1it}, x_{2i}] = 0$ and this linear regression identifies the parameters of interest.[6] It suggests a two-step estimation strategy: (i) individual-level quantile regression of y_{it} on x_{1it}, (ii) OLS regression of the fitted values from the first stage on x_{1it} and x_{2i}.

When the individual effects $\alpha(\tau, v_i)$ are endogenous (possibly correlated with x_{1it} and x_{2i}), we assume that there is a L-dimensional vector ($L \geq K$) of valid instruments z_{it} satisfying

$$\mathbb{E}[z_{it}\alpha(\tau, v_i)] = \mathbb{E}\left[z_{it}\left(Q(\tau, y_{it}|x_{1it}, x_{2i}, v_i) - x_{1it}'\beta(\tau) - x_{2i}'\gamma(\tau)\right)\right] = 0. \tag{2}$$

Note that $\beta(\tau)$ is identified in model (1) as long as there is some variation in x_{1it} over time for some individuals i. For instance, we can include the time demeaned regressors, $\dot{x}_{1it} = x_{1it} - \bar{x}_{1i}$ with $\bar{x}_{1i} = T^{-1}\sum_{t=1}^{T} x_{1it}$, in the vector of instruments z_{it} because this variable will satisfy condition (2) under strict exogeneity. On the other hand, we need additional instruments to

[5] We refer to Frölich and Melly (2013) for a discussion about conditional and unconditional treatment effects.
[6] Uncorrelation between $\alpha(\tau, v_i)$ and x_{1it} and x_{2i} is sufficient to identify the linear projection.

identify $\gamma(\tau)$. Equation (2) suggests a similar estimation strategy as in the exogenous case but with the instrumental variable estimator (or more generally the GMM estimator) in the second stage: (i) individual-level quantile regression of y_{it} on x_{1it}, (ii) GMM regression of the fitted values from the first stage on x_{1it} and x_{2i} using z_{it} as instrumental variable.

Remark 2 (Skorohod representation). The following Skorohod representation implies the model defined in equation 1.

$$\begin{aligned} y_{it} &= x_{1it}\beta(u_{it}) + x_{2i}\gamma(u_{it}) + \alpha(u_{it}, v_i) \\ &= q(x_{1it}, x_{2i}, u_{it}, v_i) \end{aligned}$$

where $q(x_{1it}, x_{2i}, u, v)$ is strictly increasing in u (while fixing the other arguments). We normalize $u_{it}|x_{1it}, x_{2i}, v_i \sim U(0,1)$ such that $q(x_{1it}, x_{2i}, u, v_i)$ is the conditional quantile function. v_i ranks the individuals while u_{it} ranks observations over time for the same individual. In this model, a sufficient condition for the instrument validity assumption (2) is $(u_{it}, v_i) \perp\!\!\!\perp z_{it}$. If the instrument does not vary over time, only $v_i \perp\!\!\!\perp z_i$ is sufficient.

Remark 3 (Heterogeneous coefficients). Our model allows only the intercept to differ between individuals.[7] We now consider a more general model where we also allow the slopes to differ between individuals:

$$y_{it} = x'_{1it}\beta(u_{it}, v_i) + x'_{2i}\gamma(u_{it}, v_i) + \alpha(u_{it}, v_i). \tag{3}$$

If we maintain the conditional strict monotonicity assumption with respect to u_{it}, this model implies that

$$Q(\tau, y_{it}|x_{1it}, x_{2i}, v_i) = x'_{1it}\beta(\tau, v_i) + x'_{2i}\gamma(\tau, v_i) + \alpha(\tau, v_i). \tag{4}$$

In the exogenous case where $(x_{1it}, x_{2i}) \perp\!\!\!\perp v_i$, this implies

$$\begin{aligned} \mathbb{E}\left[Q\left(\tau, y_{it}|x_{1it}, x_{2i}, v_i\right)|x_{1it}, x_{2i}\right] &= x'_{1it}\int \beta(\tau, v)dF_V(v) + x'_{2i}\int \gamma(\tau, v)dF_V(v) + \int \alpha(\tau, v)dF_V(v) \\ &= x'_{1it}\bar{\beta}(\tau) + x'_{2i}\bar{\gamma}(\tau) \end{aligned}$$

because we have normalized $\mathbb{E}[\alpha(\tau, v_i)] = 0$. It follows that the linear projection of $Q(\tau, y_{it}|x_{1it}, x_{2i}, v_i)$ on x_{1it} and x_{2i} identifies the average effects when these effects are heterogeneous. Thus, the linear projection identifies the coefficients $\beta(\tau)$ and $\gamma(\tau)$ when the homogenous model (1) holds and the average effect for all individual at the τ quantile of their conditional distribution when the heterogenous model (4) holds.[8] Naturally, it is also possible to modelize the heterogeneity between individuals by estimating more flexible linear projections of $Q(\tau, y_{it}|x_{1it}, x_{2i}, v_i)$. For instance, we can interact x_{1it} with indicator variables for individuals, which allows for unrestricted heterogeneity of $\beta(\tau, v_i)$, or by interacting x_{2i} with other observable characteristics.

[7] This is the same model as in Chetverikov et al. (2016), where a similar Skorohod representation is derived in their footnote 6.

[8] In the endogenous case, we obtain the instrumental variable projection instead of the standard linear projection. For instance, if x_{2i} is an endogenous binary variables and z_{it} is a binary instrument, we identify the average treatment effects for the complying individuals at the τ quantile of their conditional distribution.

Remark 4 (**Least-squares versus quantile regression in the second-stage).** As discussed in remark 3, the projection identifies the average coefficients across individuals when those are heterogeneous. It is possible to analyze the inter-individual heterogeneity if we impose (3), restrict v_i to be a scalar, and impose the strict monotonicity of $x'_{1it}\beta(u,v) + x'_{2i}\gamma(u,v) + \alpha(u,v)$ with respect to v.[9] When we normalize $v_i|x_{1it}, x_{2i} \sim U(0,1)$ and $\alpha(\tau,\theta) = 0$, we obtain in the exogeneous case, for any $0 < \theta < 1$,

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

All coefficients have two quantile indices: one for the heterogeneity across individuals and one for the heterogeneity within individuals.[10] These heterogeneous coefficients are identified by a two-step quantile regression: (i) τ quantile regression of y_{it} on x_{1it}, (ii) θ quantile regression of the fitted values from the first-stage on x_{1it} and x_{2i}. This alternative strategy identifies different parameters and is outside the scope of this paper. We focus instead on the model defined by equations (1) and (2), which is the same as in Chetverikov et al. (2016) and nests the fixed effects quantile regression model (studied e.g. in Galvao et al., 2020).

2.2 Quantile Minimum Distance Estimators

Motivated by the representation in equation (2), we suggest a quantile version of the two-steps procedure. In the first step, for each individual i and quantile τ, we regress y_{it} on time-varying variables x_{1it} and a constant using quantile regression. The intercept of the first stage regression captures both the individual effect $\alpha(\tau, v_i)$ and the term $x'_{2i}\gamma(\tau)$ as these vary only between individuals. In a second step, we regress the fitted values of the first stage on x_{1it} and x_{2i}, using GMM with instruments z_{it}.

Formally, the first stage quantile regression solves the following minimization problem for each individual and quantile separately:

$$\hat{\beta}_i(\tau) \equiv \left(\hat{\beta}_{0,i}, \hat{\beta}'_{1,i}\right)' = \underset{(b_0,b_1)\in\mathbb{R}^{K_1+1}}{\arg\min} \frac{1}{T}\sum_{t=1}^{T} \rho_\tau(y_{it} - b_0 - x'_{1it}b_1), \tag{5}$$

where $\rho_\tau(x) = (\tau - 1\{x < 0\})x$ for $x \in \mathbb{R}$ is the check function. The true vector of coefficients for individual i is given by $\beta_i(\tau) = (\alpha(\tau, v_i) + x'_{2i}\gamma(\tau), \beta(\tau)')'$.

Notation. Throughout the paper, we will use the following notation. Let $\tilde{x}_{it} = (1, x'_{1it})'$ and $x_{it} = (x'_{1it}, x'_{2i})'$. For each individual i we define the following matrices. The $T \times K_1$ matrix of time-varying regressors $X_{1i} = (x_{1i1}, x_{1i2}, \ldots, x_{1iT})'$, the $T \times K$ matrix containing all regressors $X_i = (x_{i1}, x_{i2} \ldots, x_{1iT})'$ and the $T \times L$ matrix of instruments $Z_i = (z_{i1}, z_{i2}, \ldots, z_{iT})'$. Further, we define two matrices for all observations. The $NT \times K$ matrix of regressors for all

[9] In the presence of multivariate heterogeneity, quantile regression identifies local average structural derivatives of nonseparable models, see Hoderlein and Mammen (2007).

[10] This is similar to the instrumental variable model in Chesher (2003) and Ma and Koenker (2006), which also contains two quantile indices: one for the selection equation and one for the outcome equation.

individuals $X = (X_1', \ldots, X_N')'$ and the $NT \times L$ matrix of instruments for all individuals as $Z = (Z_1', \ldots, Z_N')'$. We let Y be the $NT \times 1$ vector of the response variable. The fitted value for individual i in period t is $\hat{y}_{it}(\tau) = \hat{\beta}_{0,i}(\tau) + x_{1it}'\hat{\beta}_{1,i}(\tau)$. Denote the $T \times 1$ column vector of fitted values for individual i by $\hat{Y}_i(\tau) = (\hat{y}_{i1}(\tau), \ldots, y_{i,T}(\tau))'$, and the $NT \times 1$ vector of fitted values by $\hat{Y}(\tau) = (\hat{y}_1'(\tau), \ldots, \hat{y}_N'(\tau))'$.

Remark 5 (Alternative first-stage estimators). The Koenker and Bassett (1978) quantile regression estimator is not necessarily efficient. Newey and Powell (1990) suggest a semiparametrically efficient weighted estimator of $\beta_i(\tau)$. We prefer to use the unweighted quantile regression estimator due to the difficulty of estimating the weights and the complex interpretation of the estimates in case of misspecification. In our model (1), the variation within individuals over time is assumed to be exogenous. If this was not the case, it would be possible to use an instrumental variable quantile regression (see e.g., Chernozhukov and Hansen (2006)) in the first stage, followed by the second stage GMM regression described below.[11] We do not explore this (computationally expensive) extension in this paper.

The second stage consists in a GMM regression using $\mathbb{E}[g_i(\delta, \tau)] = 0$ a a moment condition, where $g_i(\delta, \tau) = Z_i'(\tilde{X}_i\hat{\beta}(\tau) - X_i\delta(\tau))$ and $\delta = (\beta', \gamma')'$ is the K-dimensional vector of coefficients. Thus, the moment restriction depends on the first stage and is the sample counterpart of $\mathbb{E}[Z_i'\alpha_i(\tau, v_i)]$. The closed-form expression of the second stage estimator is

$$\hat{\delta}(\hat{W}, \tau) = \left(X'Z\hat{W}(\tau)Z'X\right)^{-1} X'Z\hat{W}(\tau)Z'\hat{Y}(\tau), \tag{6}$$

where $\hat{W}(\tau)$ is a $L \times L$ symmetric weighting matrix. If $L = K$, the second step estimator in (6) simplifies to the IV estimator using Z as instrument.

Our two-step estimator is extremely simple to implement; it requires only routines performing quantile regression and GMM estimation, which are already available in many softwares. In addition, we provide general-purpose packages for both R and Stata. Quantile regression, which is computationally more demanding due to the absence of a closed-form solution, is used only in the first stage, where there are fewer observations and a limited number of parameters to estimate. The first stage is also embarrassingly parallelizable, which further increases the computational speed. For this reason, our estimator remains computationally attractive in large datasets with numerous individuals. The second stage is a straightforward GMM estimator, which includes OLS and two-stage least squares as special cases. Traditional panel data methods can also be used in the second stage. For instance, in our application, we observe individuals in a given year in a given county. The subscript i defines a county-year cell, while the subscript t defines an individual within this cell. In the second stage, we include year and county fixed effects to estimate the effect of food stamps on the birth weight distribution.

[11] An IV extension of the MD estimator of Galvao and Wang (2015) is suggested in Dai and Jin (2021).

Remark 6 (Interpretation as a minimum distance estimator). Our estimator can be interpreted as a MD estimator, where the second stage imposes restrictions on the first stage coefficients. For simplicity, we consider in this remark the case where all the regressors are exogenous and $Z = X$. The MD estimator minimizes

$$\hat{\delta}(\tau) = \arg\min_{\beta} \sum_{i=1}^{N} (\hat{\beta}_i(\tau) - R_i\delta(\tau))'\tilde{W}(\tau)(\hat{\beta}_i(\tau) - R_i\delta(\tau)), \tag{7}$$

where $\tilde{W}(\tau)$ is a $K \times K$ weighting matrix that might depend on the quantile index. If we set $\tilde{W} = \tilde{X}_i'\tilde{X}_i$ then the MD estimator is algebraically identical to using OLS in the second stage. The matrix of restrictions R_i is defined such that $\tilde{X}_i R_i = X_i$:

$$\underset{(K_1+1)\times K}{R_i} = \begin{pmatrix} x_{2i}' & 0 \\ 0 & I_{K_1} \end{pmatrix}.$$

In this sense, our estimator is a MD estimator. However, it does not correspond to the textbook definition of a "classical minimum distance" estimator.[12] In the classical MD setup, all the sampling variance arises in the first stage: if we know the first stage coefficients, we know the final coefficients. It follows that the efficient weighting matrix is the inverse of the first stage variance. In our case, the second stage also contributes to the variance due to the presence of the individual effects $\alpha(\tau, v_i)$: even if we know $\beta_i(\tau)$ (for a finite number of individuals), we cannot exactly pinpoint $\gamma(\tau)$. The individual effects play a role similar to misspecification in the classical MD, but the resulting 'bias' disappears asymptotically as the number of individuals increases. This is the second important difference: the dimension of our first stage estimates increases with the sample size while it is fixed for classical MD estimators.

The estimators suggested by Chamberlain (1994) and Galvao and Wang (2015) are special cases of our estimator in which only the first stage estimation error matters, and the efficient weighting matrix is the inverse of the first stage variance.[13] In the next section, we derive the asymptotic distribution when the estimation error of both stages matters.

2.3 Least Squares Minimum Distance Estimators

This section discusses the analogy between our minimum distance estimator and traditional least squares (panel data) estimators in the special case least squares regression is used in the first stage. We present here some equivalent ways to compute various common estimators and show that they can be estimated using our two-stages minimum distance approach. A more detailed discussion including formal statements can be found in Appendix A with proofs in Appendix B.1.

[12] See section 14.6 in Wooldridge (2010).

[13] The efficient MD estimator of Galvao and Wang (2015) is numerically identical to our estimator using an IV second stage with instrument $Z_i^*(\tau) = \tilde{X}_i(\tilde{X}_i'\tilde{X}_i V_i(\tau)\tilde{X}_i'\tilde{X}_i)^{-1}\tilde{X}_i' Z_i = (\tilde{X}_i V_i(\tau)\tilde{X}_i')^{+} Z_i$, where Z_i contains a constant, x_{1it}, and individual dummies. The symbol $^+$ denotes the Moore-Penrose inverse.

Consider first a traditional panel data model with individual effects and without time-invariant regressors

$$y_{it} = x'_{1it}\beta + \alpha_i + \varepsilon_{it}. \tag{8}$$

The least squares fixed effects estimator can be computed by subtracting from each variable its time average and applying the ordinary least squares estimator. This within transformation eliminates the potential endogeneity coming from α_i and provides a consistent estimator without imposing assumptions on the unobserved time-invariant heterogeneity. This approach is not applicable in quantile models, as there is no known transformation that eliminates the individual effects. In particular, time-demeaning or first-differencing the variables modifies the interpretation of the quantile regression coefficients because the quantiles are nonlinear operators. A second possibility to estimate fixed effects models consists in estimating the individual effects by including an indicator variable for each individual. It is well-known that this is algebraically identical to the within estimator. In quantile models, the dummy variables regression is computationally unattractive, as it requires estimating many parameters.[14] In addition, this approach does not provide a way to estimate the effect of the time-constant variables, especially when we need to exploit an instrumental variable to identify their effect.[15] A third numerically equivalent way to compute the least squares fixed effects estimator consists in dividing the problem in two steps. The first stage consists of individual-level regressions for each i. The intercept of each regression will absorb the unobserved heterogeneity α_i. The second stage aggregates the individual results by regressing the fitted values from the first stage on x_{1it} using the time-demeaned regressors, $\dot{x}_{1it} = x_{1it} - \bar{x}_{1i}$ where $\bar{x}_{1i} = T^{-1}\sum_{t=1}^{T} x_{1it}$, as an instrumental variable. As shown below, using the first-stage fitted values as dependent variable does not affect the results in least squares model. Whereas, the instrument exploits only the variation within individuals, thus yielding a within estimator. This procedure can be easily extended to quantile models, where it substantially reduces the computational burden of quantile fixed effects estimation.[16]

The two-step procedure is not specific to the fixed effects case but applies to a wide range of estimators. We include the time-constant regressors x_{2i} in the model

$$y_{it} = x'_{1it}\beta + x'_{2i}\gamma + \alpha_i + \varepsilon_{it} \tag{9}$$

and consider our minimum distance estimator with least squares in the first stage instead of

[14] This approach is nevertheless feasible thanks to the sparsity of the design matrix, see Koenker (2004), and Koenker and Ng (2005).

[15] Koenker (2004) suggests a penalized quantile regression estimator with individual effects, which can be interpreted as a random effects estimators. However, the linear dependence between the individual indicator variables and the time-constant variables implies that the effect of these variables is identified only from the individuals with fully shrunken individual effects, see Harding et al. (2020).

[16] There is a fourth possibility to compute the least squares fixed estimator, which is similar to the third one. The first stage also consists of individual-level regressions for each i. The second stage aggregates the slope coefficients directly by taking the average of the individual slopes with weights proportional to the variance of the regressors within individuals. Galvao and Wang (2015) suggests a similar estimator for the fixed effects quantile regression model. However, this averaging method does not allow for the presence of time-constant regressors and, more generally, does not exploit the between-individuals variation.

quantile regression. That is, the first stage consists of individual-level least squares regressions, including only the time-varying variables. The second stage is a linear GMM regression of the first-stage fitted values on both time-varying and time-invariant variables. This two-step estimator is algebraically identical to the one-step linear GMM regression of y_{it} on x_{it} under the mild condition that for each i the matrix of instrument lies in the column space of the matrix of first stage regressors (see Proposition 3 in Appendix A). To understand the intuition behind this results, note that the fitted values of the first-stage least squares regression can be written as $P_{X_i}Y_i$ where P_{X_i} is the least squares projection matrix of individual i. If the instrument matrix, Z_i, is in the column space of $\tilde{X}_i$, it follows $P_{X_i}Z_i = Z_i$. Therefore, $Z'Y = Z'\hat{Y}$ and the two GMM regressions are numerically identical. The matrix Z_i will lies in the column space of $\tilde{X}_i$ if the instruments are part of $\tilde{X}_i$, time-invariant, or some linear combination of the columns of the matrix of first-stage regressors. For example, demeaned time-varying regressors fall into the latter category. Since OLS and 2SLS are special cases of GMM, the same result follows.

We can numerically obtain the most common least squares panel data estimators by selecting different instrumental variables for the second step GMM regression. In essence, the instrument determines which variation we exploit in the second stage.[17] For instance, we obtain the between estimator by using the individual time-averaged variables, $\bar{x}_{1i}$ and x_{2i} as instruments. Instrumental variable approaches are available also for random effects estimation. More precisely, while FGLS is the most common estimator for the random effects model, Im et al. (1999) show that the overidentified 3SLS estimator, with instruments $\dot{x}_{1it}$, $\bar{x}_{1i}$, and x_{2i}, is numerically identical to the random effects estimator. Since 3SLS is a special case of a GMM estimator, using the first-stage fitted values as dependent variables does not change the estimates. Alternatively, the random effects estimator can be implemented using the theory of optimal instrument with a just identified 2SLS regression (see Im et al., 1999; Hansen, 2022). Finally, the Hausman-Taylor estimator (Hausman and Taylor, 1981) can be implemented by selecting the following instruments: $\dot{x}_{1it}$, the individual time average of the exogenous regressors, and potential external instruments. Interestingly, in all cases, clustering the standard errors at the level of the individuals (or at a higher level) is sufficient to capture the first stage estimation error, see Proposition 6 in Appendix B.1. These clustered standard errors are numerically identical to the standard errors obtained after using the one-step GMM estimator with y_{it} as the dependent variable.

3 Asymptotic Theory

In this section, we state the assumptions and present the asymptotic results. All the proofs are included in Appendix B.3. For simplicity of notation, in the following we write $\alpha_i(\tau)$ instead of $\alpha(\tau, v_i)$. We prove weak uniform consistency and weak convergence of the whole quantile regression process for $\tau \in \mathcal{T}$, where $\mathcal{T} \in (0, 1)$ is a compact set of quantile indices of interest.

[17]It is worth noting that the IV approach to these panel data estimators can be implemented also in one stage with y_{it} as the dependent variable.

The symbol $\ell^\infty(\mathcal{T})$ denotes the set of component-wise bounded vector values function of $\mathcal{T}$ and $\rightsquigarrow$ denotes weak convergence.

We start by writing the sampling error of $\hat{\delta}(\hat{W}, \tau)$ as a sum of a component arising from the first stage estimation error of $\beta_i(\tau)$ and a component arising from the second stage noise $\alpha_i(\tau)$:

Lemma 1 (**Sampling error**). *Assume that the model in equation (1) holds, then*

$$\hat{\delta}(\hat{W}, \tau) - \delta(\tau) = \left(S'_{ZX}\hat{W}(\tau)S_{ZX}\right)^{-1} S'_{ZX}\hat{W}(\tau)\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} z_{it}\left(\tilde{x}'_{it}(\hat{\beta}_i(\tau) - \beta_i(\tau)) + \alpha_i(\tau)\right),$$

where $S_{ZX} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} z_{it}x'_{it}$.

We now state assumptions that ensure that both components are well-behaved. For the analysis of the first stage estimator, we rely on results derived in Galvao et al. (2020) and make the assumptions required in their Theorem 2:

Assumption 1 (**Sampling**). *(i) The processes* $\{(y_{it}, x_{it}, z_{it}) : t \in \mathbb{Z}\}$ *are independent across* i. *(ii) For each* i*, the observations* $(y_{it}, x_{1it}, z_{1it})_{t=1,...,T}$ *are i.i.d. across* t.

Assumption 2 (**Covariates**). *(i) For all* $i = 1, ..., N$ *and all* $t = 1, ..., T$, $\|x_{it}\| \leq C$ *almost surely. (ii) The eigenvalues of* $\mathbb{E}[x_{1it}x'_{1it}]$ *are bounded away from zero and infinity uniformly across* i.

Assumption 3 (**Conditional distribution**). *The conditional distribution* $F_{y_{it}|x_{1it}}(y|x)$ *is twice differentiable w.r.t.* y*, with the corresponding derivatives* $f_{y_{it}|x_{1it}}(y|x)$ *and* $f'_{y_{it}|x_{1it}}(y|x)$*. Further, assume that*

$$f_{max} := \sup_{i} \sup_{y\in\mathbb{R}, x\in\mathcal{X}} |f_{y_{it}|x_{1it}}(y|x)| < \infty$$

and

$$\bar{f}' := \sup_{i} \sup_{y\in\mathbb{R}, x\in\mathcal{X}} |f'_{y_{it}|x_{1it}}(y|x)| < \infty.$$

where $\mathcal{X}$ *is the support of* x_{1it}

Assumption 4 (**Bounded density**). *There exists a constant* $f_{min} < f_{max}$ *such that*

$$0 < f_{min} \leq \inf_{i} \inf_{\tau\in\mathcal{T}} \inf_{x\in\mathcal{X}} f_{y_{it}|x_{1it}}(Q(\tau, y_{it}|x)|x)$$

These are quite standard assumptions in the quantile regression literature. In Assumption 1, we assume that the processes are independent across i; this assumption can also be relaxed by allowing for clustering between individuals. We also assume that the observations are i.i.d. over time, but this can be relaxed at the cost of a more complex notation by applying Theorem 4 in Galvao et al. (2020), which requires only stationarity and β-mixing. The estimator of the asymptotic variance that we suggest below is consistent in both cases. Assumption 2 requires that the regressors are bounded and that $\mathbb{E}[x_{1it}x_{1it}']$ is invertible. Assumptions 3 and 4 impose smoothness and boundedness of the conditional distribution, the density and its derivatives.

For the second stage GMM regression we impose the following assumptions:

Assumption 5 (Instruments). *(i) For all $i = 1, \ldots, N$ and all $t = 1, \ldots, T$, $||z_{it}|| \leq C$ a.s. (ii) For all $i = 1, \ldots, N$ and all $t = 1, \ldots, T$, $\mathbb{E}[z_{it}\alpha_i(\tau)] = 0$. (iii) For all $i = 1, \ldots, N$ and all $t = 1, \ldots, T$, y_{it} is independent of z_{it} conditional on (x_{it}, v_i). (iv) As $N \to \infty$, $N^{-1}\sum_{i=1}^{N}\mathbb{E}[z_{it}x_{it}'] \to \Sigma_{ZX}$ where the singular values of Σ_{ZX} are bounded from below and from above.*

Assumption 6 (Individuals effects).
(i) For all $i = 1, \ldots, N$, $\mathbb{E}\left[\sup_{\tau\in\mathcal{T}} |\alpha_i(\tau)|^{4+\varepsilon_C}\right] \leq C$ for $\varepsilon_C > 0$. (ii) For some (matrix-valued) function $\Omega_2 : \mathcal{T} \times \mathcal{T} \to \mathbb{R}^{L\times L}$, $N^{-1}\sum_{i=1}^{N}\mathbb{E}[\alpha_i(\tau_1)\alpha_i(\tau_2)z_{it}z_{it}'] \to \Omega_2(\tau_1, \tau_2)$ uniformly over $\tau_1, \tau_2 \in \mathcal{T}$. (iii) For all $\tau_1, \tau_2 \in \mathcal{T}$, $|\alpha_i(\tau_2) - \alpha_i(\tau_1)| \leq C|\tau_2 - \tau_1|$.

Assumption 7 (Coefficients). *For all $\tau_1, \tau_2 \in \mathcal{T}$ and $i = 1, \ldots, N$, $||\beta_i(\tau_2) - \beta_i(\tau_1)|| \leq C|\tau_2 - \tau_1|$.*

These assumptions are the same as in Chetverikov et al. (2016). For the instrumental variables, we assume that (i) they are bounded, (ii) they are not correlated with the individual effect (exclusion restriction), (iii) they do not affect the first stage estimation (this is often satisfied by construction, e.g. when the instruments do not vary within individuals or are a linear transformation of the first stage regressors), and (iv) they satisfy the relevance conditions. For the individual effects we assume that they have a finite fourth moment, and the average variance of $z_{it}\alpha_i(\tau)$ converges to a well-defined matrix. Finally, we assume that the individual effects and the coefficients are continuous functions of the quantile index.

Quantile regression is a nonlinear estimator that is potentially biased in finite samples. Hence, for consistency, T must increase to infinity as N increases to infinity. For unbiased asymptotic normality, we need the bias to shrink faster than the standard deviation of the estimator. The bias of the first stage quantile regression estimator is of order $\frac{1}{T}$. We will show that some elements of $\hat{\delta}(W, \tau)$ converge at the $\sqrt{N}$ rate such that we need that T goes to infinity more quickly than $\sqrt{N}$. On the other hand, other elements converge at the $\sqrt{NT}$ rate such that we need that T goes to infinity more quickly than N. We state these three different relative growth rates in the following assumption:

Assumption 8 (Growth rates).

(a) $\frac{\log N}{T} \to 0$,

(b) $\frac{\sqrt{N}\log T}{T} \to 0$,

(c) $\frac{N(\log T)^2}{T} \to 0$.

In our first result, we establish that our estimator is uniformly consistent. In addition to the previous stated conditions, we also assume that the estimated weighting matrix uniformly converges to a continuous function. Note that, for consistency, the required growth condition for T is extremely weak.

Theorem 1 (**Uniform consistency**). *Let the model in equation (1), Assumptions 1-7 as well as Assumption 8(a) hold. Uniformly in $\tau \in \mathcal{T}$, $\hat{W}(\tau) \underset{p}{\rightarrow} W(\tau)$ where $W(\tau)$ is strictly positive definite and, for all $\tau_1, \tau_2 \in \mathcal{T}$, $||W(\tau_2) - W(\tau_1)|| \leq C|\tau_2 - \tau_1|$. Then,*

$$\sup_{\tau \in \mathcal{T}} \|\hat{\delta}(\tau) - \delta(\tau)\| = o_p(1).$$

We now study the asymptotic distribution of our estimator. In Lemma 1 we see that the sample moment condition is made of two terms. It is useful to consider them separately; we define

$$\bar{g}^{(1)}_{NT}(\hat{\delta}, \tau) := \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} z_{it} \tilde{x}'_{it} \left(\hat{\beta}_i(\tau) - \beta_i(\tau) \right) \tag{10}$$

$$\bar{g}^{(2)}_{NT}(\hat{\delta}, \tau) := \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} z_{it} \alpha_i(\tau) \tag{11}$$

such that total moment condition is the sum of both components: $\bar{g}_{NT}(\hat{\delta}, \tau) := \bar{g}^{(1)}_{NT}(\hat{\delta}, \tau) + \bar{g}^{(2)}_{NT}(\hat{\delta}, \tau) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} g_{it}(\hat{\delta}, \tau)$. Lemma 2 establishes joint asymptotic normality for the entire moment condition processes.

Lemma 2 (**Asymptotic distribution of the sample moments**). *We assume that Assumptions 1-7 hold.*

(i) Under Assumption 8(c), as $N \to \infty$,

$$\sqrt{NT} \bar{g}^{(1)}_{NT}(\hat{\delta}, \cdot) \rightsquigarrow Z_1(\cdot), \text{ in } l^{\infty}(\mathcal{T}), \tag{12}$$

where $Z_1(\cdot)$ is a mean-zero Gaussian process with uniformly continuous sample paths and covariance function $\Omega^{(1)}(\tau, \tau') = \mathbb{E}\left[\Sigma_{ZXi} V_i(\tau, \tau') \Sigma'_{ZXi}\right]$ with $\Sigma_{ZXi} = \mathbb{E}_t[z_{it} x'_{it}]$ and $V_i(\tau, \tau')$ is the asymptotic variance-covariance matrix of $\hat{\beta}_i(\tau)$ and $\hat{\beta}_i(\tau')$:

$$V_i(\tau, \tau') = \mathbb{E}_t[f_{y|x}(Q_{y|x}(\tau|\tilde{x}_{it})|\tilde{x}_{it}) \tilde{x}_{it} \tilde{x}'_{it}]^{-1} (\min(\tau, \tau') - \tau\tau') \mathbb{E}_t[\tilde{x}_{it} \tilde{x}'_{it}] \mathbb{E}_t[f_{y|x}(Q_{y|x}(\tau'|\tilde{x}_{it}) \tilde{x}_{it} \tilde{x}'_{it}]^{-1}$$

(ii) As $N \to \infty$,

$$\sqrt{N} \bar{g}^{(2)}_{NT}(\hat{\delta}, \cdot) \rightsquigarrow Z_2(\cdot), \text{ in } l^{\infty}(\mathcal{T}), \tag{13}$$

where $Z_2(\cdot)$ is a mean-zero Gaussian process with uniformly continuous sample paths and covariance function $\Omega_2(\tau, \tau')$, which is defined in Assumption 6(ii).

(iii) Under Assumption 8(c), as $N \to \infty$, $\sup_{\tau \in \mathcal{T}} \mathrm{Cov}(Z_1(\cdot), Z_2(\cdot)) = o_p\left(\frac{1}{NT}\right)$.

$\bar{g}^{(1)}_{NT}(\hat{\delta},\cdot)$ reflects the estimation error that arises in the first stage quantile regression estimation. Since the first-stage regressors vary over time, the relevant number of observations is NT and, correspondingly, the variance is proportional to $1/NT$. On the other hand, since the bias of the first-stage quantile regression is of order $1/T$, for asymptotic unbiasedness, we must require that T goes to infinity slightly faster than N. In the proof we build on results derived in Volgushev et al. (2019) and in Galvao et al. (2020). $\bar{g}^{(2)}_{NT}(\hat{\delta},\cdot)$ reflects the estimation error due to the randomness in $\alpha_i(\tau)$. Since $\alpha_i(\tau)$ varies only between individuals, the relevant number of observations here is N and, accordingly, the variance of this moment converges at the slower rate of $1/N$. This moment can also be interpreted as the moment that would be relevant if we knew $\beta_i(\tau)$. It follows that no restriction on the number of time periods is necessary.

The asymptotic distribution of the estimator corresponds to (a linear function of) the sum of both processes $Z_1(\cdot)$ and $Z_2(\cdot)$. Since the rate of convergence of these processes are different, the rate of convergence and the first-order asymptotic approximation will be determined by the slower moment, which is $\bar{g}^{(2)}_{NT}(\hat{\delta},\cdot)$. There are, however, two interesting special cases where this moment is exactly equal to zero. First, when $\alpha_i(\tau) = 0$ for all i, then all elements of the vector $\bar{g}^{(2)}_{NT}(\hat{\delta},\cdot)$ are equal to zero and the first-order asymptotic distribution of the whole vector $\hat{\delta}(W,\cdot)$ is determined by $\bar{g}^{(1)}_{NT}(\hat{\delta},\cdot)$. This corresponds to the case where there are no individual effects. Second, this moment will also be equal to zero for the instruments such that $\bar{z}_i := T^{-1}\sum_{t=1}^{T} z_{it} = 0$ for all i. Note that all instruments that vary only within i can be normalized to have mean zero.[18] Thus, even if $\mathrm{Var}(\alpha_i(\tau)) \neq 0$, the elements of $\bar{g}_{NT}(\hat{\delta},\cdot)$ associated with the instruments that varies only within-individuals, for instance $x_{it}-\bar{x}_i$ in a fixed effects model, converge at the $1/\sqrt{NT}$ rate while the elements associated with the instruments that vary between individuals, for instance, $\bar{x}_i$ in a between model, converge at the $1/\sqrt{N}$ rate. In what follows, we distinguish between these two sorts of instruments: L_1 instruments in z_{1it} vary only within individuals, while L_2 instruments in z_{2it} also vary between individuals. If an instrument varies both between and within individuals, it must be classified in z_{2it}. We order the instruments such that $z_{it} = (z'_{1it}, z'_{2it})'$. Note that if $\mathrm{Var}(\alpha_i(\tau)) = 0$, then $L_1 = L$ and $L_2 = 0$, which implies $K_1 = K$ and $K_2 = 0$.

The rate of convergence of each element of $\hat{\delta}(W,\cdot)$ is determined by the rate of convergence of the moments that are used asymptotically to estimate this parameter. Coefficients on variables that vary only between individuals cannot be estimated using the 'fast' instruments z_{1it}, which do not vary between individuals, and will necessarily converge at the slow $1/\sqrt{N}$ rate. On the other hand, coefficient on time-varying variables can possibly be estimated using the time-varying instruments z_{1it}. We order the vector of coefficients $\delta(\tau)$ (and the corresponding regressors x_{it}) such that the first K_1 elements are identified using only z_{1it}, while the remaining K_2 elements are not identified without z_{2i}. Formally, we assume that Σ_{11}, which is the $L_1 \times K_1$ upper-left submatrix of Σ_{ZX} has full column rank. However, even if $\delta_1(\tau)$ is identified using only the fast

[18] Except the constant term.

moments, all the elements of $\hat{\delta}(W,\tau)$ will converge at the slow rate if we do not restrict the weighting matrix $\hat{W}(\tau)$ because the slow moments may contaminate $\hat{\delta}_1(W,\tau)$. A simple but crude way to avoid this consists in not using the slow moments when we estimate $\delta_1(\tau)$.

Theorem 2 (First-order asymptotic distribution). *Let Assumptions 1-7 hold. In addition, $\hat{W}(\tau) \underset{p}{\to} W(\tau)$ uniformly in $\tau \in \mathcal{T}$. We partition the $L \times L$ weighting matrix as follows*

$$W(\tau) = \begin{pmatrix} W_{11}(\tau) & W_{12}(\tau) \\ W_{21}(\tau) & W_{22}(\tau) \end{pmatrix}$$

where $W_{11}(\tau)$ is a $L_1 \times L_1$ matrix and $W_{22}(\tau)$ a $L_2 \times L_2$ matrix. For all $\tau_1, \tau_2 \in \mathcal{T}$, $||W(\tau_2) - W(\tau_1)|| \leq C|\tau_2 - \tau_1|$. Let Σ_{11} and Σ_{22} be the $L_1 \times K_1$ upper-left and $L_2 \times K_2$ bottom-right submatrices of Σ_{ZX}.

(i) Let Assumption 8(c) hold, $W_{11}(\tau)$ is strictly positive definite, $W_{12}(\tau) = 0$, $W_{21}(\tau) = 0$, and $W_{22}(\tau) = 0$. Then,

$$\sqrt{NT}(\hat{\delta}_1(\hat{W}(\cdot),\cdot) - \delta_1(\cdot)) \rightsquigarrow G_1(\cdot)Z_1(\cdot), \text{ in } l^\infty(\mathcal{T}), \tag{14}$$

where $G_1(\tau) = \left(\Sigma_{11}' W_{11}(\tau)\Sigma_{11}\right)^{-1} \Sigma_{11}' W_{11}(\tau)$.

(ii) Let Assumption 8(b) hold and $W(\tau)$ is strictly positive definite, then

$$\sqrt{N}(\hat{\delta}_2(\hat{W}(\cdot),\cdot) - \delta_2(\cdot)) \rightsquigarrow G_2(\cdot)Z_2(\cdot), \text{ in } l^\infty(\mathcal{T}), \tag{15}$$

where $G_2(\tau) = \left(\Sigma_{22}' W_{22}(\tau)\Sigma_{22}\right)^{-1} \Sigma_{22}' W_{22}(\tau)$.

The asymptotic distribution in Theorem 2(ii) is the same as in Chetverikov et al. (2016) but we were able to weaken the growth rate condition from $\frac{N^{2/3}\log T}{T} \to 0$ to $\frac{N^{1/2}\log T}{T} \to 0$ by exploiting new results in Galvao et al. (2020). $\hat{\delta}_1(\tau)$ and $\hat{\delta}_2(\tau)$ have both different rate of convergence, and require different growth conditions. These estimators are first-order efficient if $W_1(\tau) = \Omega_1(\tau)^{-1}$ and $W_2(\tau) = \Omega_2(\tau)^{-1}$. These results have several weaknesses. First, the asymptotic distribution is not uniform in $\text{Var}(\alpha_i(\tau))$ and the rate of convergence changes discontinuously if, for example, $\text{Var}(\alpha_i(\tau))$ converges to zero. Second, for $\hat{\delta}_1(W,\tau)$, we do not use the information contained in the slow moments. Consider, for example, the random effects estimator where all the regressors are time-varying. The vector of instruments consists of $x_{it} - \bar{x}_i$ and $\bar{x}_i$. Considering only the first-order asymptotic distribution, we can obtain a first-order efficient estimator by giving zero weights to the slow moments. In other words, the instruments $\bar{x}_i$ are not used because their contribution is asymptotically negligible, and the random effects estimator would be equivalent to the fixed effects estimator.[19] Third, for the $\hat{\delta}_2(W,\tau)$, variance coming from the first stage does not appear in the asymptotic distribution because it converges

[19] This is not specific to quantile models and also affects least squares models with large T (see Ahn and Moon, 2014).

to zero at a quicker rate. Consequently, inference may have poor properties. To solve these issues, we keep both moments together, implement adaptive inference that takes the first stage error into account, and use the slow moment with weights that decline at the T^{-1} rate.

Note that

$$\sqrt{N}\bar{g}_{NT}(\hat{\delta}, \cdot) \rightsquigarrow \frac{Z_1(\cdot)}{T} + Z_2(\cdot). \tag{16}$$

Following standard GMM arguments, the efficient weighting matrix is given by

$$W(\tau)^* = \left(\Omega_1(\tau)/T + \Omega_2(\tau)\right)^{-1}. \tag{17}$$

With this weight matrix, asymptotically, the fast moments get weighted infinitely more than the slow moments, so the parameters identified by the fast moments will converge at the $\sqrt{NT}$ rate. The parameters that are identified only by the slow moments will not be affected by W_1 such that their asymptotic distribution will depend only on W_2.[20]

We estimate this weighting matrix by

$$\hat{W}(\tau)^* = \hat{\Omega}(\tau)^{-1} = \left(\frac{1}{N}\sum_{i=1}^{N} Z_i'\tilde{u}_i(\tau)\tilde{u}_i(\tau)'Z_i\right)^{-1}.$$

The $T \times 1$ vector $\tilde{u}_i(\tau)$ contains the residuals from a preliminary second stage regression using an inefficient weighting matrix. The proof that $\hat{\Omega} = \frac{\Omega_1}{T} + \Omega_2 + O_p(\zeta_{NT}^2 + \zeta_{NT}T^{-1/2})$ with $\zeta_{NT}(\tau) = \frac{1}{\sqrt{NT}} + \frac{1}{\sqrt{N}} \cdot ||V_\alpha(\tau)||^{1/2}$ follows directly by the proof of Proposition 1 below and is therefore omitted.

For asymptotic normality of the adaptive estimator we need to impose the strongest growth rate condition 8(c):

Theorem 3 (Adaptive estimator). *Let Assumptions 1-7 and 8(c) hold. Then,*

$$\sqrt{N}(\hat{\delta}(\hat{\Omega}(\cdot), \cdot) - \delta(\cdot)) \rightsquigarrow G(\cdot)\left(\frac{Z_1(\cdot)}{T} + Z_2(\cdot)\right), \text{ in } l^\infty(\mathcal{T}), \tag{18}$$

We have seen that the convergence rate and the asymptotic distribution of our estimator changes substantially depending on the data generating process. For this reason, we want to suggest an adaptive inference procedure that is uniformly valid over different degree of heterogeneity and convergence rates of the estimator. Surprisingly, we find that clustering the second-stage covariance matrix at the individual level yields uniformly consistent estimator of the covariance matrix over different degrees of unobserved heterogeneity and convergence rates. Thus, inference does not require estimating the density of the quantile regression in the first stage and is computationally straightforward. Clustering automatically takes the first stage variance into account also for $\sqrt{N}$-consistent parameters, thus providing a higher-order improvement. This simple procedure might work in a broader range of situations, and it is of interest on its own. A similar bootstrap-based procedure is suggested in Fernandez-Val et al. (2022).

[20] As we can see in the proof of Proposition 2, the fact that W_1 is multiplied with T does not cause a problem because the weighting matrix matters only up to scale.

To estimate the asymptotic covariance matrix, define the $T \times 1$ vector of residuals $\hat{u}_i(\tau) = \tilde{X}_i \hat{\beta}_i(\tau) - X_i \hat{\delta}(\tau)$. Then the covariance matrix of $\hat{\delta}(\tau)$ is estimated by

$$\widehat{\text{Var}}(\hat{\delta}) = \left(X'Z\hat{W}Z'X\right)^{-1} X'Z\hat{W} \left(\sum_{i=1}^{N} Z_i'\hat{u}_i(\tau)\hat{u}_i(\tau)'Z_i\right) \hat{W}Z'X \left(X'Z\hat{W}Z'X\right)^{-1}.$$

Proposition 1 **(Consistency of the estimated covariance matrix).** *Let assumptions 1-7 and 8(c) hold. Then,*

$$\text{Var}(\sqrt{N}\hat{\delta}(\tau)) = G\left(\frac{\Omega_1(\tau)}{T} + \Omega_2(\tau)\right)G' + O_p\left(\zeta_{NT}^2 + \zeta_{NT}T^{-1/2}\right)$$

where $\zeta_{NT}(\tau) = \frac{1}{\sqrt{NT}} + \frac{1}{\sqrt{N}} \cdot ||V_\alpha(\tau)||^{1/2}$.

If there are more moment conditions than parameters to estimate ($L > K$), it is possible to test overidentifying restrictions with an overidentification test in the second stage (see e.g. Hansen, 1982). More precisely, we want to test the hypothesis $\mathbb{H}_0 : \mathbb{E}[Z_i'\alpha_i(\tau)] = 0$. Compared to a traditional GMM, our overidentification test has to deal with the possible different convergence rates of the elements of $\hat{\delta}$. We solve this issue by rescaling the efficient weight matrix by Λ_T. Let $g_i(\delta,\tau) = Z_i'\left(\hat{Y}_i(\tau) - X_i\delta(\tau)\right)$ and $\bar{g}_N(\delta,\tau) = \frac{1}{N}\sum_{i=1}^N g_i(\delta,\tau)$. Define the GMM criterion function

$$J\left(\hat{\delta}(\tau)\right) = N\bar{g}_N(\hat{\delta},\tau)'\hat{S}^{-1}(\tau)\bar{g}_N(\hat{\delta},\tau), \tag{19}$$

where $\hat{S} = \Lambda_T^{-1}\hat{\Omega}\Lambda_T^{-1}$ is the inverse of the second order optimal weighting matrix. Note that this weight matrix is uniformly efficient over different degrees of unobserved heterogeneity and different convergence rates of the moment conditions.

Proposition 2. *Under the $\mathbb{H}_0$ and Assumptions 1-6 and 8(c) as T and $N \to \infty$, $J(\hat{\delta}(\tau)) \xrightarrow{d} \chi^2_{L-K}$.*

Hence, the criterion function $J\left(\hat{\delta}(\tau)\right)$ can be used to assess the validity of the instruments. In the next section, we show how this overidentification test can be used as a generalization of the Hausman Test for the random effects estimator.

4 Traditional Quantile Panel Data Estimators

4.1 Fixed Effects, Random Effects and Between Estimators

The MD estimator can be used for many panel data models, including the fixed effects, the random effects, the between, and the Hausman-Taylor model. The first stage estimation uses only data for one individual at a time and is unaffected. In the second stage, as for least squares estimation (see section 2.3), we compute panel data estimators by selecting different instruments. Depending on the model, the instrument z_{it} will be defined so that the orthogonality condition

holds. More precisely, for fixed effects estimation, the instrument z_{it} contains the demeaned regressor $\dot{x}_{1it}$ and varies only within i. For the between estimator, z_{it} equals the individual mean of the regressors $\bar{x}_{it}$. Finally, for the pooled estimator, $z_{it} = x_{it}$.[21]
Implementing efficient estimation is one of the main challenges of quantile random effects as the model is overidentified. We suggest two different estimators. The first is an efficient GMM estimator, while the second uses optimal instruments. Given the first stage, we have the following moment restriction:

$$\mathbb{E}[Z_i'(\tilde{X}_i\hat{\beta}_i(\tau) - X_i\delta(\tau)] = 0. \tag{20}$$

If the instrument Z_i contains both the mean and the demeaned regressors, the efficient GMM will optimally weight the within and between variation. The moment condition in equation (20) contains both fast and slow moments, but the fast moments are sufficient to identify the coefficients on the time-varying regressors. The first-order efficient weighting matrix would give zero weights to the slow moment, and the random effects estimator would be identical to the fixed effects estimator. Using the second-order efficient weighting matrix, we obtain a more efficient random effects estimator by also exploiting the between variation. The weighting matrix can be computed as described in section 3. As $T \to \infty$, the relative weights given to the slow moments converge to 0, and the random effect estimator converges to the fixed effects estimator (see Baltagi, 2021; Ahn and Moon, 2014 for a similar argument in least squares models).

If we impose the stronger assumption that the moment restriction in equation (20) holds conditional on Z_i, we can use the theory of optimal instruments to derive a random effects estimator. Optimal instruments are relevant when a researcher has a conditional moment restriction of the form $\mathbb{E}[g_i(\delta,\tau)|Z_i] = 0$. When a moment condition holds conditional on Z_i, an infinite set of valid moments exist, and one could use additional moments to increase efficiency. The goal is to select the instrument that minimizes the asymptotic variance, which takes the form $Z_i^* = \mathbb{E}[g_i(\delta,\tau)g_i(\delta,\tau)'|Z_i]^{-1}R_i(\delta,\tau)$, where $R_i(\delta,\tau) = \mathbb{E}[\frac{\partial}{\partial\delta}g_i(\delta,\tau)|Z_i]$ (see, e.g., Chamberlain, 1987 and Newey, 1993). To implement the random effect estimator with optimal instruments, we set $Z_i = X_i$. Under the additional assumption that $\mathbb{E}[\alpha_i^2(\tau)|X_i] = \sigma_\alpha^2(\tau)$, the optimal instrument simplifies to $Z_i^*(\tau) = \left(\tilde{X}_i\frac{V_i(\tau)}{T}\tilde{X}_i' + \mathbf{l}_T'\mathbf{l}_T\sigma_\alpha^2(\tau)\right)^+ X_i$, where $V_i(\tau)$ is the asymptotic variance from the first stage for an individual i and $^+$ denotes the Moore-Penrose inverse.[22] If $\mathbb{E}[\alpha_i^2(\tau)|X_i] = \sigma_\alpha^2(\tau)$, the random effect estimator based on optimal instruments is efficient.

A few remarks about the optimal instruments follow. First, under standard random effects assumptions, the optimal instrument applied to mean random effects models is numerically

[21] The fixed effects estimator, in general, does not allow estimating γ, as the effect of time-invariant variables is not identified separately from the individual effects. In some situations, it is still possible to estimate γ by strengthening the assumption on the time-invariant regressors x_{2i} without changing the assumptions on time-varying regressors x_{1it}. If x_{2i} is uncorrelated with α_i, it is possible to consistently estimate γ by regressing the fitted values for each quantile τ on x_{it} using demeaned x_{1it} and x_{2i} as instruments. Therefore, our two-step approach allows us to consistently estimate the effect of time-invariant regressors using the same approach as with linear regression.

[22] Since the matrix $(\tilde{X}_i\frac{V_i(\tau)}{T}\tilde{X}_i' + \mathbf{l}_T'\mathbf{l}_T\sigma_\alpha^2(\tau))$ is singular, we use the Moore-Penrose inverse.

identical to the FGLS estimator. Second, in least squares models, using the moment restrictions with the true outcome or the first stage fitted values imply the same optimal instrument. To put it differently, under random effects assumptions, the matrix $\tilde{X}_i \frac{V_i}{T} \tilde{X}_i' + \mathbf{l}_T' \mathbf{l}_T \sigma_\alpha^2$ simplifies to the usual random effects structure. These results are summarized in Proposition 5 in Appendix B.1. Third, if $\sigma_\alpha = 0$, this estimator is identical to the efficient MD estimator (see Proposition 7 in Appendix B.2). Fourth, the optimal instrument depends on T analogously to the second-order optimal weighting matrix of the GMM estimator. As T increases, the first stage variance converges to zero, and the generalized inverse will give infinitely more weights to the within variation and asymptotically converge to the fixed effects estimator.

To make the optimal instrument approach operational, we need a consistent estimator of Z_i^*. In the following, we assume that $\mathbb{E}[\alpha_i^2(\tau)|X_i] = \sigma_\alpha^2(\tau)$ and we suggest estimators for $V_i(\tau)$ and $\sigma_\alpha^2(\tau)$. Compared to the classical random effects structure, we use the first stage variance.[23] This formula has two main advantages. First, it is straightforward to compute $\hat{V}_i$. Second, it is possible to allow for dependence in the errors in the first stage regressions. The first stage variance can be estimated by $\hat{V}_i(\tau) = \hat{A}_i^{-1}(\tau)\hat{B}_i(\tau)\hat{A}_i^{-1}(\tau)$ where $\hat{A}_i(\tau) = \tau(1-\tau)\frac{1}{T}\sum_{t=1}^T \tilde{x}_{it}\tilde{x}_{it}'$ and $B_i(\tau)$ can be computed using the Kernel Density estimator of Powell (1991):

$$\hat{B}_i(\tau) = \frac{1}{Nh}\sum_{i=1}^N K\left(\frac{y_{it} - \tilde{x}_{it}'\beta_i(\tau)}{h}\right)\tilde{x}_{it}\tilde{x}_{it}', \tag{21}$$

where $K(\cdot)$ is the uniform kernel $K(u) = \frac{1}{2}I(|u| \leq 1)$. Alternatively, $V_i(\tau)$ can be estimated by bootstrapping the first stage for each individual separately. We estimate $\sigma_\alpha(\tau)$ using the estimator suggested by Nerlove (1971):

$$\hat{\sigma}_\alpha^2(\tau) = \frac{N}{N-1}\sum_{i=1}^N(\hat{\alpha}_i - \bar{\hat{\alpha}}_i)^2, \tag{22}$$

where $\bar{\hat{\alpha}}_i = \frac{1}{N}\sum_{i=1}^N \hat{\alpha}_i$ and the α_i are estimated by a preliminary least squares dummy variable regression of $\hat{y}_{it}(\tau)$ on x_{it}.[24]

Compared to the optimal instrument approach, efficient GMM relies on weaker conditions, and it is simpler to implement as it does not require a direct estimation of $V_i(\tau)$. Instead, it requires only the consistent estimation of the efficient weighting matrix, which is simpler to estimate.

4.2 Hausman and Taylor Model

The Hausman-Taylor model allows to find instrumental variables from inside the model. It is a middle ground between fixed effects and random effects. On the one hand, the random effects estimator relies on the orthogonality between $\alpha_i(\tau)$ and x_{it}. On the other hand, the fixed

[23] Using the first stage variance will not impose equality on the estimated densities of the errors $\hat{f}_{Y_i - \tilde{X}_i\beta_i}(0)$ in the second stage. Thus, observations will be weighted differently, depending on the first stage variance.

[24] The formula can be modified to accommodate unbalanced panels.

effects estimator only identifies the effect of time-varying variables. To estimate the effect of time-constant variables, Hausman and Taylor (1981) assume that some elements of X_i might be correlated with $\alpha_i(\tau)$. We consider model (1) but we partition X into four types of variables, $X = [X_{1i}^x \ X_{1i}^n \ X_{2i}^x \ X_{2i}^n]$, where the superscript x indicates that the variable is exogenous, and the superscript n indicates that it might be endogenous. Thus,

$$\mathbb{E}[X_{1i}^x \alpha_i(\tau)] = 0$$
$$\mathbb{E}[X_{2i}^x \alpha_i(\tau)] = 0.$$

The assumptions imply that we can estimate $\delta(\tau)$ using the instrument $Z_i = (\dot{X}_{1i}^x, \dot{X}_{1i}^n, \bar{X}_{1i}^x, X_{2i}^x)$. While X_{2i}^n is potentially endogenous, the within variation is uncorrelated with $\alpha_i(\tau)$ as it varies only between i. Identification requires that there are at least as many instruments as parameters to estimate. Hence, we need $dim(x_{1it}^x) \geq dim(x_{2i}^n)$. If the model is overidentified, it is possible to implement efficient GMM, and if conditional moment restrictions are available, optimal instruments can be implemented. The optimal instrument is then $Z_i^*(\tau) = \mathbb{E}\left[\left(\tilde{X}_i(\hat{\beta}_i(\tau) - \beta(\tau)) + \alpha_i(\tau)\right)\left(\tilde{X}_i(\hat{\beta}_i(\tau) - \beta(\tau)) + \alpha_i(\tau)\right)' |Z_i\right]^{-1} \mathbb{E}[X_i|Z_i]$. Implementation of the optimal instrument is not straightforward as it requires the estimation of $\mathbb{E}[X_i|Z_i]$, usually estimated nonpametrically (see Newey, 1993). In this paper, we do not contribute in this direction. In the special case where there is no X_{1i}^n, so that all time-varying regressors are exogenous, the optimal instrument approach can more easily be implemented as the first stage includes only exogenous variables.

4.3 Hausman Test

Consistency of the random effects estimator relies on stronger orthogonality conditions compared to the fixed effects estimator. Under these stronger assumptions, both estimators are consistent, but the fixed effects is inefficient. Hausman (1978) suggested a test for the null hypothesis of random effects against the alternative of fixed effects. This subsection explains how we can use the overidentification test presented in section 3 as a quantile version of the Hausman test for our two-step estimator. Various generalizations of the Hausman test have been suggested in the literature (see, e.g., Chamberlain, 1982; Mundlak, 1978; Wooldridge, 2019). Arellano (1993) considers an heteroskedasticity and autocorrelation robust generalization based on a Wald test. Further, he shows that Chamberlain-type tests can also be computed as a Wald test, testing KT moment restrictions.[25] Ahn and Low (1996) propose a GMM test based on a 3SLS regression as an equivalent method for the Hausman test. In section 4, we suggest second-order efficient GMM as a possibility to perform random effects estimation. One worry in implementing a Hausman test in this setting is that the test would converge to a degenerate distribution as the random effect estimator converges asymptotically to the fixed effects estimator.

[25] These types of tests are not applicable in our context, as $T \to \infty$.

Results in Ahn and Moon (2014) show that this is not the case for mean models. Hence, the overidentification test suggested in section 3 directly extends to the random effect estimator.

The assumption of correct specification of the first stage is maintained both under the null and the alternative hypotheses. Compared to the fixed effects estimator, consistency of the random effects estimator additionally requires that X_i is uncorrelated with $\alpha_i(\tau)$ so that $\mathbb{E}[\dot{X}'_{1i}\alpha_i(\tau)] = 0$ and $\mathbb{E}[\bar{X}'_i\alpha_i(\tau)] = 0$ are a valid moment conditions. By contrast, the fixed effects rely only on the moment condition $\mathbb{E}[\dot{X}'_{1i}\alpha_i(\tau)] = 0$. Consequently, the overidentification test suggested in section 3 can be used as a test of the $\mathbb{H}_0 : \mathbb{E}[\dot{X}'_{1i}\alpha_i(\tau)] = 0$ and $\mathbb{E}[\bar{X}'_i\alpha_i(\tau)] = 0$, which is a test of the random effects orthogonality conditions. Compared to the traditional Hausman test, our test does not rely on the assumption of conditional homoskedasticity of the errors and is robust to clustering.

4.4 Simulations

This section presents simulation results for the different panel data estimators and the Hausman-type test presented in the previous subsections. These simulations focus on the estimation of $\beta(\tau)$, while the next section includes simulation for $\gamma(\tau)$. We consider the following data generating process where all variables are scalars:

$$y_{it} = \beta x_{1it} + \alpha_i + (1 + 0.1x_{1it})\nu_{it}. \tag{23}$$

We let $\beta = 1$ and $\nu_{it} \sim \mathcal{N}(0, 1)$. The regressor is defined by $x_{1it} = h_i + 0.5u_{it}$, where $u_{it} \sim \mathcal{N}(0, 1)$ and

$$\begin{pmatrix} h_i \\ \alpha_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \Lambda \\ \Lambda & 1 \end{pmatrix}\right).$$

If $\Lambda \neq 0$, x_{1it} is correlated with α_i. For the simulation of the panel data estimators, we let $\Lambda = 0$ so that all estimators are consistent. In contrast, in the Monte Carlo study of the Hausman test we let $\Lambda = \{0, 0.1, 0.2, 0.3, 0.4\}$. The true coefficient takes the values $\beta(\tau) = \beta + 0.1F^{-1}(\tau)$ where F is the standard normal CDF. We consider the samples with $T = \{10, 25, 200\}$ and $N = \{25, 200\}$ and focus on the set of quantiles $\mathcal{T} = \{0.1, 0.5, 0.9\}$. All simulation results are based on 10,000 replications. Table 1 shows the bias and the standard deviations and Table 2 shows the standard errors. Simulation results of the rejection probabilities of the Hausman test are in Table 3.

As shown in Table 1, the estimator performs well also when both N and T are small. The RE-GMM (random effects implemented by GMM) estimator performs similarly to the RE-OI (random effects implemented with optimal instruments) estimator and, in some cases, even better. As expected, asymptotically, the RE-GMM, the RE-OI, and the fixed effects (FE) estimators become indistinguishable as T increases. Whereas with small T, there is an apparent gain in using a random effects estimator. From the standard deviations, it is possible to see the different rates of convergence of the estimators. The precision of the fixed effects and random effects estimators increases in similar magnitude when N or T increases. In contrast, the

Quantile	Pooled	BE	FE	RE–OI	RE–GMM
			(N, T) = (25, 10)		
0.1	0.009	0.002	0.037	0.044	0.014
	(0.193)	(0.235)	(0.261)	(0.177)	(0.178)
0.5	0.000	0.000	-0.001	0.000	0.000
	(0.182)	(0.224)	(0.172)	(0.168)	(0.143)
0.9	-0.010	-0.003	-0.039	-0.045	-0.015
	(0.195)	(0.235)	(0.259)	(0.181)	(0.180)
			(N, T) = (200, 10)		
0.1	0.011	0.005	0.040	0.046	0.019
	(0.068)	(0.080)	(0.092)	(0.067)	(0.061)
0.5	0.001	0.001	0.001	0.001	0.001
	(0.063)	(0.076)	(0.059)	(0.063)	(0.047)
0.9	-0.010	-0.003	-0.040	-0.045	-0.018
	(0.067)	(0.080)	(0.091)	(0.068)	(0.060)
			(N, T) = (25, 25)		
0.1	0.003	0.000	0.015	0.016	0.008
	(0.175)	(0.222)	(0.141)	(0.120)	(0.124)
0.5	-0.003	-0.004	0.000	-0.002	-0.002
	(0.171)	(0.218)	(0.102)	(0.106)	(0.099)
0.9	-0.009	-0.007	-0.017	-0.018	-0.013
	(0.177)	(0.223)	(0.138)	(0.120)	(0.124)
			(N, T) = (200, 25)		
0.1	0.006	0.004	0.015	0.017	0.011
	(0.061)	(0.075)	(0.049)	(0.042)	(0.041)
0.5	0.000	0.000	0.000	0.000	0.000
	(0.059)	(0.073)	(0.036)	(0.036)	(0.032)
0.9	-0.006	-0.004	-0.015	-0.017	-0.012
	(0.061)	(0.075)	(0.049)	(0.042)	(0.041)
			(N, T) = (25, 200)		
0.1	0.001	0.002	0.002	0.002	0.002
	(0.163)	(0.211)	(0.049)	(0.047)	(0.056)
0.5	0.001	0.001	0.000	0.000	0.001
	(0.163)	(0.210)	(0.035)	(0.035)	(0.045)
0.9	0.000	0.001	-0.002	-0.002	-0.002
	(0.163)	(0.211)	(0.049)	(0.046)	(0.056)
			(N, T) = (200, 200)		
0.1	0.000	0.000	0.002	0.002	0.002
	(0.058)	(0.073)	(0.017)	(0.016)	(0.017)
0.5	0.000	0.000	0.000	0.000	0.000
	(0.058)	(0.072)	(0.013)	(0.012)	(0.012)
0.9	-0.001	-0.001	-0.002	-0.002	-0.002
	(0.058)	(0.073)	(0.017)	(0.017)	(0.017)

Note:
The table reports bias and standard deviation (in parentheses) of the simulations for $\beta(\tau)$ from 10,000 Monte Carlo simulations.

Table 1: Bias and Standard Deviation of $\hat{\beta}(\tau)$

Quantile	Pooled	BE	FE	RE–OI	RE–GMM
		(N, T) = (25, 10)			
0.1	0.201	0.215	0.254	0.158	0.159
0.5	0.188	0.204	0.166	0.147	0.125
0.9	0.201	0.215	0.254	0.158	0.159
		(N, T) = (200, 10)			
0.1	0.067	0.079	0.091	0.064	0.059
0.5	0.063	0.075	0.060	0.059	0.046
0.9	0.067	0.079	0.091	0.064	0.059
		(N, T) = (25, 25)			
0.1	0.183	0.203	0.138	0.112	0.111
0.5	0.177	0.198	0.100	0.099	0.088
0.9	0.183	0.203	0.138	0.113	0.111
		(N, T) = (200, 25)			
0.1	0.061	0.074	0.049	0.042	0.041
0.5	0.060	0.072	0.036	0.036	0.032
0.9	0.061	0.074	0.049	0.042	0.041
		(N, T) = (25, 200)			
0.1	0.171	0.194	0.048	0.046	0.047
0.5	0.170	0.194	0.035	0.034	0.036
0.9	0.171	0.194	0.048	0.046	0.047
		(N, T) = (200, 200)			
0.1	0.058	0.071	0.017	0.016	0.017
0.5	0.057	0.071	0.013	0.012	0.012
0.9	0.058	0.071	0.017	0.016	0.017

Note:
The table reports standard errors of the simulations for $\beta(\tau)$ from 10,000 Monte Carlo simulations. The standard errors are clustered at the individual level.

Table 2: Standard Errors of $\hat{\beta}(\tau)$

standard deviation of the Pooled and between (BE) estimator decreases only when N increases. The pooled and the between estimators have the smallest bias, but in most cases, also the largest variance.

The standard errors in Table 2 are close to the standard deviations of the simulations, suggesting that our inference procedure performs well also in finite samples. With $T = 10$, the standard errors tend to be slightly undersized in the random effects estimators. The difference is small and decreases quickly as the sample size increase.

Table 3 shows the rejection probabilities of the overidentification test for different values of λ. When $\lambda = 0$, the $\mathbb{H}_0$ is satisfied, so we should reject the null at a rate close to 5%. If $\lambda \neq 0$, X_i is correlated with α_i some moment conditions used by the RE-GMM estimator are not valid. In this case, higher rejection probabilities suggest a more powerful test. The first column shows that the empirical sizes of the test are close to the theoretical levels in most sample sizes. The power of the test is higher in large samples and increases the larger the correlation between $\bar{x}_{1i}$ and the unobserved heterogeneity α_i. An increase in N substantially improves the power of the test, while a larger number of time periods T improves the results to a lesser extent. In general,

Quantile	λ 0.0	0.1	0.2	0.3	0.4
	(N, T) = (25, 10)				
0.1	0.052	0.058	0.077	0.118	0.181
0.5	0.057	0.073	0.117	0.195	0.306
0.9	0.050	0.067	0.095	0.147	0.224
	(N, T) = (200, 10)				
0.1	0.062	0.085	0.276	0.578	0.844
0.5	0.050	0.177	0.533	0.872	0.987
0.9	0.058	0.193	0.483	0.782	0.949
	(N, T) = (25, 25)				
0.1	0.060	0.075	0.121	0.209	0.342
0.5	0.064	0.087	0.152	0.269	0.430
0.9	0.059	0.081	0.140	0.231	0.363
	(N, T) = (200, 25)				
0.1	0.051	0.167	0.555	0.898	0.994
0.5	0.051	0.232	0.691	0.963	0.999
0.9	0.049	0.231	0.646	0.938	0.997
	(N, T) = (25, 200)				
0.1	0.086	0.119	0.212	0.366	0.567
0.5	0.101	0.138	0.248	0.417	0.615
0.9	0.085	0.118	0.218	0.374	0.570
	(N, T) = (200, 200)				
0.1	0.054	0.262	0.773	0.986	1.000
0.5	0.055	0.276	0.792	0.989	1.000
0.9	0.053	0.273	0.787	0.987	1.000

Note:
The table reports rejection probabilities of the Hausman test. The results are based on 10,000 Monte Carlo simulations. The first column, shows the empirical size, while the other columns show the power of the test.

Table 3: Hausman Test

the test performs better both in terms of size and power when N is large, which is most often the case in empirical applications. Although the random effects estimator converges to the fixed effects estimator as T increases, so that random effects coefficients on time-varying regression will be consistent even if $\lambda \neq 0$, the size and power of the test do not deteriorate as T increases. This result is consistent with the findings in Ahn and Moon (2014).

5 Grouped (IV) Quantile Regression Model

In this section, we discuss a particular case of our model in which i indexes groups and t indexes any ordering between groups. The model is of practical relevance when a researcher has micro-data on a sample that can be divided into groups. For instance, groups could be schools, and students in these schools are the individuals. Variables are divided between group-level and individual-level instead of time-varying and time-constant. Individual-level variables include students' characteristics, while school characteristics are group-level variables. Similarly, we

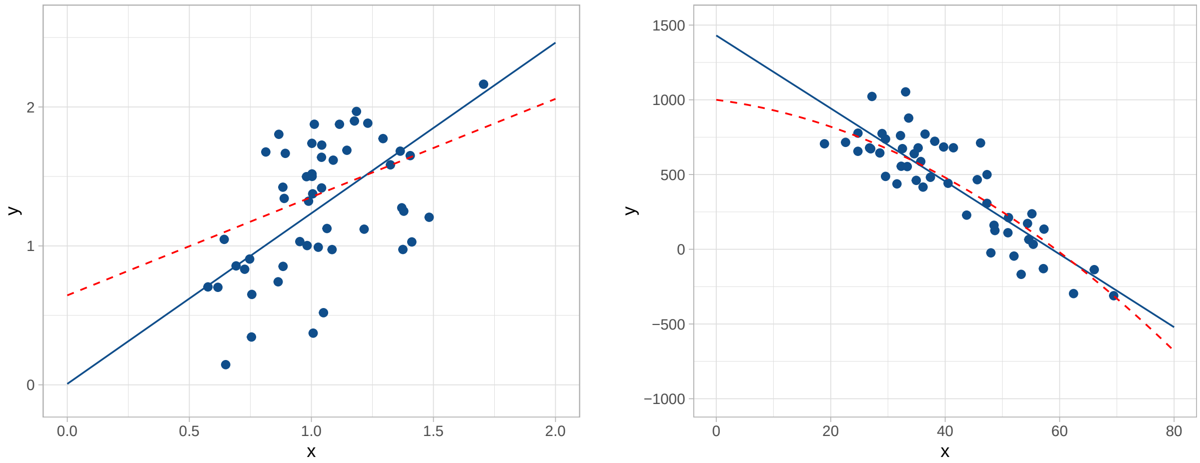

Figure 1: An illustration of first stage estimations

The figure illustrates the venerability of the intercept to extrapolation and misspecification. Both panels show generated data for one group and a first-stage fit. Panel (a) uses the same DGP as in the simulation of CLP. The solid line shows the regression line estimated by median regression, and the dashed red line is the true regression line. Panel (b) uses a different DGP where $y = 1000 - 5x - 0.2x^2 + u$, where $u \sim N(0, 200), x = w \cdot z$, $w \sim U(25, 55)$, $z \sim N(1, 0.2)$. The solid line is estimated by least squares regression without the quadratic term (misspecified). The dashed red line is the true regression line.

might define groups as county-year cells, and individual observations could be economic agents in these counties. In these models, empirical researchers might include fixed effects at the city or county level in the second stage. An estimator for these models was suggested by Chetverikov et al. (2016).

5.1 Chetverikov et al. (2016)

Chetverikov et al. (2016) considers two different models. The first one is identical to model (1). The second model assumes that

$$
\begin{aligned}
Q(\tau, y_{it}|x_{1it}, x_{2i}, v_i) =& \beta_{0,i}(\tau) + x'_{1it}\beta_i(\tau) \quad (24)\\
\beta_{0,i}(\tau) =& x'_{2i}\gamma(\tau) + \alpha(\tau, v_i). \quad (25)
\end{aligned}
$$

Compared to model (1), the coefficient on x_{1it} is allowed to vary over i. They suggest a two-step estimator. The first stage consists in regressing y_{it} on x_{1it} and a constant using quantile regression separately for each group i and quantile τ. In the second stage, they regress the *intercept* from the first stage on x_{2i}. Their estimator focuses on estimating $\gamma(\tau)$ and does not directly provide an estimate of $\beta(\tau)$.

This estimation procedure of Chetverikov et al. (2016) is subject to vulnerabilities. First, it is not invariant to reparametrizations of the individual level variables. Second, the estimator extrapolates the intercept in the first stage, making it vulnerable to misspecification. Figure 1 provides an illustration of these issues. The figure shows two groups from two different samples.

Each panel shows an estimated first-stage regression line (solid blue line) and the true regression line (dashed red line). Panel (a) shows that if the support of the covariates does not cover zero in all groups, the intercept is extrapolated. As shown, a small estimation error in the slope parameter can lead to large estimation errors in the intercept. Further, the value of the intercept, as well as its estimation error, changes with the reparametrization of the individual-level covariates, so that reparametrization lead to different results. By imposing equality of the slopes across groups, our estimator is invariant to reparametrization of the covariates. Panel (b) shows how model misspecification in the first stage can lead to a large estimation error with the CLP estimator. The model is misspecified, as it fits a linear regression model instead of a quadratic one. The consequences of misspecification are substantially larger outside the support of the covariates, e.g., at $x_{1,it} = 0$. In comparison, the misspecification error in the fitted values is negligible. For these reasons, the estimation procedure of Chetverikov et al. (2016) by keeping only the constant looses precision in finite samples.

We want to show that asymptotically our estimator has lower variance than the CLP estimator uniformly over different values of $\text{Var}(\alpha_i(\tau))$. From section 3 we know that the variance comprises two terms, one that accounts for first-stage error and the other accounts for the second-stage noise. The sum of these two terms determines the asymptotic behavior of the estimator. Thus, to show that our MD estimator is more precise, it suffices to show that both components of the variance are smaller. If $\alpha_i(\tau) = 0$ for all i, there is no second stage noise, both estimators are $\sqrt{NT}$-consistent, and only the variance coming from the first stage matters. On the other hand, if $\alpha_i > 0$ for some i, both estimators are $\sqrt{N}$-consistent, and the variance coming from the first stage does not enter the first-order asymptotic distribution. Thus, in the latter case, we will consider the estimators as if we knew the true first stage.

In the following, we assume that the more widely used model (1) is correct and focus on a case with exogenous regressors. We consider our MD estimator implemented using optimal instruments, as this estimator simultaneously minimizes both components of the variance. To apply optimal instruments, we need to impose the stronger assumption that $\mathbb{E}[\alpha_i(\tau)|X_i] = 0$. Below, we provide some results without this assumption.

Consider first the variance arising from the first stage, which shows up in the first order asymptotic distribution of $\hat{\delta}_1(\tau)$. To study this part of the variance, we can assume, without loss of generality, that $\alpha_i(\tau) = 0$ for all i. The optimal instrument is $Z_i^* = \left(\tilde{X}_i V_i(\tau) \tilde{X}_i'\right)^+ X_i$, which yields an estimator that is algebraically identical to the efficient minimum distance estimator (see proposition 7). Remark 6 show the minimum distance representation of our estimator. The CLP estimator can also be written as a minimum distance estimator that minimizes

$$\frac{1}{N} \sum_{i=1}^{N} \left(\hat{\beta}_i(\tau) - \tilde{R}_i \delta(\tau)\right)' \left(\hat{\beta}_i(\tau) - \tilde{R}_i \delta(\tau)\right), \tag{26}$$

where

$$\underset{(K_1+1)\times(K_1\cdot G+K_2)}{\tilde{R}_i} = \begin{pmatrix} x'_{2i} & 0 \\ 0 & j'_i \otimes I_{K_1} \end{pmatrix},$$

and j_i is a N-dimensional vector of zeros with a 1 in the i position. The restriction matrix $\tilde{R}_i$ is different from the restriction matrix of our estimator, as it does not impose equality of the first stage coefficients implied by the model. Further, Chetverikov et al. (2016) uses an identity matrix as a weighting matrix so that their estimator is inefficient relative to an efficient MD estimator with restriction matrix $\tilde{R}_i$. Since our estimator imposes the additional (correct) restriction, our efficient MD estimator has a smaller variance than any alternative (efficient) MD estimator with restriction matrix $\tilde{R}_i$, including the CLP estimator. In the special case of quantile independence, the weighting matrix of the efficient MD estimator reduces to $\hat{W}_i = \tilde{X}'_i\tilde{X}_i$, which corresponds to using OLS in the second stage. In this case, our estimator with an OLS second stage is efficient and will have a lower variance than the CLP estimator.

Next, we focus on the component of the variance coming from the second stage error. For this term, we can assume that we know the true first stage. We start by noting that we numerically obtain the CLP estimator by regressing the first stage fitted values on x_{2i}, $x_{1it} \cdot d_1, \ldots, x_{1it} \cdot d_N$ with instruments x_{2i}, $\dot{x}_{1it} \cdot d_1, \ldots, \dot{x}_{1it} \cdot d_N$ where d_i is a group indicator. In the special case where we know true first stage, we can recover the CLP point estimates if we regress the fitted values on x_{1it} and x_{2i} with instruments $\dot{x}_{1it}$ and x_{2it} without the interactions. From this representation, it is clear that the CLP estimator only exploits the within variation of x_{1it}. Differently, our estimator, uses the entire variation of x_{1it} efficiently. If we know the true first stage, the optimal instrument implied by the conditional moment restriction is $Z_i^* = \mathbb{E}[\alpha_i(\tau)^2|X_i]^{-1}X_i$, which implies that our second stage is a GLS regression which is efficient.

One backdrop of this analysis is that it relies on the stronger conditional moment restriction $\mathbb{E}[\alpha_i(\tau)|X_i] = 0$. Nonetheless, we can show that regardless the value of $\mathrm{Var}(\alpha_i(\tau))$, we can implement an estimator that is more precise than the CLP estimator. More precisely, if $\alpha_i = 0$ for all i, the efficient minimum distance is optimal. While, if $\alpha_i \geq 0$ using an efficient GMM estimator instruments $\dot{x}_{1it}, \bar{x}_{1i}, x_{2i}$ yields more precise point estimates as it exploits all moment conditions efficiently. Compared to the CLP estimator, this GMM estimator exploits the between variation of x_{1it} by including $\bar{x}_{it}$ in the instrument set. By adding an instrument, asymptotically, our estimator will have a weakly lower variance (see Proposition 4.51 in White, 2001).

We have seen that if model (1) is correct, our estimator will, in general, outperform the estimator of CLP estimator because (i) we exploit the exogeneity of the between variation in the time-varying variables and (ii) we impose equality of the slope coefficients. This makes the estimator invariant to reparametrization of the time-varying variables and increases precision. The CLP estimator is consistent for the treatment effect at $x_{1it} = 0$. If the treatment effect is heterogeneous in x_{1it}, then the QTE at 0 may not be particularly interesting. In such a case, one could parametrize the treatment effect on the random slope and estimate the effect of the intercept and the effect on the slope separately. In a second step, both estimates could be combined to get, for instance, an average (in x_{1it}) QTE. Using our approach, we can also allow for heterogeneous effects by including interaction terms between x_{1it} and x_{2i}. We obtain a more precise estimator by estimating all the parameters simultaneously and imposing all the assumptions. On the other hand, if the model (24)-(25) is correct, we should not exploit the between-variation in the time-varying variables and use the demeaned individual-level regressors as instruments. If the slopes are systematically correlated with the treatment variable, the treatment effect is heterogeneous, and CLP estimates the quantile treatment effects at $x_{1it} = 0$, which may not be particularly interesting. Instead, we estimate the best linear approximation of the treatment effect and could easily incorporate interaction terms between the treatment and the time-varying variables to analyze the heterogeneity. Differently, if the slopes are not correlated with the treatment, the treatment effect is homogeneous, and imposing the equality in the coefficients makes the estimator invariant to reparametrization.

5.2 Simulations

This subsection presents Monte Carlo simulations comparing the MD and CLP estimators. The simulations are based on the same data generating process and sample sizes as in Chetverikov et al. (2016). That is, $(T, N) = \{(25, 25), (200, 25), (25, 200), (200, 200)\}$. For both estimator we use a OLS (or 2SLS) second stage. The generated data include one time-invariant regressor, one time-varying regressor and one instrument. Heterogeneity is introduced via a rank variable u_{it}. Since the effect of the individual-level covariates is constant across groups, $\beta(\tau) = (\beta_{i,0}(\tau), \beta_i(\tau)')' = (\beta_0(\tau), \beta(\tau)')'$, where $\beta_{i,0}(\tau) = \beta_0(\tau)$ is the constant of the first stage. The data is generated as follows:

$$y_{it} = \beta_0(u_{it}) + x_{1it}\beta(u_{it}) + x_{2i}\gamma(u_{it}) + \alpha_i(u_{it}), \tag{27}$$

$$z_i = x_{2i} + \eta_i + \nu_i, \tag{28}$$

$$\alpha_i(u_{it}) = u_{it}\eta_i - \frac{u_{it}}{2}, \tag{29}$$

where x_{1it}, x_{2i} and ν_i are distributed $\exp(0.25 \cdot N[0, 1])$ and η_i as well as the rank variable u_{it} are $U[0, 1]$ distributed. The data generating process implies that $\mathbb{E}[\alpha(u_{it})|x_{2i}] = \mathbb{E}[u_{it}\eta_i - \frac{u_{it}}{2}|x_{2i}] = \mathbb{E}[\frac{u_{it}}{2} - \frac{u_{it}}{2}|x_{2i}] = 0$. At quantiles $\tau \in (0, 1)$, the true parameters $\gamma(\tau)$ and $\beta(\tau)$ equal $\sqrt{\tau}$ and, $\alpha_1(\tau) = \frac{\tau}{2}$. Consequently, $\gamma(u_{it}) = \beta(u_{it}) = \sqrt{u_{it}}$ and $\beta_0(u_{it}) = \frac{u_{it}}{2}$. It is worth mentioning that the data generating process of Chetverikov et al. (2016) has a weak instrument when N is small.[26] One should consider this when looking at the simulation results. In empirical research, it is straightforward to construct confidence intervals that remain valid even if identification is weak.

[26] With $N = 25$ in over 40% of the draws, the F-statistics of the first stage of the 2SLS estimations is below 10. The issue disappears when $N = 200$.

	Baseline			Exogenous			Endogenous		
Quantile	MD	CLP	Rel. MSE	MD	CLP	Rel. MSE	MD	CLP	Rel. MSE
				(N, T) = (25, 25)					
0.1	0.022	-0.011	0.051	0.022	-0.010	0.052	0.049	0.001	0.404
	(0.192)	(0.858)		(0.195)	(0.860)		(3.218)	(5.062)	
0.5	-0.010	-0.001	0.061	-0.011	0.000	0.088	-0.017	0.039	0.318
	(0.166)	(0.673)		(0.204)	(0.691)		(3.098)	(5.491)	
0.9	-0.019	-0.003	0.049	-0.020	-0.004	0.216	-0.052	-0.011	0.409
	(0.094)	(0.435)		(0.227)	(0.490)		(3.239)	(5.065)	
				(N, T) = (200, 25)					
0.1	0.024	0.003	0.060	0.024	0.004	0.063	0.023	0.006	0.057
	(0.066)	(0.284)		(0.067)	(0.285)		(0.106)	(0.456)	
0.5	-0.006	-0.001	0.059	-0.006	0.000	0.086	-0.009	-0.003	0.071
	(0.056)	(0.232)		(0.069)	(0.238)		(0.097)	(0.366)	
0.9	-0.017	-0.004	0.060	-0.017	-0.003	0.223	-0.022	-0.009	0.142
	(0.031)	(0.145)		(0.075)	(0.164)		(0.086)	(0.234)	
				(N, T) = (25, 200)					
0.1	0.003	-0.002	0.059	0.003	-0.001	0.066	-0.027	-0.076	0.130
	(0.070)	(0.289)		(0.074)	(0.291)		(2.025)	(5.618)	
0.5	-0.001	-0.002	0.060	-0.001	-0.001	0.233	-0.082	-0.094	0.580
	(0.060)	(0.247)		(0.134)	(0.278)		(3.485)	(4.575)	
0.9	-0.002	0.000	0.061	-0.001	0.001	0.769	-0.118	-0.114	1.126
	(0.030)	(0.121)		(0.217)	(0.247)		(3.780)	(3.561)	
				(N, T) = (200, 200)					
0.1	0.003	-0.003	0.057	0.003	-0.003	0.062	0.002	-0.004	0.058
	(0.024)	(0.100)		(0.025)	(0.101)		(0.039)	(0.162)	
0.5	-0.001	0.000	0.059	-0.001	-0.001	0.222	-0.004	-0.004	0.141
	(0.020)	(0.084)		(0.044)	(0.093)		(0.051)	(0.136)	
0.9	-0.002	0.000	0.067	-0.003	-0.001	0.762	-0.009	-0.007	0.617
	(0.010)	(0.040)		(0.071)	(0.082)		(0.074)	(0.095)	

Note:
The table reports mean bias, standard deviation and relative MSE from the simulations for $\gamma(\tau)$ from 10000 Monte Carlo simulations using the MD estimator and the CLP estimator. The relative MSE gives the MSE of the MD estimator relative to that of the CLP estimator.

Table 4: Bias, Standard Deviation and MSE of $\hat{\gamma}(\tau)$

The simulations consider three cases. In the first one (baseline), $\alpha_i(\tau) = 0$ for all i and all τ. In this case, conditioning on the individual does not affect the quantile function. Further, as $\alpha_i(\tau) = 0$, the estimators with a least square second stage are $\sqrt{NT}$-consistent. In the second case, there are individual specific effects ($\alpha_i(\tau) \neq 0$), which are uncorrelated with the regressors. The individual heterogeneity is multiplied with the rank variable u_{it}. Thus, in the lower tail of the distribution, we see a faster convergence rate. In the third case, $\alpha_i(\tau)$ is correlated with the regressor of interest, such that x_{2i} is endogenous. In this case, we use 2SLS in the second stage. We perform 10,000 Monte Carlo replications for the set of quantiles $\tau \in \{0.1, 0.5, 0.9\}$. Since the CLP estimator does not directly provide an estimate for $\beta(\tau)$, we present only results for $\gamma(\tau)$.

Table 4 reports the bias, standard deviation and relative MSE of the estimators. The relative MSE reports the MSE of the MD estimator relative to that of the CLP estimator. Thus, a number smaller than 1 indicates that the MD estimator has a lower MSE. The CLP estimator

	Baseline			Exogenous			Endogenous		
	Rel. length	Coverage Rate		Rel. length	Coverage Rate		Rel. length	Coverage Rate	
Quantile	MD/CLP	MD	CLP	MD/CLP	MD	CLP	MD/CLP	MD	CLP
				(N, T) = (25, 25)					
0.1	0.232	0.941	0.938	0.235	0.939	0.938	0.227	0.966	0.972
0.5	0.244	0.940	0.942	0.301	0.942	0.945	0.262	0.964	0.972
0.9	0.223	0.942	0.949	0.501	0.940	0.946	0.373	0.957	0.972
				(N, T) = (200, 25)					
0.1	0.230	0.932	0.947	0.233	0.932	0.948	0.230	0.942	0.953
0.5	0.245	0.947	0.944	0.296	0.945	0.946	0.267	0.952	0.949
0.9	0.220	0.925	0.947	0.475	0.941	0.945	0.368	0.953	0.952
				(N, T) = (25, 200)					
0.1	0.241	0.943	0.940	0.256	0.943	0.943	0.240	0.968	0.974
0.5	0.242	0.937	0.944	0.496	0.938	0.944	0.370	0.949	0.971
0.9	0.248	0.948	0.941	0.884	0.934	0.945	0.771	0.939	0.955
				(N, T) = (200, 200)					
0.1	0.241	0.944	0.944	0.254	0.947	0.945	0.246	0.951	0.950
0.5	0.244	0.946	0.945	0.483	0.952	0.948	0.377	0.957	0.951
0.9	0.246	0.942	0.953	0.872	0.950	0.950	0.772	0.954	0.955

Note:
Results based on 10000 Monte Carlo simulations. The table provides the coverage rate and median length of the confidence intervals of $\gamma(\tau)$. The relative length provides the length of the confidence interval of the MD estimator relative to that of the CLP estimator. Robust standard errors are used for the CLP estimator, and clustered standard errors at the group level are used for the MD estimator.

Table 5: Properties of the 95% Confidence Intervals

seems to have a smaller bias than the MD estimator when $T = 25$. When T increases to 200, the difference disappears. There are remarkable differences in the variance of the estimators. The standard deviation of the MD estimator is four times smaller compared to that of the CLP estimator in the baseline case. The difference is somewhat smaller in the exogenous and endogenous cases but remains substantial.[27] This difference in precision explains the large discrepancies in MSE. The MSE of the CLP estimator is over 10 times larger than that of the MD estimator when $\alpha_i(\tau) = 0$ and remains substantially larger in all scenarios considered. If $\alpha_i(\tau) = 0$, quantile regression is a consistent estimator for $\beta(\tau)$. Although not shown here, simulation results comparing our estimator with traditional quantile regression show that the two estimators are indistinguishable in terms of bias and variance in large samples.

Table 5 show the performance of the 95% confidence intervals suggested with our inference procedure. The table reports the coverage rate of the confidence intervals and the median length of the confidence interval of our estimator relative to that of the CLP estimator. Our suggested inference procedure has coverage close to 95% in all cases. Compared to the CLP estimator, our confidence bands are substantially shorter. In most cases, our estimator yields confidence bands less than half the length of those for the CLP estimator.

[27]The standard deviations in the endogenous case with $N = 25$ should be interpreted with caution due to the weak instrument.

6 Empirical Application: The Effect of the Food Stamps Program on Birth Weight

In this section, we apply our minimum distance approach to estimate the impact of the food stamp program on the birth weight distribution using grouped data. We complement the analysis of Almond et al. (2011) by providing distributional effects. Food stamps constitute an important means-tested program that gives entitled households coupons they can redeem at approved retail food stores. The Food Stamp Act (FSA) was introduced in 1964 and enabled counties to start their own federally funded food stamp program (FSP). In the subsequent years, counties increasingly adopted such programs, and in 1973, an amendment to the FSA required all counties to establish a FSP by 1975. Thus, the share of counties with an FSP increased steadily from 1964 to 1974, and identification exploits the variation in the timing of the adoption across counties. Almond et al. (2011) use data from 1968 (when about 40% of the counties had introduced the program) to 1977 (two years after the FSP was implemented everywhere) to analyze the effect of the program.

Given the negative consequences of low birth weight, besides estimating the effect of the policy on average weight, Almond et al. (2011) estimate the effect on the probability that birth weight falls below a certain threshold. As discussed in Melly and Santangelo (2015) this procedure leads to biased results unless there is no time effect or group effect, or the outcome is uniformly distributed.

In this section, we use the subscripts j, c, and t to denote the birth, the county, and the trimester of birth, respectively.[28] The variable of interest, the FSP, is coded 1 if there was a food stamp program in place three months before birth. Thus, the treatment is assigned to county-month cells, and in around 1% of cases, it also varies within groups.

We consider the following model separately for blacks and whites:

$$Q(\tau, bw_{jct}|fsp_{ct}, x_{1jct}, x_{2ct}, v_{ct}) = fsp_{ct}\gamma(\tau) + x_{1jct}\beta(\tau) + x_{2ct}\gamma_2(\tau) + \alpha(\tau, v_{ct}), \tag{30}$$

where $Q(\tau, bw_{jct}|fsp_{ct}, x_{1jct}, x_{2ct}, v_{ct})$ is the conditional quantile function of the outcome given all the variables. fsp_{ct} is a variable indicating whether there is a food stamp program in place, x_{1jct} are variables related to the individual births, such as gender, mother age, and its square as well as the legitimacy status of the birth.[29] x_{2ct} are annual county-level controls (real per capita income, government transfers to individuals, medical spending, and retirement and disability payments) and 1960 county-level characteristics (county population and the shares of urban population, black population, and of farmland) interacted with a linear time trend. $\alpha_{ct}(\tau, v_{ct})$ is a group level unobserved heterogeneity and v_{ct} is an unrestricted random vector.

[28] Using the same notation as in the paper, the i units are county-trimester combinations (which we call groups), and the t units index individual births within a county in a given trimester. However, in this section, we use three subscripts for clarity.

[29] To be precise, we should write fsp_{jct} as the treatment variable, in a few instances, takes different values within a group.

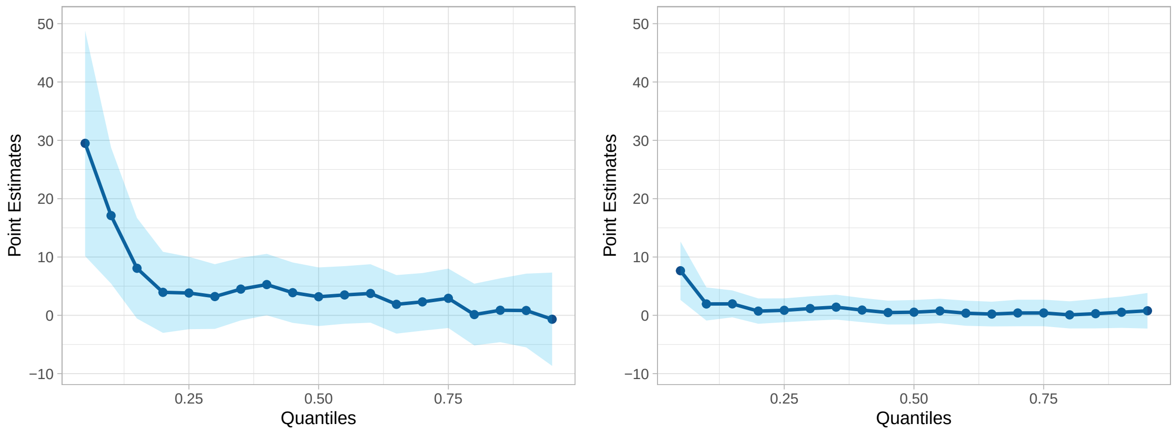

Figure 2: Impact of Food Stamp Introduction on the Distribution of Birth Weight

The figure shows the impact of the food stamp introduction on the conditional distribution of birth weight. The panels show point estimates and 95% confidence bands (shaded area) computed using standard errors clustered at the county level. The panel on the left (right) shows the effects for blacks (whites). The regressions include county, time, and state-year fixed effects.

We model the unobserved heterogeneity with county, state-year fixed effects, and time fixed effects.

Figure 2 illustrates the results. The estimations are performed using a sample of 2,751,607 individual observations divided into 16'857 groups for blacks and 16,030,197 individual births divided into 79,948 groups in the sample of whites.[30] The results for black are in the left panel, while the results for whites are in the right panel. As shown, the effect is substantially larger among blacks. The results suggest a positive effect of the food stamp program on the lower tail of the conditional distribution. The estimates suggest that the food stamp program is associated with an increase in birth weight by almost 30 grams for blacks at the 5th percentile of the conditional distribution. For whites, there seems to be an effect only at the left tail of the distribution, and the effects are small. For blacks, the coefficients are large in the left tail and remain positive, albeit of small magnitude, until the 75% percentile. However, for higher quantiles, the effects are not statistically different from zero.

7 Conclusion

This paper suggests a MD estimator for quantile panel data models. The estimator is of practical relevance with classical panel data settings where the units are observed over time and with grouped data, where individuals are divided into groups, and the treatment varies at the

[30] We have a different number of groups compared to Almond et al. (2011) due to multiple reasons. First, they give higher weights to births in groups where only 50% of the births are included in the natality data; thus, when they drop groups with less than 25 births, the number of births in these groups is inflated. Second, since they take the group average, they keep births with missing values for birth weight. We drop those births as we work with individual-level data.

group level. The coefficient on the time-varying and time-invariant variables can be estimated. The estimator is computationally fast and straightforward to compute and consists of a first stage individual level quantile regressions, followed by a GMM regression with the fitted values as the dependent variable. We show that our two-step procedure applied to linear estimators is algebraically identical to traditional one-step estimators. We suggest a quantile counterpart to traditional panel data estimators, including the pooled, the fixed effects, and the random effects estimator. In the second stage, both internal and external instruments can be used in a Hausman and Taylor or traditional instrumental variables framework. Further, an overidentification test can be implemented if the model is overidentified.

The rate of convergence of the moment condition used to identify a parameter determines the convergence rate of its estimator. Using an efficient weighting matrix, the coefficients of the time-varying variables are $\sqrt{NT}$-consistent, while the coefficients of the time-invariant variables are $\sqrt{N}$-consistent. For the coefficient converging at the faster rate, only the variance coming from the first stage enters the first-order asymptotic distribution. On the other hand, since T diverges to infinity, the first stage variance would not appear in the first-order asymptotic variance of the coefficients converging at the slower rate. In other words, the first-order asymptotic distribution is the same as if we knew the true first stage. We suggest an inference procedure that is uniformly valid regardless of the convergence rate of the estimator and, importantly, takes the first stage variance into account, thus providing more accurate inference. Monte Carlo simulations show that our estimator and the suggested standard errors perform well in finite samples. Compared to the grouped estimator of Chetverikov et al. (2016), the MD estimator has a much smaller MSE. Finally, in an empirical application, we study the effect of the food stamp program on birth weight, and we find large effects for blacks in the lower tail of the conditional distribution.

References

Ahn, S. C. and S. Low (1996): "A reformulation of the Hausman test for regression models with pooled cross-section-time-series data," *Journal of Econometrics*, 71, 309–319.

Ahn, S. C. and H. R. Moon (2014): "Large-N and Large-T Properties of Panel Data Estimators and the Hausman Test," in *Festschrift in Honor of Peter Schmidt*, Springer.

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

Alvarez, J. and M. Arellano (2003): "The time series and cross-section asymptotics of dynamic panel data estimators," *Econometrica*, 71, 1121–1159.

Angrist, J., V. Chernozhukov, and I. Fernández-Val (2006): "Quantile Regression under Misspecification, with an Application to the U.S. Wage Structure," *Econometrica*, 74, 539–563.

Angrist, J. D. and K. Lang (2004): "Does school integration generate peer effects? Evidence from Boston's metco program," *American Economic Review*, 94, 1613–1634.

Arellano, M. (1993): "On the testing of correlated effects with panel data," *Journal of Econometrics*, 59, 87–97.

Autor, D. H., D. Dorn, and G. H. Hanson (2013): "The China syndrome: Local labor market effects of import competition in the United States," *American Economic Review*, 103, 2121–2168.

Baltagi, B. H. (2021): *Econometric Analysis of Panel Data*, Springer, 6 ed.

Brown, B. M. (1971): "Martingale central limit theorems," *The Annals of Mathematical Statistics*, 59–66.

Canay, I. A. (2011): "A Simple Approach to Quantile Regression for Panel Data," *Econometrics Journal*, 14, 368–386.

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

——— (1987): "Asymptotic efficiency in estimation with conditional moment restrictions," *Journal of Econometrics*, 34, 305–334.

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

Chernozhukov, V., I. Fernández-Val, and B. Melly (2013): "Inference on Counterfactual Distributions," *Econometrica*, 81, 2205–2268.

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

——— (2006): "Instrumental quantile regression inference for structural and treatment effect models," *Journal of Econometrics*, 132, 491–525.

Chesher, A. (2003): "Identification in nonseparable models," *Econometrica*, 71, 1405–1441.

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

Dai, X. and L. Jin (2021): "Minimum distance quantile regression for spatial autoregressive panel data models with fixed effects," *PLoS ONE*, 16, 1–13.

Fernandez-Val, I., W. Y. Gao, Y. Liao, and F. Vella (2022): "Dynamic Heterogeneous Distribution Regression Panel Models, with an Application to Labor Income Processes," 1–45.

Frölich, M. and B. Melly (2013): "Unconditional Quantile Treatment Effects Under Endogeneity," *Journal of Business and Economic Statistics*, 31, 346–357.

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

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

Galvao, A. F. and K. Kato (2016): "Smoothed quantile regression for panel data," *Journal of Econometrics*, 193, 92–112.

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

Gu, J. and S. Volgushev (2019): "Panel Data Quantile Regression with Grouped Fixed Effects," *Journal of Econometrics*, 213, 68–91.

Hahn, J. and G. Kuersteiner (2002): "Asymptotically unbiased inference for a dynamic panel model with fixed effects when both n and T are large," *Econometrica*, 70, 1639–1657.

Hansen, B. E. (2022): *Econometrics*, Princeton University Press.

Hansen, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054.

Harding, M., C. Lamarche, and M. H. Pesaran (2020): "Common correlated effects estimation of heterogeneous dynamic panel quantile regression models," *Journal of Applied Econometrics*, 35, 294–314.

Hausman, J. and W. E. Taylor (1981): "Panel Data and Unobservable Individual Effects," *Econometrica*, 49, 1377.

Hausman, J. A. (1978): "Specification Tests in Econometrics," *Econometrica*, 46, 1251–1271.

Hoderlein, S. and E. Mammen (2007): "Identification of marginal effects in nonseparable models without monotonicity," *Econometrica*, 75, 1513–1518.

Im, K. S., S. C. Ahn, P. Schmidt, and J. M. Wooldridge (1999): "Efficient estimation of panel data models with strictly exogenous explanatory variables," *Journal of Econometrics*, 93, 177–201.

Kato, K., A. Galvao Jr., and G. V. Montes-Rojas (2012): "Asymptotics for Panel Quantile Regression Models with Individual Effects," *Journal of Econometrics*, 170, 76–91.

Knight, K. (1998): "Limiting Distributions for L1 Regression Estimators under General Conditions," *Annals of Statistics*, 26, 755–770.

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

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

Koenker, R. and P. Ng (2005): "A frisch-newton algorithm for sparse quantile regression," *Acta Mathematicae Applicatae Sinica*, 21, 225–236.

Ma, L. and R. Koenker (2006): "Quantile regression methods for recursive structural equation models," *Journal of Econometrics*, 134, 471–506.

Melly, B. and G. Santangelo (2015): "The changes-in-changes model with covariates," 1–32.

Mundlak, Y. (1978): "On the Pooling of Time Series and Cross Section Data," *Econometrica*, 46, 69–85.

Nerlove, M. (1971): "Further Evidence on the Estimation of Dynamic Economic Relations from a Time Series of Cross Sections," *Econometrica*, 39, 359.

Newey, W. K. (1993): "Efficient estimation of models with conditional moment restrictions," in *Handbook of Statistics*, Elsevier, vol. 11, chap. 16, 419–454.

Newey, W. K. and J. L. Powell (1990): "Efficient Estimation of Linear and Type I Censored Regression Models Under Conditional Quantile Restrictions," *Econometric Theory*, 6, 295–317.

Phillips, P. C. and H. R. Moon (1999): "Linear regression limit theory for nonstationary panel data," *Econometrica*, 67, 1057–1111.

Powell, J. L. (1991): "Estimation of Monotonic Regression Models under Quantile Restriction," in *Nonparametric and Semiparametric Model in Economics*, Cambridge: Cambridge University Press.

Volgushev, S., S.-K. Chao, and G. Cheng (2019): "Distributed inference for quantile regression processes," *The Annals of Statistics*, 47, 1634–1662.

White, H. (2001): *Asymptotic theory for econometricians*, Academic press.

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, 137–150.

A Least Squares Panel Data Models

This section complements subsection 2.3 by discussing more in detail the relationship between least squares estimator and the minimum distance approach. Throughout the section, we define the $T \times (K_1+1)$ matrix of first-stage regressors $\tilde{X}_{1i} = (\tilde{x}_i, \tilde{x}_{i2}, \dots, \tilde{x}_{iT})'$, and the $NT \times K_1$ matrix of time-varying regressors $X_1 = (X_{1i}', \dots, X_{1N}')'$. Further, we use the matrices $P_i = l(l'l)^{-1}l'$ and $Q_i = I_i - P_i$, where l is a $T \times 1$ vector of ones. Thus, $P_i X_i = \bar{X}_i$ and $Q_i X_{1i} = \dot{X}_i$. We consider a linear version of our estimator, where OLS instead of quantile regression is used in the first stage. We consider model (9). In this section, we show that mean models can be estimated using a two-step procedure. Notation is the same as in the paper, except that the fitted values are computed using an OLS regression. More precisely, the vector of fitted values of individual i is

$$\hat{Y}_i = \tilde{X}_i \hat{\beta}_i = \tilde{X}_i \left(\tilde{X}_i' \tilde{X}_i\right)^{-1} \tilde{X}_i' Y_i.$$

The following Proposition states the equivalence of the two-step procedure using the fitted values and the conventional one-step estimator in mean models.

Proposition 3. *Denote $\hat{\delta}_{GMM}^{MD}$ the coefficient vector of a linear GMM regression of $\hat{Y}$ on X with instrument Z. Let $\hat{\delta}_{GMM}$ be the coefficient vector of the same GMM regression but with regressand Y. If $\tilde{X}_i$ lies in the column space of Z_i, $\hat{\delta}_{GMM}^{MD} = \hat{\delta}_{GMM}$.*

The proof of this Proposition and all subsequent proofs are in Appendix B.1. Proposition 3 implies that any linear model can be computed by a two-step estimator as long as the matrix of instruments of individual i, Z_i lies in the column space of the matrix of first-stage regressors of individual i, $\tilde{X}_i$.[31] This result applies to a wide range of estimators. Since OLS is a special case of GMM, the result for pooled OLS follows directly, while the result for the within estimator is summarized in the following Corollary.

Corollary 1. *Denote $\hat{\delta}_{FE}^{MD}$ the coefficient vector of a 2SLS regression of $\hat{Y}$ on $\tilde{X}$ with instruments $\dot{X}_1$. Let $\hat{\delta}_{FE}$ be the coefficient vector of the within estimator, that is, of a regression of $\dot{Y}$ on $\dot{X}_1$. Then $\hat{\delta}_{FE}^{MD} = \hat{\delta}_{FE}$.*

The between estimator is usually computed by regressing $\bar{Y}$ on $\bar{X}$. Alternatively, it can be estimated by an IV regression of Y (or $\hat{Y}$) on X using $\bar{X}$ as an instrument, where it exploits only the variation between individuals.

Corollary 2. *Denote $\hat{\delta}_{BE}^{MD}$ the coefficient vector of a 2SLS regression of $\hat{Y}$ on X with instruments $\bar{X}$. Let $\hat{\delta}_{BE}$ be the coefficient vector of the between estimator, that is, of a regression of $\bar{Y}$ on $\bar{X}$. Then $\hat{\delta}_{BE}^{MD} = \hat{\delta}_{BE}$.*

It is worth noting that the IV approach to these panel data estimators also works in one stage with Y as the dependent variable. Further, it is possible to estimate between (within)

[31] Since $\tilde{X}_i$ includes a constant, the presence of time-invariant variables in Z_i will not affect its column space.

models using average (demeaned) fitted values and regressors.
The pooled OLS and the between estimators can estimate both β and γ but are not efficient. The random effects estimator optimally combines between and the within variation to find a more efficient estimator. While FGLS is the most common estimator for the random effects model, Im et al. (1999) show that the overidentified 3SLS estimator, with instruments $Z_i = (\dot{X}_{1i}, \bar{X}_i)$, is identical to the random effects estimator. The 3SLS estimator is a special case of GMM with weighting matrix $W = \mathbb{E}[Z_i'\tilde{\Omega}Z_i]$ where $\tilde{\Omega}$ follows the usual random effects covariance structure. Thus, by Proposition 3, the random effects estimator can also be computed in two steps using the fitted values in the second stage.

Corollary 3. *Denote $\hat{\delta}_{RE}^{MD}$ the coefficient vector of a 3SLS regression of $\hat{Y}$ on X with instruments $(\dot{X}_{1i}, \bar{X}_i)$. Let $\hat{\delta}_{RE}$ be the coefficient vector of a random effects regression of Y on X. Then $\hat{\delta}_{RE}^{MD} = \hat{\delta}_{RE}$.*

Alternatively, the random effects estimator can be implemented using the theory of optimal instruments and a just identified 2SLS regression. Starting from a conditional moment restriction, the idea of optimal instruments is to select an instrument and weights that minimize the asymptotic variance (see, e.g. Newey, 1993). Relevant to our two-step procedure, under homoskedasticity of the errors, the conditional moments $\mathbb{E}[Y_i - X_i\delta|X_i] = 0$ and $\mathbb{E}[\hat{Y}_i - X_i\delta|X_i] = 0$ imply the same optimal instrument (see Proposition 5 in Appendix B.1).

The Hausman-Taylor model (Hausman and Taylor, 1981) is a middle ground between the fixed effects and the random effects models where some regressors are assumed to be uncorrelated with α_i. In contrast, no restriction is placed on the relationship between the other regressors and the unobserved heterogeneity. The matrix of regressors X is partitioned as $X = [X_1^x\ X_1^n\ X_2^x\ X_2^n]$ where X_1^x and X_2^x are orthogonal to α_i. No assumption is placed on the relationship between α_i and X_1^n and X_2^n. The model can be estimated by IV using instruments $Z = (\dot{X}_1^x, \dot{X}_1^n, \bar{X}_1^x, X_2^x)$ (see, e.g., Hansen, 2022). Thus, it follows by Proposition 3 that the Hausman Taylor model can be estimated in two stages.

Proposition 4. *Denote $\hat{\delta}_{HT}^{MD}$ the coefficient vector of a 2SLS regression of $\hat{Y}$ on X with instruments $(\dot{X}_1^x, \dot{X}_1^n, \bar{X}_1^x, X_2^x)$. Let $\hat{\delta}_{HT}$ be the coefficient vector of the Hausman Taylor Estimator based on a regression Y on X. Then $\hat{\delta}_{HT}^{MD} = \hat{\delta}_{HT}$.*

B Proofs

B.1 Linear Models

Proof of Proposition 3. Define the projection matrix $\tilde{P} = \tilde{X}_i(\tilde{X}_i'\tilde{X}_i)^{-1}\tilde{X}_i'$. Since Z_i is in the column space of $\tilde{X}_i$,

$$\tilde{P}Z_i = Z_i \tag{31}$$

The MD estimator with a GMM second stage is:

$$\hat{\delta}^{MD}_{GMM} = \left(X'Z\hat{W}Z'X\right)^{-1} X'Z\hat{W}Z'\hat{Y}.$$

For $\hat{\delta}^{MD}_{GMM}$ to be equal to $\hat{\delta}_{GMM}$, it suffices that $Z'\hat{Y} = Z'Y$. Note that

$$\begin{aligned}
Z'\hat{Y} &= \sum_{i=1}^{N} Z_i'\hat{Y}_i \\
&= \sum_{i=1}^{N} Z_i'\tilde{X}_i\hat{\beta}_i \\
&= \sum_{i=1}^{N} Z_i'\tilde{X}_i(\tilde{X}_i'\tilde{X}_i)^{-1}\tilde{X}_i'Y_i \\
&= \sum_{i=1}^{N} (\tilde{P}Z_i)'Y_i \\
&= \sum_{i=1}^{N} Z_i'Y_i = Z'Y,
\end{aligned}$$

where the third line uses $\hat{Y}_i = \tilde{X}_i\hat{\beta}_i$, the fourth line uses the definition of the OLS estimator in the first stage and the last line uses equation (31). Thus, it follows directly that $\hat{\delta}_{MD}$ equals $\hat{\delta}_{GMM}$. ∎

Proof of Corollary 1. First, note that since $Q_iX_{1i} = \dot{X}_{1i}$, $\dot{X}_{1i}$ lies in the column space of X_{1i}. Then, we apply Proposition 3 and since $K = L$, the 2SLS estimator reduces to the IV estimator. It follows that a 2SLS (or IV) regression of $\hat{Y}$ on X_{1i} with instrument Z_i is algebraically identical to a 2SLS regression with Y_i as dependent variable. Then,

$$\begin{aligned}
\hat{\delta}^{MD}_{FE} &= \left(\sum_{i=1}^{N} Z_i'X_{1i}\right)^{-1} \sum_{i=1}^{N} Z_i'Y_i \\
&= \left(\sum_{i=1}^{N} \dot{X}_{1i}'X_{1i}\right)^{-1} \sum_{i=1}^{N} \dot{X}_{1i}'Y_i \\
&= \left(\sum_{i=1}^{N} X_{1i}'Q_iX_{1i}\right)^{-1} \sum_{i=1}^{N} X_{1i}'Q_iY_i \\
&= \left(\sum_{i=1}^{N} \dot{X}_{1i}'\dot{X}_{1i}\right)^{-1} \sum_{i=1}^{N} \dot{X}_{1i}'\dot{Y}_i = \hat{\delta}_{FE},
\end{aligned}$$

where the second line follows since $Z_i = \dot{X}_{1i}$, the third and last line by $Q_iX_{1i} = \dot{X}_{1i}$, $Q_iY_i = \dot{Y}_i$ and since Q_i is idempotent. ∎

Proof of Corollary 2. First, note that since $P_i\tilde{X}_i = \bar{X}_i$, $\bar{X}_i$ lies in the column space of $\tilde{X}_i$. Then, we apply Proposition 3 and since $K = L$, the 2SLS estimator reduces to an IV estimator. It follows that a 2SLS regression of $\hat{Y}$ on X_i with instrument Z_i is algebraically identical to a 2SLS regression with Y_i as dependent variable. Then,

$$\begin{aligned}
\hat{\delta}_{BE}^{MD} &= \left(\sum_{i=1}^{N} Z_i'X_i\right)^{-1} \sum_{i=1}^{N} Z_i'Y_i \\
&= \left(\sum_{i=1}^{N} \bar{X}_i'X_i\right)^{-1} \sum_{i=1}^{N} \bar{X}_i'Y_i \\
&= \left(\sum_{i=1}^{N} X_i'P_iX_i\right)^{-1} \sum_{i=1}^{N} X_i'P_iY_i \\
&= \left(\sum_{i=1}^{N} \bar{X}_i'\bar{X}_i\right)^{-1} \sum_{i=1}^{N} \bar{X}_i'\bar{Y}_i = \hat{\delta}_{BE}
\end{aligned}$$

where the second line follows since $Z_i = \bar{X}_i$, the third and last line by $P_iX_i = \bar{X}_i$, $P_iY_i = \bar{Y}_i$ and, since P_i is idempotent. ∎

Proposition 5. *Assume $\mathbb{E}[\varepsilon_{it}^2|X_i] = \sigma_\varepsilon^2$ and $\mathbb{E}[\alpha_i^2|X_i] = \sigma_\alpha^2$. The conditional moments $\mathbb{E}[\hat{Y}_i - X_i\delta|X_i] = 0$ and $\mathbb{E}[Y_i - X_i\delta|X_i] = 0$ imply the same optimal instrument.*

Proof. The optimal instrument takes the form $Z_i^* = \mathbb{E}[g_i(\delta)g_i(\delta)'|Z_i]^{-1}R_i(\delta,\tau)$, where $R_i(\delta,\tau) = \mathbb{E}[\frac{\partial}{\partial\delta}g_i(\delta,\tau)|Z_i]$. For both moment conditions, $R_i(\delta,\tau)$ is identical. Then for the first moment restriction, we have:

$$\begin{aligned}
\mathbb{E}[(\hat{Y}_i - X_i\delta)(\hat{Y}_i - X_i\delta)'|X_i] &= \mathbb{E}[(\tilde{X}_i(\hat{\beta}_i - \beta) + \tilde{X}_i\beta - X_i\delta)(\tilde{X}_i(\hat{\beta}_i - \beta) + \tilde{X}_i\beta - X_i\delta)'|X_i] \quad (32)\\
&= \mathbb{E}[(\tilde{X}_i(\hat{\beta}_i - \beta) + \alpha_i)(\tilde{X}_i(\hat{\beta}_i - \beta) + \alpha_i)'|X_i] \\
&= \tilde{X}_i\frac{V_i}{T}\tilde{X}_i' + \mathbf{l}_T\mathbf{l}_T'\sigma_\alpha^2.
\end{aligned}$$

The matrix $\tilde{X}_i\frac{V_i}{T}\tilde{X}_i' + \mathbf{l}_T\mathbf{l}_T'\sigma_\alpha^2$ is singular, so that we suggest using the Moore-Penrose inverse to construct the optimal instrument.

For the second moment restriction, we have:

$$\begin{aligned}
\mathbb{E}[(Y_i - X_i\delta)(Y_i - X_i\delta)'|X_i] &= \mathbb{E}[(\alpha_i + \varepsilon_{it})(\alpha_i + \varepsilon_{it})'|X_i] \\
&= \left(\mathbf{I}_T\sigma_\varepsilon^2 + \mathbf{l}_T\mathbf{l}_T'\sigma_\alpha^2\right).
\end{aligned}$$

Then note that $\left(\mathbf{I}_T\sigma_\varepsilon^2 + \mathbf{l}_T\mathbf{l}_T'\sigma_\alpha^2\right)^{-1} = (\tilde{X}_i\tilde{X}_i^+\sigma_\varepsilon^2 + \mathbf{l}_T'\mathbf{l}_T\sigma_\alpha^2)^+ = (\tilde{X}_i(\tilde{X}_i'\tilde{X}_i)^{-1}\tilde{X}_i'\sigma_\varepsilon^2 + \mathbf{l}_T'\mathbf{l}_T\sigma_\alpha^2)^+ = (\tilde{X}_i\frac{V_i}{T}\tilde{X}_i' + \mathbf{l}_T'\mathbf{l}_T\sigma_\alpha^2)^+X_i$, where $V_i = (\frac{1}{T}\tilde{X}_i'\tilde{X}_i)^{-1}\sigma_\varepsilon^2$ and since for a full column rank matrix $\tilde{X}_i$, $\tilde{X}_i\tilde{X}_i^+ = I_T$ and $\tilde{X}_i^+ = (\tilde{X}_i'\tilde{X}_i)^{-1}\tilde{X}_i'$. ∎

Proposition 6. *Denote $\hat{V}_\delta$ the clustered covariance matrix of $\hat{\delta}$ estimated by a GMM regression of Y on X with instrument Z. Let $\hat{V}_{\delta^{MD}}$ be the clustered covariance matrix of $\hat{\delta}^{MD}$ estimated by GMM regression of $\hat{Y}$ on X with instrument Z, where $\hat{Y}$ are estimated by an OLS first-stage. Let the clusters be at weakly higher level than i. Then, $\hat{V}_{\delta^{MD}} = \hat{V}_\delta$.*

We show that correct standard errors can be obtained using a two-stage approach by clustering the standard errors in the second stage at a level weakly higher than the individuals i. Let $g = 1, \dots, G$ index the clusters and assume that each of the clusters has N_g observations. This nests the case where one wishes to cluster at the individual level or at a higher level. For example, if i are county-year combinations, one might cluster at the county-level.
For an estimator $\hat{\delta}$ the clustered covariance matrix is estimated by

$$\hat{V}_\delta = \left(\frac{1}{G}\sum_{g=1}^{G} X_g' Z_g \hat{W} \frac{1}{G}\sum_{g=1}^{G} Z_g' X_g\right)^{-1} \frac{1}{G}\sum_{g=1}^{G} X_g' Z_g \hat{W} \left(\frac{1}{G}\sum_{g=1}^{G} Z_g' \tilde{u}_g \tilde{u}_g' Z_g\right)$$
$$\cdot \hat{W} \frac{1}{G}\sum_{g=1}^{G} Z_g' X_g \left(\frac{1}{G}\sum_{g=1}^{G} X_g' Z_g \hat{W} \frac{1}{G}\sum_{g=1}^{G} Z_g' X_g\right)^{-1},$$

where $\tilde{u}_g$ is a N_g-dimensional vector of estimated errors for the observations in cluster g.

Proof. Define $Z_g = (z_{1g}, \dots, z_{n_g g})'$, $X_g = (x_{1g}, \dots, x_{n_g g})'$, $Y_g = (y_{1g}, \dots, y_{n_g g})'$ and $\hat{Y}_g = (\hat{y}_{1g}, \dots, \hat{y}_{n_g g})'$. The first and third terms of the expression are identical for both estimators. Thus, we focus on the middle term. Let $\hat{u}_g = Y_g - X_g\hat{\delta}$ be the vector of residuals from the regression using Y as dependent variable, and let $\hat{u}_g^{MD} = \hat{Y}_g - X_g\hat{\delta}^{MD}$ be the vector of residuals of the estimator using the fitted values as regressand. We show that $Z_g'\hat{u}_g = Z_g'\hat{u}_g^{MD}$ for all g. By Proposition 3, $\hat{\delta}^{MD} = \hat{\delta}$. Thus, the fitted values of both estimators are identical. Next, define $\breve{X}_g = \operatorname{diag}\{\tilde{x}_{1g}, \dots, \tilde{x}_{n_g g}\}$ and recall that regressing Y_g on $\breve{X}_g$ is the same as performing G separate regressions. Let $\breve{\beta}_g$ be the coefficient vector of a OLS regression of Y_g on $\breve{X}_g$. Note that Z_g is in the column space of $\tilde{X}_g$. Define the projection matrix $\breve{P} = \breve{X}_g(\breve{X}_g'\breve{X}_g)^{-1}\breve{X}_g'$. Since Z_i is in the column space of $\breve{X}_g$,

$$\breve{P} Z_g = Z_g. \tag{33}$$

Then,

$$\begin{aligned}
Z_g'\hat{u}_g^{MD} &= Z_g'\left(\hat{Y}_g - X_g\hat{\delta}^{MD}\right) \\
&= Z_g'\tilde{X}_g\breve{\beta}_g - Z_g X_g\hat{\delta} \\
&= Z_g'\tilde{X}_g(\breve{X}_g'\breve{X}_g)^{-1}\breve{X}_g' Y_g - Z_g X_g\hat{\delta} \\
&= Z_g'(Y_g - X_g\hat{\delta}) = Z_g'\hat{u}_g,
\end{aligned}$$

where the fourth line follows by (33). Since this holds for all g, the desired result follows directly.

∎

B.2 Optimal Instruments and Minimum Distance

In this subsection, we show that if $\alpha_i(\tau) = 0$ for all i and τ, efficient minimum distance can be implemented by optimal instruments. From equation (32) we have that if $\alpha_i(\tau) = 0$ for all i and all τ, $\mathbb{E}[(\tilde{X}_i\hat{\beta}_i(\tau) - X_i\delta(\tau))(\tilde{X}_i\hat{\beta}_i(\tau) - X_i\delta(\tau))'|X_i] = \tilde{X}_i\frac{V_i(\tau)}{T}\tilde{X}_i'$. This implies the optimal instrument $Z_i^* = (\tilde{X}_i\frac{V_i(\tau)}{T}\tilde{X}_i')^+X_i$. Since T is a scalar, using $Z_i^*(\tau) = (\tilde{X}_iV_i(\tau)\tilde{X}_i')^+X_i$ leads to the same results.

Proposition 7. *The IV regression with instrument $Z_i^*(\tau) = (\tilde{X}_iV_i(\tau)\tilde{X}_i')^+X_i$ equals the efficient MD estimator.*

Proof.

$$\begin{aligned}
\hat{\delta}_{EMD}(\tau) &= \left(\sum_{i=1}^N R_i'\hat{V}_i^{-1}(\tau)R_i\right)^{-1}\left(\sum_{i=1}^N R_i'\hat{V}_i^{-1}(\tau)\hat{\beta}_i(\tau)\right)\\
&= \left(\sum_{i=1}^N X_i'\tilde{X}_i\left(\tilde{X}_i'\tilde{X}_i\hat{V}_i(\tau)\tilde{X}_i'\tilde{X}_i\right)^{-1}\tilde{X}_i'X_i\right)^{-1}\left(X_i'\tilde{X}_i\left(\tilde{X}_i'\tilde{X}_i\hat{V}_i(\tau)\tilde{X}_i'\tilde{X}_i\right)^{-1}\tilde{X}_i'\hat{Y}_i(\tau)\right)\\
&= \left(\sum_{i=1}^N X_i'\left(\tilde{X}_i\hat{V}_i(\tau)\tilde{X}_i'\right)^{+}X_i\right)^{-1}\left(X_i'\left(\tilde{X}_i\hat{V}_i(\tau)\tilde{X}_i'\right)^{+}\hat{Y}_i(\tau)\right) = \hat{\delta}_{OI}(\tau).
\end{aligned}$$

The second line follows by the relationship between $\tilde{X}_i$ and X_i, that is $\tilde{X}_iR_i = X_i$ and the third line follows since for a full column rank matrix $\tilde{X}_i$, $\tilde{X}_i^+ = (\tilde{X}_i'\tilde{X}_i)^{-1}\tilde{X}_i'$. ∎

B.3 Asymptotic Results

B.3.1 Proof of Theorems 1-3 and Lemmas 1-2

Proof of lemma 1. Starting from the definition of the estimator we obtain

$$\begin{aligned}
\hat{\delta}(\tau) &= \left(X'Z\hat{W}(\tau)Z'X\right)^{-1}X'Z\hat{W}(\tau)Z'\hat{y}(\tau)\\
&= \left(S_{ZX}'\hat{W}(\tau)S_{ZX}\right)^{-1}S_{ZX}'\hat{W}(\tau)\frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T z_{it}\tilde{x}_{it}'\hat{\beta}_i(\tau)\\
&= \left(S_{ZX}'\hat{W}(\tau)S_{ZX}\right)^{-1}S_{ZX}'\hat{W}(\tau)\frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T z_{it}\left(\tilde{x}_{it}'\left(\hat{\beta}_i(\tau)-\beta_i(\tau)\right)+\tilde{x}_{it}'\beta_i(\tau)\right)\\
&= \left(S_{ZX}'\hat{W}(\tau)S_{ZX}\right)^{-1}S_{ZX}'\hat{W}(\tau)\frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T z_{it}\left(\tilde{x}_{it}'\left(\hat{\beta}_i(\tau)-\beta_i(\tau)\right)+x_{it}'\delta(\tau)+\alpha_i(\tau)\right)\\
&= \delta(\tau)+\left(S_{ZX}'\hat{W}(\tau)S_{ZX}\right)^{-1}S_{ZX}'\hat{W}(\tau)\frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T z_{it}\left(\tilde{x}_{it}'\left(\hat{\beta}_i(\tau)-\beta_i(\tau)\right)+\alpha_i(\tau)\right).
\end{aligned}$$

∎

Proof of Theorem 1. We start from Lemma 1 and show that the last factor converges to zero while the other factors converge to finite values.

First, it follows from Assumptions 1(ii), 2(i) and 5(i) that $\text{Var}\left(\frac{1}{T}\sum_{t=1}^{T} z_{it}x'_{it}\right) = o_p\left(\frac{1}{T}\right)$ and $\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} z_{it}x'_{it}\right] = \mathbb{E}[z_{it}x'_{it}]$. Hence, by Assumption 1(i), $\text{Var}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T}\sum_{t=1}^{T} x_{it}z'_{it}\right) = o_p\left(\frac{1}{NT}\right)$. By Chebyshev's inequality,

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T} z_{it}x'_{it} - \mathbb{E}[z_{it}x'_{it}]\right) \underset{p}{\to} 0.$$

In addition, by Assumption 5(iii), $N^{-1}\sum_{i=1}^{N}\mathbb{E}[z_{it}x'_{it}] \to \Sigma_{ZX}$. It follows that

$$S_{ZX} \underset{p}{\to} \Sigma_{ZX}$$

Uniformly in $\tau \in \mathcal{T}$, $\hat{W}(\tau) \underset{p}{\to} W(\tau)$ where $W(\tau)$ is uniformly continuous. Together with the boundedness of Σ_{ZX} and the invertibility of $\Sigma'_{ZX}W\Sigma_{ZX}$, it follows that

$$\sup_{\tau\in\mathcal{T}}\left(S'_{ZX}\hat{W}(\tau)S_{ZX}\right)^{-1} S'_{ZX}\hat{W}(\tau) \underset{p}{\to} \left(\Sigma'_{ZX}W(\tau)\Sigma'_{ZX}\right)^{-1}\Sigma'_{ZX}W(\tau) \tag{34}$$

By lemma 3, $\hat{\beta}_i(\tau)$ is consistent for $\beta_i(\tau)$ uniformly in i and τ. Together with the boundedness of x_{it} in Assumption 2(i) and of z_{it} in Assumption 5(i), it follows that

$$\sup_{\tau\in\mathcal{T}}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} z_{it}\tilde{x}'_{it}(\hat{\beta}_i(\tau)-\beta_i(\tau)) \underset{p}{\to} 0. \tag{35}$$

By Assumption 5(ii), $\mathbb{E}[z_{it}\alpha_i(\tau)] = 0$ uniformly in τ. By Assumption 5, $\text{Var}(z_{it}\alpha_i(\tau))$ is uniformly bounded. In addition, z_{it} is bounded and $\alpha_i(\tau)$ is uniformly continuous in τ. Hence,

$$\sup_{\tau\in\mathcal{T}}\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} z_{it}\alpha_i(\tau) \underset{p}{\to} 0. \tag{36}$$

The result of the proposition follows from equations 34, 35, and 36. ■

Assumption 9 (Weighting matrix). *$\hat{W}(\tau) \underset{p}{\to} W(\tau)$ uniformly in $\tau \in \mathcal{T}$ such that*

$$W(\tau) = \begin{pmatrix} W_1(\tau)T & 0 \\ 0 & 0\end{pmatrix} + W_2(\tau) = \begin{pmatrix} W_1(\tau)T & 0 \\ 0 & 0\end{pmatrix} + \begin{pmatrix} W_{11}(\tau) & W_{12}(\tau) \\ W_{21}(\tau) & W_{22}(\tau)\end{pmatrix},$$

where the $L_1 \times L_1$ matrix $W_1(\tau)$ and the $L_2 \times L_2$ matrix $W_{22}(\tau)$ are symmetric and strictly positive-definite. For all $\tau_1, \tau_2 \in \mathcal{T}$, $||W(\tau_2) - W(\tau_1)|| \leq C|\tau_2 - \tau_1|$.

Theorem 1′ (Uniform consistency with an alternative weighting matrix). *Let the model in equation (1), Assumptions 1-7, Assumption 8(a) and Asummption 9 hold. Then,*

$$\sup_{\tau\in\mathcal{T}}\|\hat{\delta}(\tau)-\delta(\tau)\| = o_p(1)$$

Proof of Theorem 1′. The proof is similar to the proof of Theorem 1 except that $\Sigma'_{ZX} W \Sigma_{ZX}$ is not necessarily invertible. We partition the matrix Σ_{ZX} as follows

$$\Sigma_{ZX} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} = \begin{pmatrix} \Sigma_{11} & 0 \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

where Σ_{11} is $L_1 \times K_1$, Σ_{12} is $L_1 \times K_2$, Σ_{21} is $L_2 \times K_1$ and Σ_{22} is $L_2 \times K_2$. Note that $\Sigma_{12} = 0$:

$$\Sigma_{12} = \mathbb{E}\left[z_{1it} x'_{2it}\right] = \mathbb{E}_i\left[\mathbb{E}_t\left[z_{1it} x'_{2it}\right]\right] = \mathbb{E}_i\left[\mathbb{E}_t\left[z_{1it} x_{2i}\right]\right] = \mathbb{E}_i\left[\mathbb{E}_t\left[z_{1it}\right] x_{2i}\right] = 0$$

To simplify the notation we suppress the dependency of W on τ in the rest of the proof. Note that

$$\begin{aligned} \Sigma'_{ZX} W \Sigma_{ZX} &= \begin{pmatrix} \Sigma'_{11} & \Sigma'_{21} \\ 0 & \Sigma'_{22} \end{pmatrix} \begin{pmatrix} W_1 T + W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix} \begin{pmatrix} \Sigma_{11} & 0 \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \\ &:= \begin{pmatrix} A_{11} T + B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} \end{aligned}$$

where $A_{11} = \Sigma'_{11} W_1 \Sigma_{11}$, $B_{11} = \Sigma'_{11} W_{11} \Sigma_{11} + \Sigma'_{11} W_{12} \Sigma_{21} + \Sigma'_{21} W_{21} \Sigma_{11} + \Sigma'_{21} W_{22} \Sigma_{21}$, $B_{12} = \Sigma'_{11} W_{12} \Sigma 22 + \Sigma'_{21} W_{22} \Sigma_{22}$, $B_{21} = \Sigma_{22} W_{21} \Sigma_{11} + \Sigma_{22} W_{22} \Sigma_{21}$, and $B_{22} = \Sigma'_{22} W_{22} \Sigma_{22}$. Note that A_{11} and B_{22} are invertible by Assumptions 5 and 9. By the inverse of a partitioned matrix, we obtain

$$\left(\Sigma'_{ZX} W \Sigma_{ZX}\right)^{-1} = \begin{pmatrix} \left(A_{11} T + B_{11} - B_{12} B_{22}^{-1} B_{21}\right)^{-1} & -\left(A_{11} T + B_{11}\right)^{-1} B_{12} \left(B_{22} - B_{21}(A_{11} T + B_{11})^{-1} B_{12}\right)^{-1} \\ -B_{22}^{-1} B_{21} (A_{11} T + B_{11} - B_{12} B_{22}^{-1} B_{21})^{-1} & \left(B_{22} - B_{21}(A_{11} T + B_{11})^{-1} B_{12}\right)^{-1} \end{pmatrix}$$

Similarly,

$$\Sigma'_{ZX} W = \begin{pmatrix} \Sigma'_{11} W_1 T + \Sigma'_{11} W_{11} + \Sigma'_{21} W_{21} & \Sigma'_{11} W_{12} + \Sigma'_{21} W_{22} \\ \Sigma'_{22} W_{21} & \Sigma'_{22} W_{22} \end{pmatrix}$$

As $T \to \infty$, applying L'Hospital's rule for the first column, we obtain

$$\left(\Sigma'_{ZX} W \Sigma_{ZX}\right)^{-1} \Sigma'_{ZX} W \to \begin{pmatrix} A_{11}^{-1} \Sigma'_{11} W_1 & 0 \\ -B_{22}^{-1} B_{21} A_{11}^{-1} \Sigma'_{11} W_1 + B_{22}^{-1} \Sigma'_{22} W_{21} & B_{22}^{-1} \Sigma'_{22} W_{22} \end{pmatrix}$$

All the terms in this matrix are finite by the invertibility of A_{11} and B_{22}. The rest of the proofs follows as in the proof of Theorem 1. ∎

Lemma 3 (Uniform consistency of $\hat{\beta}_i(\tau)$)**.** *Under Assumptions 1-4 and 8(i), we have*

$$\sup_{\tau \in \mathcal{T}} \max_{1 \le i \le N} \|\hat{\beta}_i(\tau) - \beta_i\| = o_p(1).$$

Proof of Lemma 3. Angrist et al. (2006) show uniform consistency in τ but not in i of the quantile regression estimator (see their Theorem 3) while Galvao and Wang (2015) show uniform consistency in i but not in τ (see their Lemma 1). We show here uniformity in both dimensions by following the steps of the proof in Galvao and Wang (2015) and extending it. We define $\mathbb{Q}_{iT}(\tau, \beta) :=$

$\frac{1}{T}\sum_{t=1}^{T}\rho_\tau(y_{it}-\tilde{x}_{it}'\beta)-\rho_\tau(y_{it}-\tilde{x}_{it}'\beta_i(\tau))$ and $Q_i(\tau,\beta) := E[\rho_\tau(y_{it}-\tilde{x}_{it}'\beta)-\rho_\tau(y_{it}-\tilde{x}_{it}'\beta_i(\tau))]$. Angrist et al. (2006) show that the empirical process for each individual i is stochastically equicontinuous because $|\mathbb{Q}_{iT}(\tau',\beta')-\mathbb{Q}_{iT}(\tau'',\beta'')| \le C_1\cdot|\tau'-\tau''|+C_2\cdot\|\beta'-\beta''\|$ where $C_1 = 2\cdot C\cdot\sup_{\beta\in\mathcal{B}}\|\beta\|$ and $C_2 = 2\cdot C$. Note that C_1 and C_2 are not functions of either i or τ.

Fix any $\delta > 0$. Let $B_i(\delta,\tau) := \{\beta : \|\beta-\beta_i(\tau)\| \le \delta\}$, the ball with center $\beta_i(\tau)$ and radius δ. For each $\beta \notin B_i(\delta,\tau)$, define $\tilde{\beta} = r_i\beta + (1-r_i)\beta_i(\tau)$ where $r_i = \frac{\delta}{\|\beta-\beta_i(\tau)\|}$. So $\tilde{\beta} \in \partial B_i(\delta,\tau) := \{\beta : \|\beta-\beta_i(\tau)\| = \delta\}$, the boundary of $B_i(\delta,\tau)$. Since $\mathbb{Q}_{iT}(\beta,\tau)$ is convex in β for all τ, and $\mathbb{Q}_{it}(\beta_i(\tau),\tau) = 0$, we have

$$r_i\mathbb{Q}_{iT}(\beta,\tau) \ge \mathbb{Q}_{iT}(\tilde{\beta},\tau) = Q_i(\tilde{\beta},\tau) + \mathbb{Q}_{iT}(\tilde{\beta},\tau) - Q_i(\tilde{\beta},\tau) > \epsilon_\delta + \mathbb{Q}_{iT}(\tilde{\beta},\tau) - Q_i(\tilde{\beta},\tau) \quad (37)$$

uniformly in i and τ, where

$$\epsilon_\delta := \inf_{\tau\in\mathcal{T}}\inf_{1\le i\le N}\inf_{\|\beta-\beta_i(\tau)\|=\delta}\mathbb{E}\left[\int_0^{\tilde{x}_{it}'(\beta-\beta_i(\tau))}[1(y_{it}-\tilde{x}_{it}'\beta_i(\tau)\le s) - 1(y_{it}-\tilde{x}_{it}'\beta_i(\tau)\le 0)]ds\right]$$

by the identity of Knight (1998) and $\epsilon_\delta > 0$ by Assumptions 3 and 4.

Thus, we have the following

$$\begin{aligned}\left\{\sup_{\tau\in\mathcal{T}}\max_{1\le i\le N}\|\hat{\beta}_i(\tau)-\beta_i(\tau)\| > \delta\right\} &\overset{(a)}{\subseteq} \{\exists\tau_i\in\mathcal{T}, \exists\beta_i\notin B_i(\delta,\tau_i) : \mathbb{Q}_{iT}(\beta_i,\tau_i)\le 0\}\\ &\overset{(b)}{\subseteq} \cup_{i=1}^{N}\left\{\sup_{\tau\in\mathcal{T}}\sup_{\beta_i\in B_i(\delta,\tau_i)}|\mathbb{Q}_{iT}(\beta_i,\tau_i)-Q_i(\beta_i,\tau_i)| \ge \epsilon_\delta\right\}\end{aligned}$$

Relation (a) holds because, by definition, $\hat{\beta}_i(\tau)$ minimizes $\mathbb{Q}_{iT}(\beta,\tau)$, and $\mathbb{Q}_{iT}(\beta_i(\tau),\tau) = 0$. Relation (b) holds by the rightmost inequality of line (37). Then, it follows that

$$\begin{aligned}\mathrm{P}\left\{\sup_{\tau\in\mathcal{T}}\max_{1\le i\le N}\|\hat{\beta}_i(\tau)-\beta_i(\tau)\| > \delta\right\} &\le \mathrm{P}\left\{\cup_{i=1}^{N}\left\{\sup_{\tau\in\mathcal{T}}\sup_{\beta_i\in B_i(\delta,\tau)}|\mathbb{Q}_{iT}(\beta_i,\tau)-Q_i(\beta_i,\tau)| \ge \epsilon_\delta\right\}\right\}\\ &\le \sum_{i=1}^{N}\mathrm{P}\left\{\sup_{\tau\in\mathcal{T}}\sup_{\beta_i\in B_i(\delta\tau)}|\mathbb{Q}_{iT}(\beta_i,\tau)-Q_i(\beta_i,\tau)| \ge \epsilon_\delta\right\}\\ &\le N\max_{1\le i\le N}\mathrm{P}\left\{\sup_{\tau\in\mathcal{T}}\sup_{\beta_i\in B_i(\delta,\tau)}|\mathbb{Q}_{iT}(\beta_i,\tau)-Q_i(\beta_i,\tau)| \ge \epsilon_\delta\right\}\end{aligned}$$

Therefore, if we can show that

$$\max_{1\le i\le N}\mathrm{P}\left\{\sup_{\tau\in\mathcal{T}}\sup_{\beta_i\in B_i(\delta,\tau)}|\mathbb{Q}_{iT}(\beta_i,\tau)-Q_i(\beta_i,\tau)| \ge \epsilon_\delta\right\} = o\left(\frac{1}{N}\right)$$

the proof of the lemma will be completed.

Without loss of generality, we assume $\beta_i(\tau) = 0$ for all i and $\tau\in\mathcal{T}$. Then the balls $B_i(\delta,\tau)$ for all i and $\tau\in\mathcal{T}$ are identical and we denote them by $B(\delta)$. Because the closed ball $B(\delta)$ is compact, there exist K balls with center β^j, $j = 1,...,K$, and radius $\frac{\epsilon}{3C_2}$ such that the collection of them covers $B(\delta)$. For any $\epsilon > 0$, we can find a finite K that satisfies this condition and is independent of i and τ. Therefore, for any $\beta \in B(\delta)$, there is some $j \in \{1, ..., K\}$ such that

$$\begin{aligned}
|\mathbb{Q}_{iT}(\beta,\tau) - Q_i(\beta,\tau)| - |\mathbb{Q}_{iT}(\beta^j,\tau) - Q_i(\beta^j,\tau)| &\leq |\mathbb{Q}_{iT}(\beta,\tau) - Q_i(\beta,\tau) - \mathbb{Q}_{iT}(\beta^j,\tau) + Q_i(\beta^j,\tau)| \\
&\leq |\mathbb{Q}_{iT}(\beta,\tau) - \mathbb{Q}_{iT}(\beta^j,\tau)| + |Q_i(\beta,\tau) - Q_i(\beta^j,\tau)| \\
&\leq C_2 \frac{\epsilon}{3C_2} + C_2 \frac{\epsilon}{3C_2} = \frac{2\epsilon}{3}
\end{aligned}$$

uniformly in i and $\tau \in \mathcal{T}$. The third line is justified by the stochastic equicontinuity of $\mathbb{Q}_{iT}(\beta,\tau)$.

It then follows that, for any $\epsilon > 0$,

$$\sup_{\tau \in \mathcal{T}} \sup_{\beta \in B(\delta)} |\mathbb{Q}_{iT}(\beta,\tau) - Q_i(\beta,\tau)| \leq \sup_{\tau \in \mathcal{T}} \max_{1 \leq j \leq K} |\mathbb{Q}_{iT}(\beta^j,\tau) - Q_i(\beta^j,\tau)| + \frac{2\epsilon}{3}$$

and

$$\begin{aligned}
\mathrm{P}\left\{\sup_{\tau \in \mathcal{T}} \sup_{\beta \in B(\delta)} |\mathbb{Q}_{iT}(\beta,\tau) - Q_i(\beta,\tau) > \epsilon\right\} &\leq \mathrm{P}\left\{\sup_{\tau \in \mathcal{T}} \max_{1 \leq j \leq K} |\mathbb{Q}_{iT}(\beta^j,\tau) - Q_i(\beta^j,\tau)| + \frac{2\epsilon}{3} > \epsilon\right\} \\
&= \mathrm{P}\left\{\sup_{\tau \in \mathcal{T}} \max_{1 \leq j \leq K} |\mathbb{Q}_{iT}(\beta^j,\tau) - Q_i(\beta^j,\tau)| > \frac{\epsilon}{3}\right\} \\
&\leq \sup_{\tau \in \mathcal{T}} \sum_{j=1}^{K} \mathrm{P}\left\{|\mathbb{Q}_{iT}(\beta^j,\tau) - Q_i(\beta^j,\tau)| > \frac{\epsilon}{3}\right\}
\end{aligned}$$

For each $\tau \in \mathcal{T}$, $\mathbb{Q}_{iT}(\beta^j,\tau)$ is the sample mean of T i.i.d. terms bounded in absolute values by $2 \cdot C \cdot \delta$. By Hoeffding's inequality, it follows that

$$\begin{aligned}
\sum_{j=1}^{K} \mathrm{P}\left\{|\mathbb{Q}_{iT}(\beta^j,\tau) - Q_i(\beta^j,\tau)| > \frac{\epsilon}{3}\right\} &\leq 2K \exp\left\{-\frac{2T\epsilon^2}{3^2 2^2 C^2 \delta^2}\right\} \\
&= 2K \exp\left\{-\frac{T\epsilon^2}{18 C^2 \delta^2}\right\} \\
&= O(\exp(-T))
\end{aligned}$$

This upper bound is deterministic and not a function of τ such that it also applies to the supremum over τ. Since $\frac{\log N}{T} \to 0$ by Assumption 8(a), it follows that $O(\exp(-T)) = o(1/N)$. ∎

Proof of Lemma 2. **Part (i)**

Lemma 3 in Galvao et al. (2020) provides the uniform Bahadur representation for the individual-level quantile regression coefficient under our assumptions:

$$\hat{\beta}_i(\tau) - \beta_i(\tau) = \frac{1}{T} \sum_{t=1}^{T} \phi_{i,\tau}(\tilde{x}_{it}, y_{it}) + R_{iT}^{(1)}(\tau) + R_{iT}^{(2)}(\tau), \tag{38}$$

where

$$\phi_{i,\tau}(\tilde{x}_{it}, y_{it}) = -B_{i,\tau}^{-1} \tilde{x}_{it}(1(y_{it} \leq \tilde{x}_{it}\beta_i(\tau)) - \tau) \tag{39}$$

with $B_{i,\tau} = \mathbb{E}[f_{y|x}(Q_{y|x}(\tau|\tilde{x}_{it})\tilde{x}_{it}\tilde{x}_{it}']$ and

$$\sup_i \sup_{\tau\in\mathcal{T}} \left\| R_{iT}^{(2)}(\tau) \right\| = O_p\left(\frac{\log T}{T}\right) \tag{40}$$

$$\sup_i \sup_{\tau\in\mathcal{T}} \left\| \mathbb{E}\left[R_{iT}^{(1)}(\tau)\right] \right\| = O\left(\frac{\log T}{T}\right) \tag{41}$$

$$\sup_i \sup_{\tau\in\mathcal{T}} \left\| \mathbb{E}\left[\left(R_{iT}^{(1)}(\tau) - \mathbb{E}[R_{iT}^{(1)}(\tau)]\right)\left(R_{iT}^{(1)}(\tau) - \mathbb{E}[R_{iT}^{(1)}(\tau)]\right)'\right] \right\| = O\left(\left(\frac{\log T}{T}\right)^{3/2}\right) \tag{42}$$

It follows that

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} z_{it}\tilde{x}_{it}'\left(\hat{\beta}_i(\tau) - \beta_i(\tau)\right) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T} z_{it}\tilde{x}_{it}'\right)\left(\frac{1}{T}\sum_{t=1}^{T}\phi_{i,\tau}(\tilde{x}_{it}, y_{it})\right) \tag{43}$$

$$+\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T} z_{it}\tilde{x}_{it}'\right) R_{iT}^{(1)}(\tau) \tag{44}$$

$$+\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T} z_{it}\tilde{x}_{it}'\right) R_{iT}^{(2)}(\tau) \tag{45}$$

Consider first the third term (45). By assumptions 2(i) and 5(i), x_{it} and z_{it} are bounded by C such that the sample mean of their product is also bounded. Therefore, (40) implies that

$$\sup_{\tau\in\mathcal{T}} \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T} z_{it}\tilde{x}_{it}'\right) R_{iT}^{(2)}(\tau) = O_p\left(\frac{\log T}{T}\right) \tag{46}$$

Consider now the second term of (43). Since $\mathrm{Var}\left(R_{iT}^{(1)}(\tau)\right) = o\left(\frac{1}{T}\right)$ by (42), x_{it} and z_{it} are bounded by assumptions 2(i) and 5(i), and observations are independent across individuals, it follows that $\mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T} z_{it}\tilde{x}_{it}'\right) R_{iT}^{(1)}(\tau)\right) = o_p\left(\frac{1}{NT}\right)$. In addition, by (41), $\sup_i \sup_{\tau\in\mathcal{T}} \mathbb{E}\left[R_{iT}^{(1)}\right] = O\left(\frac{\log T}{T}\right)$ such that $\sup_{\tau\in\mathcal{T}} \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T}\sum_{i=1}^{T} z_{it}\tilde{x}_{it}'\right) R_{iT}^{(1)}(\tau)\right] = O\left(\frac{\log T}{T}\right)$. Putting this together, by the Chebyshev inequality and under Assumption 8(c),

$$\sup_{\tau\in\mathcal{T}} \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T} z_{it}\tilde{x}_{it}'\right) R_{iT}^{(1)}(\tau) = o_p\left(\frac{1}{\sqrt{NT}}\right) \tag{47}$$

It follows that both remainder terms are $o_p\left(\frac{1}{\sqrt{NT}}\right)$ uniformly over τ.

Consider now the term (43). Let $\Sigma_{ZXi} = \mathbb{E}_t[z_{it}\tilde{x}_{it}']$, i.e. Σ_{ZXi} is the expected value of $z_{it}\tilde{x}_{it}'$ for individual i.

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T}\sum_{i=1}^{T} z_{it}\tilde{x}_{it}'\right)\left(\frac{1}{T}\sum_{t=1}^{T}\phi_{i,\tau}(\tilde{x}_{it}, y_{it})\right) =$$

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T} z_{it}\tilde{x}_{it}' - \Sigma_{ZXi}\right)\left(\frac{1}{T}\sum_{t=1}^{T}\phi_{i,\tau}(\tilde{x}_{it}, y_{it})\right) + \frac{1}{N}\sum_{i=1}^{N}\Sigma_{ZXi}\left(\frac{1}{T}\sum_{t=1}^{T}\phi_{i,\tau}(\tilde{x}_{it}, y_{it})\right) \tag{48}$$

By the boundedness of z_{it} and x_{it} and the independence of the observations over time, it follows that $\left\|\frac{1}{T}\sum_{i=1}^{T} z_{it}\tilde{x}_{it}' - \Sigma_{ZXi}\right\| = o(1)$ uniformly in i. In addition, $\text{Var}\left(\frac{1}{T}\sum_{t=1}^{T}\phi_{i,\tau}(\tilde{x}_{it}, y_{it})\right) = O\left(\frac{1}{T}\right)$. Hence,

$$\text{Var}\left(\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T} z_{it}\tilde{x}_{it}' - \Sigma_{ZXi}\right)\left(\frac{1}{T}\sum_{t=1}^{T}\phi_{i,\tau}(\tilde{x}_{it}, y_{it})\right)\right) = o\left(\frac{1}{NT}\right)$$

The model in equation (1) and Assumption 5(iv) imply that $\mathbb{E}\left[1(y_{it} \le \tilde{x}_{it}\beta_i(\tau))|\tilde{x}_{it}, z_{it}, v_i\right] = \tau$, which implies that

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T} z_{it}\tilde{x}_{it}' - \Sigma_{ZXi}\right)\left(\frac{1}{T}\sum_{t=1}^{T}\phi_{i,\tau}(\tilde{x}_{it}, y_{it})\right)\right] = 0$$

uniformly in τ. Therefore, by Chebyshev inequality,

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T} z_{it}\tilde{x}_{it}' - \Sigma_{ZXi}\right)\left(\frac{1}{T}\sum_{t=1}^{T}\phi_{i,\tau}(\tilde{x}_{it}, y_{it})\right) = o_p\left(\frac{1}{\sqrt{NT}}\right) \tag{49}$$

uniformly in τ.

Since all other terms are $o_p\left(\frac{1}{\sqrt{NT}}\right)$ uniformly over τ, the limiting distribution of the process $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} z_{it}\tilde{x}_{it}'\left(\hat{\beta}_i(\tau) - \beta_i(\tau)\right)$ is the same as the limiting distribution of

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^{N}\Sigma_{ZXi}\left(\frac{1}{T}\sum_{t=1}^{T}\phi_{i,\tau}(\tilde{x}_{it}, y_{it})\right)\\ &\qquad = \frac{1}{N}\sum_{i=1}^{N}\Sigma_{ZXi}\left(\frac{-B_{i,\tau}^{-1}}{T}\sum_{t=1}^{T}\tilde{x}_{it}(1(y_{it} \le \tilde{x}_{it}\beta_i(\tau)) - \tau)\right) := \frac{1}{N\sqrt{T}}\sum_{i=1}^{N} s_i(\tau)\end{aligned} \tag{50}$$

At a given quantile τ, this is a sample mean over N independent (but not necessarily identically distributed) observations denoted by $s_i(\tau)$. The model in equation (1) and Assumption 5(iv) imply that $\mathbb{E}\left[1(y_{it} \le \tilde{x}_{it}\beta_i(\tau))|\tilde{x}_{it}, z_{it}, v_i\right] = \tau$, which implies that $\mathbb{E}[s_i(\tau)] = 0$ uniformly in i and τ. In addition,

$$\text{Var}(s_i(\tau)) = \mathbb{E}[\Sigma_{ZXi}\,\text{Var}(\phi_{i,\tau})\Sigma_{ZXi}'] = \mathbb{E}[\Sigma_{ZXi}B_{i,\tau}^{-1}\tau(1-\tau)\mathbb{E}[x_{it}x_{it}'|v_i]B_{i,\tau}^{-1}\Sigma_{ZXi}'] \tag{51}$$

Pointwise asymptotic normality follows by an application of the Lindeberg CLT.

Next we note that $\left\{\Sigma_{ZXi}\left(\frac{-B_{i,\tau}^{-1}}{T}\sum_{t=1}^{T}\tilde{x}_{it}(1(y_{it} \le \tilde{x}_{it}\beta) - \tau)\right), \tau \in \mathcal{T}, \beta \in \mathcal{B}\right\}$ is a Donsker class for any compact set $\mathcal{B}$. This follows by noting that $\{1(y_{it} \le \tilde{x}_{it}\beta_i(\tau)), \tau \in \mathcal{T}, \beta \in \mathcal{B}\}$ is a VC subgraph class and hence a bounded Donsker class. Hence,

$$\left\{\frac{1}{T}\sum_{t=1}^{T}\tilde{x}_{it}(1(y_{it} \le \tilde{x}_{it}\beta) - \tau), \tau \in \mathcal{T}, \beta \in \mathcal{B}\right\}$$

is also bounded Donsker with a square-integrable envelope $2 \cdot \max_{t \in 1,\ldots,T}|\tilde{x}_{it}| \le 2 \cdot C$. The whole function is then Donsker by the boundedness of Σ_{ZXi} and B_i^{-1}. The weak convergence result follows by application of the functional central limit theorem for independent but not identically distributed random variables, see for instance Theorem 3 in Brown (1971).

Part (ii) follows directly by Lemma 3 in Chetverikov et al. (2016).

Part (iii) The first moment is asymptotically equivalent to (up to a term, which is uniformly $o_p(\frac{1}{NT})$)

$$\frac{1}{N}\sum_{i=1}^{N}\Sigma_{ZXi}\left(\frac{-B_{i,\tau}^{-1}}{T}\sum_{t=1}^{T}\tilde{x}_{it}(1(y_{it}\leq\tilde{x}_{it}\beta_i(\tau))-\tau)\right).$$

We have already shown that both moments have mean zero. By Assumption 1, the observations are independent across i and t such that we only need to consider the correlation between both moments for the same individual and time period.

$$\begin{aligned}
&\mathrm{Cov}(\tilde{x}_{it}(1(y_{it}\leq\tilde{x}_{it}\beta_i(\tau))-\tau), z_{it}\alpha_i(\tau'))\\
&\quad=\mathbb{E}[\tilde{x}_{it}(1(y_{it}\leq\tilde{x}_{it}\beta_i(\tau))-\tau)z_{it}'\alpha_i(\tau')]\\
&\quad=\mathbb{E}[\mathbb{E}[-\Sigma_{ZXi}B_{i,\tau}^{-1}\tilde{x}_{it}(1(y_{it}\leq\tilde{x}_{it}\beta_i(\tau))-\tau)z_{it}'\alpha_i(\tau')|x_{it},z_{it},v_i]]\\
&\quad=\mathbb{E}[\tilde{x}_{it}\mathbb{E}[(1(y_{it}\leq\tilde{x}_{it}\beta_i(\tau))-\tau)|x_{it},z_{it},v_i]z_{it}'\alpha_i(\tau')]=0
\end{aligned}$$

It follows that

$$\sup_{\tau,\tau'\in\mathcal{T}}\|\mathrm{Cov}(\bar{g}_{NT}^{(1))}(\delta,\tau),\bar{g}_{NT}^{(1))}(\delta,\tau'))\|=o_p\left(\frac{1}{NT}\right)$$

■

Proof of Theorem 2. From the definition of the estimator,

$$\hat{\delta}(\tau)-\delta(\tau)=\left(S_{ZX}'\hat{W}S_{ZX}\right)^{-1}S_{ZX}'\hat{W}\bar{g}_{NT}(\delta,\tau)$$

As shown in the proof of Theorem 1, $S_{ZX}\to\Sigma_{ZX}$. We can partition the matrix Σ_{ZX} as follows

$$\Sigma_{ZX}=\begin{pmatrix}\Sigma_{11}&\Sigma_{12}\\\Sigma_{21}&\Sigma_{22}\end{pmatrix}=\begin{pmatrix}\Sigma_{11}&0\\\Sigma_{21}&\Sigma_{22}\end{pmatrix}$$

where Σ_{11} is $L_1\times K_1$, Σ_{12} is $L_1\times K_2$, Σ_{21} is $L_2\times K_1$ and Σ_{22} is $L_2\times K_2$. Note that $\Sigma_{12}=0$:

$$\Sigma_{12}=\mathbb{E}\left[z_{1it}x_{2it}'\right]=\mathbb{E}_i\left[\mathbb{E}_t\left[z_{1it}x_{2it}'\right]\right]=\mathbb{E}_i\left[\mathbb{E}_t\left[z_{1it}x_{2i}\right]\right]=\mathbb{E}_i\left[\mathbb{E}_t\left[z_{1it}\right]x_{2i}\right]=0$$

To simplify the notation we suppress the dependency of W on τ in the rest of the proof. Note that

$$\begin{aligned}
\Sigma_{ZX}'W\Sigma_{ZX}&=\begin{pmatrix}\Sigma_{11}'&\Sigma_{21}'\\0&\Sigma_{22}'\end{pmatrix}\begin{pmatrix}W_1&0\\0&W_{22}\end{pmatrix}\begin{pmatrix}\Sigma_{11}&0\\\Sigma_{21}&\Sigma_{22}\end{pmatrix}\\
&:=\begin{pmatrix}\Sigma_{11}'W_1\Sigma_{11}&B_{12}\\B_{21}&B_{22}\end{pmatrix}
\end{aligned}$$

where $A_{11}=\Sigma_{11}'W_1\Sigma_{11}$, $B_{11}=\Sigma_{11}'W_{11}\Sigma_{11}+\Sigma_{11}'W_{12}\Sigma_{21}+\Sigma_{21}'W_{21}\Sigma_{11}+\Sigma_{21}'W_{22}\Sigma_{21}$, $B_{12}=\Sigma_{11}'W_{12}\Sigma22+\Sigma_{21}'W_{22}\Sigma_{22}$, $B_{21}=\Sigma_{22}W_{21}\Sigma_{11}+\Sigma_{22}W_{22}\Sigma_{21}$, and $B_{22}=\Sigma_{22}'W_{22}\Sigma_{22}$. Note that

A_{11} and B_{22} are invertible by Assumptions 5 and 9. By the inverse of a partitioned matrix, we obtain

$$\left(\Sigma'_{ZX}W\Sigma_{ZX}\right)^{-1} =$$
$$\begin{pmatrix} \left(A_{11}T + B_{11} - B_{12}B_{22}^{-1}B_{21}\right)^{-1} & -\left(A_{11}T + B_{11}\right)^{-1}B_{12}\left(B_{22} - B_{21}(A_{11}T + B_{11})^{-1}B_{12}\right)^{-1} \\ -B_{22}^{-1}B_{21}(A_{11}T + B_{11} - B_{12}B_{22}^{-1}B_{21})^{-1} & \left(B_{22} - B_{21}(A_{11}T + B_{11})^{-1}B_{12}\right)^{-1} \end{pmatrix}$$

Similarly,

$$\Sigma'_{ZX}W = \begin{pmatrix} \Sigma'_{11}W_1T + \Sigma'_{11}W_{11} + \Sigma'_{21}W_{21} & \Sigma'_{11}W_{12} + \Sigma'_{21}W_{22} \\ \Sigma'_{22}W_{21} & \Sigma'_{22}W_{22} \end{pmatrix}$$

As $T \to \infty$, applying L'Hospital's rule for the first column, we obtain

$$\left(\Sigma'_{ZX}W\Sigma_{ZX}\right)^{-1}\Sigma'_{ZX}W \to \begin{pmatrix} A_{11}^{-1}\Sigma'_{11}W_1 & 0 \\ -B_{22}^{-1}B_{21}A_{11}^{-1}\Sigma'_{11}W_1 + B_{22}^{-1}\Sigma'_{22}W_{21} & B_{22}^{-1}\Sigma'_{22}W_{22} \end{pmatrix}$$

All the terms in this matrix are finite by the invertibility of A_{11} and B_{22}.

Next, we partition the matrix S_{ZX} to separate the z_{1it} from the z_{2it} components as well as the x_{1it} from the x_{2it} components:

$$S_{ZX} = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} = \begin{pmatrix} S_{11} & 0 \\ S_{21} & S_{22}, \end{pmatrix}$$

where S_{11} is $L_1 \times K_1$, S_{12} is $L_1 \times K_2$, S_{21} is $L_2 \times K_1$ and S_{22} is $L_2 \times K_2$. Note that $S_{12} = 0$:

$$S_{12} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} x_{2it}z_{1it} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} x_{2i}z_{1it} = \frac{1}{NT}\sum_{i=1}^{N} x_{2i}\sum_{t=1}^{T} z_{1it} = \frac{1}{N}\sum_{i=1}^{N} x_{2i}\bar{z}_{1i} = 0.$$

This means that the fast moments (time-varying-instruments) cannot identify the coefficients on the time-constant covariates.

It follows that

$$\begin{aligned} \Lambda_T^{-1}S'_{ZX}\hat{W}S_{ZX}\Lambda_T^{-1} &= \begin{pmatrix} S'_{11}/\sqrt{T} & S'_{21}/\sqrt{T} \\ 0 & S'_{22} \end{pmatrix}\left(\begin{pmatrix} \hat{W}_1T & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} \hat{W}_{11} & \hat{W}_{12} \\ \hat{W}_{21} & \hat{W}_{22} \end{pmatrix}\right)\begin{pmatrix} S_{11}/\sqrt{T} & 0 \\ S_{21}/\sqrt{T} & S_{22} \end{pmatrix} \\ &= \begin{pmatrix} S'_{11}\hat{W}_1S_{11} & 0 \\ 0 & S'_{22}\hat{W}_{22}S_{22} \end{pmatrix} + o_p(1). \end{aligned}$$

and

$$\begin{aligned} \Lambda_T^{-1}S'_{ZX}\hat{W}\Lambda_T^{-1} &= \begin{pmatrix} S'_{11}/\sqrt{T} & S'_{21}/\sqrt{T} \\ 0 & S'_{22} \end{pmatrix}\left(\begin{pmatrix} \hat{W}_1T & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} \hat{W}_{11} & \hat{W}_{12} \\ \hat{W}_{21} & \hat{W}_{22} \end{pmatrix}\right)\begin{pmatrix} I_{L_1}/\sqrt{T} & 0 \\ 0 & I_{L_2} \end{pmatrix} \\ &= \begin{pmatrix} S'_{11}\hat{W}_1 & 0 \\ 0 & S'_{22}\hat{W}_{22} \end{pmatrix} + o_p(1). \end{aligned}$$

Using the results derived in the proof of Proposition 1, we obtain

$$\begin{aligned} &\left(\Lambda_T^{-1}S'_{ZX}\hat{W}S_{ZX}\Lambda_T^{-1}\right)^{-1}\Lambda_T^{-1}S'_{ZX}\hat{W}\Lambda_T^{-1} \\ &\qquad \underset{p}{\to} \begin{pmatrix} \left(\Sigma'_{11}W_1\Sigma_{11}\right)^{-1}\Sigma'_{11}W_1 & 0 \\ 0 & \left(\Sigma'_{22}W_{22}\Sigma_{22}\right)^{-1}\Sigma'_{22}W_{22} \end{pmatrix} = G. \end{aligned} \tag{52}$$

Combining this result and Lemma 2 with Slutzky's lemma, we obtain

$$\Lambda_{NT}(\hat{\delta}(\tau) - \delta(\tau)) \underset{d}{\to} N(0, G\Omega G').$$

■

B.3.2 Covariance Matrix

Proof of Proposition 1.

$$\widehat{\text{Var}}(\sqrt{N}(\hat{\delta}-\delta)) = \left(S'_{ZX}\hat{W}S_{ZX}\right)^{-1} S'_{ZX}\hat{W}\left(\frac{1}{NT^2}\sum_{i=1}^{N} Z_i'\hat{u}_i\hat{u}_i'Z_i\right)\hat{W}S_{ZX}\left(S'_{ZX}\hat{W}S_{ZX}\right)^{-1}$$

By theorem 3 it follows that

$$\left(S'_{ZX}\hat{W}(\tau)S_{ZX}\right)^{-1} S'_{ZX}\hat{W}(\tau) = G(\cdot) + o_p(1).$$

For the term in the middle, we have that

$$\begin{aligned}
\frac{1}{NT^2}\sum_{i=1}^{N} Z_i'\hat{u}_i\hat{u}_i'Z_i =& \frac{1}{NT^2}\sum_{i=1}^{N}\left(Z_i'\left(\tilde{X}_i(\hat{\beta}_i-\beta_i)+X_i(\delta-\hat{\delta})+\alpha_i\right)\cdot\left(\tilde{X}_i(\hat{\beta}_i-\beta_i)+X_i(\delta-\hat{\delta})+\alpha_i\right)'Z_i\right)\\
=&\frac{1}{NT^2}\sum_{i=1}^{N}\Big(Z_i'\Big(\tilde{X}_i(\hat{\beta}_i-\beta_i)(\hat{\beta}_i-\beta_i)'\tilde{X}_i'+\alpha_i\alpha_i'+X_i(\hat{\delta}-\delta)(\hat{\delta}-\delta)'X_i'\\
&+X_i(\delta-\hat{\delta})(\hat{\beta}_i-\beta_i)'\tilde{X}_i'+\tilde{X}_i(\hat{\beta}_i-\beta_i)(\delta-\hat{\delta})'X_i'\\
&+\alpha_i(\hat{\beta}_i-\beta_i)'\tilde{X}_i'+\tilde{X}_i(\hat{\beta}_i-\beta_i)\alpha_i'+\alpha_i(\delta-\hat{\delta})'X_i'+X_i(\delta-\hat{\delta})\alpha_i'\Big)Z_i\Big).
\end{aligned}$$

Next, we want to show that all but the first two terms converge to zero quickly. We consider each term separately. Let $\zeta_{NT}(\tau) = \frac{1}{\sqrt{NT}} + \frac{1}{\sqrt{N}}\cdot||V_\alpha(\tau)||^{1/2}$, where $V_\alpha(\tau) = G\Omega_2G'$ and note that $(\delta-\hat{\delta}) = O_p\left(\zeta_{NT}(\tau)\right)$ and $(\hat{\beta}_i-\beta_i) = O_P\left(\frac{1}{T^{1/2}}\right)$.

For the first term, note that by the proof of Lemma 2(I), it follows that

$$\begin{aligned}
\frac{1}{NT^2}\sum_{i=1}^{N} Z_i'\tilde{X}_i(\hat{\beta}_i-\beta_i)(\hat{\beta}_i-\beta_i)'\tilde{X}_i' =& \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T} z_{it}\tilde{x}_{it}'(\hat{\beta}_i-\beta_i)\right)\left(\frac{1}{T}\sum_{t=1}^{T} z_{it}\tilde{x}_{it}'(\hat{\beta}_i-\beta_i)\right)'\\
=&\mathbb{E}\left[\left(\Sigma_{ZXi}\frac{1}{T}\sum_{t=1}^{T}\phi_{i,\tau}(\tilde{x}_{it},z_{it})\right)\left(\Sigma_{ZXi}\frac{1}{T}\sum_{t=1}^{T}\phi_{i,\tau}(\tilde{x}_{it},z_{it})\right)'\right]\\
&+o_p\left((NT)^{-1}\right)\\
=&\frac{\Omega_1}{T}+o_p\left((NT)^{-1}\right),
\end{aligned}$$

For the second term

$$\frac{1}{N}\sum_{i=1}^{N}\bar{z}_i\bar{z}_i\alpha_i^2 = \text{Var}(\bar{z}_i\alpha_i) + O_p\left(\frac{\text{Var}(\bar{z}_i\alpha_i)}{\sqrt{N}}\right) = \Omega_2 + O_p\left(\frac{1}{\sqrt{N}}\right)\cdot||V_\alpha||,$$

For the third term, we have that

$$\begin{aligned}
\frac{1}{NT^2}\sum_{i=1}^{N} Z_i'X_i(\hat{\delta}-\delta)(\hat{\delta}-\delta)'X_i'Z_i =& \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T} z_{it}x_{it}'\right)(\hat{\delta}-\delta)(\hat{\delta}-\delta)'\left(\frac{1}{T}\sum_{t=1}^{T} x_{it}z_{it}'\right)\\
=&O_p(\zeta_{NT}^2).
\end{aligned}$$

The fifth term is just the transpose of the fourth term. Thus we need to look at only one. we will consider only the fourth one.

$$\begin{aligned}\frac{1}{NT^2}\sum_{i=1}^{N} Z_i' X_i(\delta-\hat{\delta})(\hat{\beta}_i-\beta_i)'\tilde{X}_i' Z_i =&\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T} z_{it}x_{it}'\right)(\delta-\hat{\delta})(\hat{\beta}_i-\beta_i)'\left(\frac{1}{T}\sum_{t=1}^{T}\tilde{x}_{it}z_{it}'\right)\\ =&O_p\left(\zeta_{NT}\cdot T^{-1/2}\right).\end{aligned}$$

For the sixth (and seventh) term(s), we have that

$$\frac{1}{NT^2}\sum_{i=1}^{N} Z_i'\alpha_i(\hat{\beta}_i-\beta_i)'\tilde{X}_i' Z_i = \frac{1}{N}\sum_{i=1}^{N}\bar{z}_i\alpha_i(\hat{\beta}_i-\beta_i)'\left(\frac{1}{T}\sum_{t=1}^{T}\tilde{x}_{it}z_{it}'\right) = O_p\left(\zeta_{NT}\cdot T^{-1/2}\right).$$

Finally, for the eighth (and ninth) term(s), it follows that

$$\frac{1}{NT^2}\sum_{i=1}^{N} Z_i'\alpha_i(\delta-\hat{\delta})'X_i' Z_i = \frac{1}{NT^2}\sum_{i=1}^{N}\bar{z}_i\alpha_i(\delta-\hat{\delta})'\left(\frac{1}{T}\sum_{t=1}^{T}x_{it}z_{it}'\right) = O_p\left(\zeta_{NT}^2\right).$$

Hence,

$$\left(\frac{1}{NT^2}\sum_{i=1}^{N} Z_i'\hat{u}_i\hat{u}_i' Z_i\right) = \frac{\Omega_1}{T}+\Omega_2+O_p(\zeta_{NT}^2+\zeta_{NT}T^{-1/2}),$$

and the final result follows directly by the continuous mapping theorem. ■

B.3.3 Overidentification Test

Proof of Proposition 2. First, we want to rewrite the J-statistics, in a way that accounts for the different convergence rates of the moments conditions:

$$\begin{aligned}J(\hat{\delta}) &= N\bar{g}_{NT}(\hat{\delta})'\hat{S}^{-1}\bar{g}_{NT}(\hat{\delta})\\ &= N\left(\Lambda_{NT}\bar{g}_{NT}(\hat{\delta})\right)'\Lambda_{NT}^{-1}\hat{S}^{-1}\Lambda_{NT}^{-1}\Lambda_{NT}\bar{g}_{N}(\hat{\delta})\\ &= \left(\Lambda_{NT}\bar{g}_{NT}(\hat{\delta})\right)'\Lambda_{T}^{-1}\hat{S}^{-1}\Lambda_{T}^{-1}\Lambda_{NT}\bar{g}_{NT}(\hat{\delta}).\\ &= \left(\Lambda_{NT}\bar{g}_{NT}(\hat{\delta})\right)'\hat{\Omega}^{-1}\Lambda_{NT}\bar{g}_{NT}(\hat{\delta}).\end{aligned}$$

where $\hat{\Omega}^{-1}=\left(\Lambda_T\hat{S}\Lambda_T\right)^{-1}$.

Second, we want to show that $\bar{g}_{NT}(\hat{\delta}) = \hat{B}\left(\bar{g}_{NT}(\delta)+\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T}\sum_{t=1}^{T} z_{it}\tilde{x}_{it}'(\hat{\beta}_i-\beta_i)\right)$. Recall that $\hat{Y}_i = X_i\delta+\alpha_i+\tilde{X}_i(\hat{\beta}_i-\beta_i)$. Hence, we can write

$$\begin{aligned}Z_i'\hat{Y}_i &= Z_i'X_i\delta + Z_i'\alpha_i + Z_i'\tilde{X}_i(\hat{\beta}_i-\beta)\\ S_{Z\hat{Y}} &= S_{ZX}\delta+\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T}\sum_{t=1}^{T} z_{it}\alpha_i+\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T}\sum_{t=1}^{T} z_{it}\tilde{x}_{it}'(\hat{\beta}_i-\beta_i)\\ S_{Z\hat{Y}} &= S_{ZX}\delta+\bar{g}_{NT}(\delta).\end{aligned}$$

Then, note that

$$
\begin{aligned}
\bar{g}_{NT}(\hat{\delta}) &= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T}\sum_{t=1}^{T} z_{it}(\hat{y}_{it} - x_{it}'\hat{\delta}) \\
&= S_{Z\hat{Y}} - S_{ZX}\hat{\delta} \\
&= S_{Z\hat{Y}} - S_{ZX}\left(S_{ZX}\hat{S}^{-1}S_{Z}X\right)^{-1} S_{ZX}\hat{S}^{-1}S_{Z\hat{Y}} = \hat{B}S_{Z\hat{Y}},
\end{aligned}
$$

where $\hat{B} = \left(I_L - S_{ZX}\left(S_{ZX}'\hat{S}^{-1}S_{ZX}\right)^{-1} S_{ZX}'\hat{S}^{-1}\right)$.
Thus,

$$
\begin{aligned}
\bar{g}_{NT}(\hat{\delta}) &= \hat{B}S_{Z\hat{Y}} \\
&= \left(I_L - S_{ZX}\left(S_{ZX}'\hat{S}^{-1}S_{ZX}\right)^{-1} S_{ZX}'\hat{S}^{-1}\right)(S_{ZX}\delta + \bar{g}_{NT}(\delta)) \\
&= \hat{B}\bar{g}_{NT}(\delta).
\end{aligned}
$$

Since Ω is positive definite there exist a matrix C such that $\hat{\Omega}^{-1} = C'C$.
We define $A \equiv C\Lambda_T S_{ZX}'$ and $M \equiv I_L - A(A'A)^{-1}A'$.
In this third part, we show that

$$
\hat{B}'\Lambda_{NT}\hat{\Omega}^{-1}\Lambda_{NT}\hat{B} = \Lambda_{NT}C'MC\Lambda_{NT}.
$$

Note that

$$
\begin{aligned}
C\Lambda_{NT}\hat{B} &= C\Lambda_{NT}\left(I_L - S_{ZX}\left(S_{ZX}'\hat{S}^{-1}S_{ZX}\right)^{-1} S_{ZX}'\hat{S}^{-1}\right) \\
&= \left(C\Lambda_{NT} - C\Lambda_{NT}S_{ZX}\left(S_{ZX}'\hat{S}^{-1}S_{ZX}\right)^{-1} S_{ZX}'\hat{S}^{-1}\right) \\
&= \left(C\Lambda_{NT} - C\Lambda_{NT}S_{ZX}\left(S_{ZX}'\Lambda_T C'C\Lambda_T S_{ZX}\right)^{-1} S_{ZX}'\Lambda_T C'C\Lambda_T\right) \\
&= \left(C\Lambda_{NT} - C\Lambda_{T}S_{ZX}\left(S_{ZX}'\Lambda_T C'C\Lambda_T S_{ZX}\right)^{-1} S_{ZX}'\Lambda_T C'C\Lambda_{NT}\right) \\
&= \left(I_L - A\left(A'A\right)^{-1}A'\right)C\Lambda_{NT} \\
&= MC\Lambda_{NT}.
\end{aligned}
$$

Where the third line uses $\hat{\Omega}^{-1} = \Lambda_T^{-1}\hat{S}^{-1}\Lambda_T^{-1} = C'C$. The fourth line follows because $\Lambda_{NT} = \sqrt{N}\Lambda_T$. In the last two lines, we use the definitions of A and M.
M is symmetric and idempotent. Thus

$$
\begin{aligned}
\hat{B}'\Lambda_{NT}\hat{\Omega}^{-1}\Lambda_{NT}\hat{B} &= \hat{B}'\Lambda_{NT}C'C\Lambda_{NT}\hat{B} \\
&= (C\Lambda_{NT}\hat{B})'C\Lambda_{NT}\hat{B} \\
&= (MC\Lambda_{NT})'MC\Lambda_{NT} \\
&= \Lambda_{NT}C'MC\Lambda_{NT}.
\end{aligned}
$$

The rank of M is the trace of M, which is $L-K$.
Since $\Omega = \Lambda_T S \Lambda_T$ is positive definite, there exist a matrix Q such that

$$Q'Q = \Omega^{-1}$$

and the probability limit of C is Q. We define $v \equiv C\Lambda_{NT}\bar{g}_{NT}(\delta)$.
It follows that

$$\Lambda_{NT}\bar{g}_{NT}(\delta) \xrightarrow{d} N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\Omega_{11} & 0\\ 0 & \Omega_{22}\end{pmatrix}\right) \sim N(0,\Omega).$$

Thus it follows directly that

$$v \xrightarrow{d} N(0, Q\Omega Q') = N(0, Q(Q'Q)^{-1}Q') = N(0, I_L).$$

Now we can come back to our test statistic:

$$\begin{aligned}
J(\hat{\delta}) &= N\bar{g}_{NT}(\hat{\delta})'\hat{S}^{-1}\bar{g}_{NT}(\hat{\delta})\\
&= \left(\Lambda_{NT}\bar{g}_{NT}(\hat{\delta})\right)'\hat{\Omega}^{-1}\Lambda_T^{-1}\bar{g}_{NT}(\hat{\delta})\\
&= \left(\Lambda_{NT}\hat{B}\bar{g}_{N}(\delta)\right)'\hat{\Omega}^{-1}\Lambda_{NT}\left(\hat{B}\bar{g}_{NT}(\delta)\right)\\
&= \bar{g}_{NT}(\delta)'\hat{B}'\Lambda_{NT}\hat{\Omega}^{-1}\Lambda_{NT}\hat{B}\bar{g}_{NT}(\delta)\\
&= \bar{g}_{NT}(\delta)'\Lambda_{NT}C'MC\Lambda_{NT}\bar{g}_{NT}(\delta)\\
&= \left[C\Lambda_{NT}\bar{g}_{NT}(\delta)\right]' M \left[C\Lambda_{NT}\bar{g}_{NT}(\delta)\right].
\end{aligned}$$

Since M is idempotent with rank L, it follows that

$$J(\hat{\delta}) \xrightarrow{d} \chi^2_{L-K}.$$

■