Inference on Time-Invariant Variables using Panel Data: A Pre-Test Estimator with an Application to the Returns to Schooling

Jean-Bernard Chatelain, Kirsten Ralf

▶ **To cite this version:**

Jean-Bernard Chatelain, Kirsten Ralf. Inference on Time-Invariant Variables using Panel Data: A Pre-Test Estimator with an Application to the Returns to Schooling. 2010. hal-00492039

HAL Id: hal-00492039
https://paris1.hal.science/hal-00492039

Preprint submitted on 14 Jun 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Inference on Time-Invariant Variables using Panel Data: A Pre-Test Estimator with an Application to the Returns to Schooling

Jean-Bernard Chatelain*and Kirsten Ralf†

January 15, 2010

Abstract

This paper proposes a new pre-test estimator of panel data models including time invariant variables based upon the Mundlak-Krishnakumar estimator and an "unrestricted" Hausman-Taylor estimator. The paper evaluates the biases of currently used restricted estimators, omitting the average-over-time of at least one endogenous time-varying explanatory variable. Repeated Between,

*CES, Centre d'Economie de la Sorbonne, Université Paris I Pantheon Sorbonne, Paris School of Economics and CEPREMAP. 106-112 Boulevard de l'Hôpital 75647 Paris Cedex 13. Email: jean-bernard.chatelain@univ-paris1.fr

†ESCE Ecole Supérieure du Commerce Extérieur, Paris and CEPREMAP.

Ordinary Least Squares, Two stage restricted Between and Oaxaca-Geisler estimator, Fixed Effect Vector Decomposition, Generalized least squares may lead to wrong conclusions regarding the statistical significance of the estimated parameter values of time-invariant variables.

JEL classification numbers: C01, C22, C23.

Keywords: Time-Invariant Variables, Panel data, Time-Series Cross-Sections, Pre-Test Estimator, Mundlak Estimator, Fixed Effects Vector Decomposition.

1. Introduction

We observe in sample time-invariant variables such as variables of origin, rare events or rarely changing variables which have been measured only once within the period of interest. Let us give a few examples of time invariant variables: colonial, legal or political system, international conflicts, institutional and governance indicators, initial gross domestic product per head when testing the convergence of incomes in growth regressions; geographical position for cross country data in gravity models of foreign trade and foreign direct investments; years of schooling, gender and race when testing wage income using survey data. These variables are often highly relevant in a theoretical model predicting correlations with a cross-sections and time-varying variable of interest. Because publishing results that do not reject the null hypothesis of no effect of these time invariant variable on the time varying variable of interest is generally not accepted by journal editors, applied researchers are likely to select

estimators that reject the null hypothesis among the pool of available estimators.

We are then interested in drawing inference with respect to the statistical significance of these time-invariant variables (observed for N individuals) for explaining the variance of a time and individual varying variable (observed for N individuals during T periods). Since a time-invariant variable has no variance in the time direction, it can only explain the variance of a time and individual varying variable in its individual direction. The reason why it matters for inference is that "*the effect of a random component can only be averaged out if the sample increases in the direction of that random (time or individual) component*" (Kelejian and Stephan 1983, see also Hsiao 2003, pp. 51-53.). For example, the pooled ordinary least square (OLS) estimator with time-invariant explanatory variables leads to inference on time-invariant variables using the time dimension NT. But "*we have to deal with the unfortunate fact that there is not quite so much information in N individuals observed T times as there is with NT individuals.*" (Johnston and Di Nardo (1997), pp. 395.).

This leads to the suggestion of a number of possible estimators for the parameter values with different characteristics. The available estimators of time invariant variables using panel data or time series cross sections are found in Hausman and Taylor (1981), Hsiao (2003), Baltagi, Bresson and Pirotte (2003), Oaxaca and Geisler (2003), Krishnakumar (2006) extension of Mundlak (1978) and Plümper and Troeger (2007) Fixed Effect Vector Decomposition Estimator (FEVD). For example, the FEVD estimator has been widely used over the recent years (for example, Blaydes (2006)

and Goodrich (2006)). By contrast, the Mundlak-Krishnakumar (2006) estimator remained unnoticed by practitionners.

This paper suggests that inference on time-invariant variables without a pre-test Mundlak-Krishnakumar estimator may lead to conclude wrongly that time-invariant variables are statistically significant. This paper includes four contributions:

(1) It proposes a pre-test estimator based upon the Mundlak-Krishnakumar estimator and a modified Hausman-Taylor estimator, extending the pre-test estimator proposed Baltagi, Bresson and Pirotte (2003).

(2) The Mundlak-Krishnakumar model is not currently used for doing inference on time invariant variables. *It means that some currently published papers where the data generating process could be approximated by the Mundlak-Krishnakumar theoretical model report, for example, restricted estimators omitting the average over time of endogenous time-varying variables.* The paper computes the determinants of the bias of the estimated parameters and of the estimated standard errors of using other estimators instead of the pre-test estimator. It presents an illustration of these omitted variable biases on a returns to schooling classic database (Baltagi and Khanti-Akom (1990), Cornwell and Rupert (1988), Baltagi (2008 and 2009), Cameron and Trivedi (2009), chapter 8).

(3 and 4) The paper explains why the Oaxaca-Geisler (2003) and FEVD estimators may provide misleading estimated standard errors of the parameters of time-invariant variables. Mitze (2009) and Breusch, Ward, Nguyen and Kompas (2010) found that

FEVD estimated standard errors of the estimated parameters of time invariant variables were (too) small. The explanation that we propose is that the FEVD estimator uses the within root mean square error for the estimator of these standard errors. In so doing, it contradicts the Frisch-Waugh-Lovell theorem, where the same orthogonal projection matrix has to be used in the parameter estimate and in the estimator of its variance (following a similar argument, Greene (2010) refers to Aitken's theorem).

The overall picture is that when one uses this pre-test estimator, results may differ a lot from currently published inference which neglects a pre-test stage. There are two reasons for that. First, an omitted variable bias of the average over time of endogenous time-varying variables may occur. Secondly, the estimator of the standard error of the parameters of time invariant variables may be biased because of an excessive weight on within mean square error and on within degrees of freedom (depending on the time dimension T). As the extent of these potential biases are not known when the data sets are not available for replication, this casts serious doubts on the emphasis put on inference on time invariant variable in panel data formerly published in academic journals.

This paper proceeds as follows. Section 2 defines our pre-test estimator based upon the Mundlak-Krishnakumar and the Hausman and Taylor estimator. Section 3 compares the outcome of different estimators using a typical time series cross section data set of gasoline demand. Section 4 analyses the differences of our pre-test estimator with the currently available estimators, in particular, the Oaxaca-Geisler, the two

stage, the FEVD and the usual random effects estimators. Section 5 concludes.

2. A Pre-Test Estimator including Time-invariant Variables and Correlated Individual Effects

2.1. Mundlak-Krishnakumar Estimator

The model of time-series cross-sections regression estimates the following equation

$$y_{it} = \mathbf{X}_{it}\beta + \mathbf{Z}_i\gamma + \alpha_i + \varepsilon_{it} \tag{2.1}$$

where y_{it} denotes the endogenous variable, $\mathbf{X}_{it}$ is a $NT \times k$ matrix of cross-sections time-series data, $\mathbf{Z}_i$ is a $NT \times g$ matrix of time invariant variables, β and γ are k and g vectors of coefficients associated with time-varying and time-invariant observable variables respectively. Subscripts indicate variation over individuals ($i = 1, ..., N$) and time ($t = 1, ..., T$). Observations are ordered first by individual and then by time, so that α_i and each column of $\mathbf{Z}_i$ are NT vectors having blocks of T identical entries within each individual $i = 1, ..., N$. The disturbance ε_{it} is assumed to be uncorrelated with the columns of ($\mathbf{X}_{it}, \mathbf{Z}_i$,$\alpha_i$) and has zero mean and constant variance σ_ε^2 conditional on $\mathbf{X}_{it}$ and $\mathbf{Z}_i$. The individual effect α_i is assumed to be a time-invariant random variable, distributed independently across individuals with variance σ_α^2. The primary focus of the literature is the potential correlation of α_i with the columns of ($\mathbf{X}_{it}, \mathbf{Z}_i$).

Mundlak [1978] and Krishnakumar [2006] introduce an auxiliary regression which takes explicitly into account a linear relation between the explanatory variables and the individual random effects:

$$\alpha_i = \mathbf{X}_{i.}\pi + \mathbf{Z}_i\phi + \alpha_i^M \tag{2.2}$$

where it is assumed that the disturbance $\alpha_i^M \sim (0, \sigma^2_{\alpha^M})$ and where $\mathbf{X}_{i.}$ is the average over time for each individual of time varying variables, such that $E\left(\alpha_i^M \mid \mathbf{X}_{i.}\right) = 0$, π and ϕ are k and g vectors of coefficients associated with the average over time of time-varying variables and time-invariant observable variables, respectively. Clearly $\pi = 0$ and $\phi = 0$ if and only if the time varying and time invariant variables are uncorrelated with the random individual effects. Combining the auxiliary regression with the initial regression yields:

$$y_{it} = \mathbf{X}_{it}\beta + \mathbf{Z}_i\left(\gamma + \phi\right) + \mathbf{X}_{i.}\pi + \alpha_i^M + \varepsilon_{it} \tag{2.3}$$

When $E\left(\alpha_i \mid \mathbf{Z}_i\right) = \phi \neq 0$, one may only estimate the sum $\gamma + \phi$ and may not identify γ and ϕ separately, without additional prior information. The prior information the Hausman and Taylor (1981) procedure uses is the ability to distinguish columns of $\mathbf{X}_{it}$ and $\mathbf{Z}_i$ which are asymptotically uncorrelated with α_i from those which are not. For fixed T, let

$$\begin{aligned}
\operatorname*{plim}_{N\to\infty} \frac{1}{N}\mathbf{X}_{1it}^{'}\alpha_i &= \pi_1 = \mathbf{0}, \ \operatorname*{plim}_{N\to\infty} \frac{1}{N}\mathbf{X}_{2it}^{'}\alpha_i = \pi_2 \neq \mathbf{0} \\
\operatorname*{plim}_{N\to\infty} \frac{1}{N}\mathbf{Z}_{1i}^{'}\alpha_i &= \phi_1 = \mathbf{0}, \ \operatorname*{plim}_{N\to\infty} \frac{1}{N}\mathbf{Z}_{2i}^{'}\alpha_i = \phi_2 \neq \mathbf{0}
\end{aligned}$$

where $\mathbf{X}_{it} = [\mathbf{X}_{1it}, \mathbf{X}_{2it}]$ and $\mathbf{Z}_i = [\mathbf{Z}_{1i}, \mathbf{Z}_{2i}]$ are split into two sets of variables such that $\mathbf{X}_1$ is $NT \times k_1$, $\mathbf{X}_2$ is $NT \times k_2$, $\mathbf{Z}_1$ is $NT \times g_1$, $\mathbf{Z}_2$ is $NT \times g_2$ with $k_1 + k_2 = k$ and $g_1 + g_2 = g$. $\mathbf{X}_1$ and $\mathbf{Z}_1$ are assumed exogenous and not correlated with α_i and ε_{it}, while $\mathbf{X}_2$ and $\mathbf{Z}_2$ are endogenous due to their correlation with α_i but not with ε_{it}. The pre-test estimator proposed in the next section differs from the Hausman and Taylor [1981] estimator in that no prior information is assumed about how to split π, but that the information is obtained using t-tests on each estimated parameter $\widehat{\pi}$ of equations (2.2) or equation (2.3).

It will prove helpful to recall the menu of conventional estimators for (β, γ) in equation (2.3). Letting i_T denote a T vector of ones, two orthogonal projection operators can be defined as:

$$\mathbf{B} = \mathbf{I}_N \otimes \frac{1}{T} i_T i_T^{'}, \ \mathbf{W} = \mathbf{I}_{NT} - \mathbf{B}$$

which are idempotent matrices of rank N and $NT-N$ respectively. With data grouped by individuals, $\mathbf{B}$ transforms a vector of observations into a vector of group means, such that $\mathbf{B}y_{it} = y_{i.}$. Similarly, $\mathbf{W}$ produces a vector of deviation from group means:

i.e., $\mathbf{W}y_{it} = y_{it} - y_{i.}$. For any time-invariant vector of observations, $\mathbf{W}z_i = \mathbf{O}$. The orthogonality of the two projection operators is such $\mathbf{BW} = \mathbf{WB} = \mathbf{O}_{NT}$. The space of observations ($\mathbb{R}^{NT}$) has an orthogonal decomposition in two spaces: the subspace of observations transformed by the between operator (dimension N) and the subspace of observations transformed by the within operator (dimension $NT-N$). In other words, the covariance of the between transformed variable (time average for each individual i, with t a time index and T observations) $x_{i.} = \frac{1}{T}\sum_{t=1}^{t=T} x_{it}$ with respect to a within transformation of another variable ($y_{it} - y_{i.}$) is always equal to zero:

$$cov\left(y_{it} - y_{i.}, x_{i.}\right) = 0 \text{ and } cov\left(y_{it} - y_{i.}, y_{i.}\right) = 0 \tag{2.4}$$

Hence, the total sum of squares SST of y_{it} is the sum of total sum of squares SST_W for the within transformed variable and T times the total sum of square SST_B for the between transformed variable over N observations:

$$SST\left(y_{it}\right) = SST\left(y_{it} - y_{i.}\right) + SST\left(y_{i.}\right) = SST_W + T \cdot SST_B. \tag{2.5}$$

Transform model (3) by the Within projection operator, we obtain:

$$y_{it} - y_{i.} = \mathbf{WX}_{it}\beta + \mathbf{WZ}_i\left(\gamma + \phi\right) + \mathbf{WX}_{i.}\pi + \mathbf{W}\alpha_i^M + \mathbf{W}\varepsilon_{it} \tag{2.6}$$

$$y_{it} - y_{i.} = \left(\mathbf{X}_{it} - \mathbf{X}_{i.}\right)\beta + \varepsilon_{it} - \varepsilon_{i.} \tag{2.7}$$

Least square estimates of β in the within transformed equation are Gauss-Markov for the transformed equation and define the within-groups estimator:

$$\widehat{\beta}_W = \left(\mathbf{X}'_{it}\mathbf{F}\mathbf{X}_{it}\right)^{-1}\mathbf{X}'_{it}\mathbf{F}y_{it} \text{ with } \mathbf{F} = \mathbf{W}. \tag{2.8}$$

Since the columns of $\mathbf{W}\mathbf{X}_{it}$ are uncorrelated with $\mathbf{W}\varepsilon_{it}$, $\widehat{\beta}_W$ is unbiased and consistent for β regardless of possible correlation between α_i and the columns of $\mathbf{X}_{it}$ or $\mathbf{Z}_i$. The OLS analysis of variance in the within subspace of observations (SSM is the sum of squares of the model, SSE is the sum of squares of the error) is: $SST_W = SSM_W + SSE_W$. The sum of squared residuals (denoted form this equation can be used to obtain an unbiased and consistent estimate of σ^2_ε:

$$MSE_W = \widehat{\sigma^2_\varepsilon} = \frac{SSE_W}{NT - N - k} \tag{2.9}$$

To make use of between-group variation, transform model (3) by the Between projection operator obtaining

$$y_{i.} = \mathbf{X}_{i.}\left(\beta + \pi\right) + \mathbf{Z}_i\left(\gamma + \phi\right) + \alpha_i^M + \varepsilon_{i.}$$

When $E\left(\alpha_i \mid \mathbf{X}_{it}, \mathbf{Z}_i\right) = 0 = \pi = \phi$, least squares estimates of β and γ in the between transformed regression define the between groups estimators estimators (denoted $\widehat{\beta}_B$ and $\widehat{\gamma}_B$) which are unbiased and consistent for β and γ, using N observa-

tions. When $E(\alpha_i \mid \mathbf{X}_{it}) \neq 0$ or when $\pi \neq 0$, $\widehat{\beta}_B$ is biased and inconsistent. When $E(\alpha_i \mid \mathbf{Z}_i) \neq 0$ or when $\phi \neq 0$, $\widehat{\gamma}_B$ is biased and inconsistent. The OLS analysis of variance in the between subspace of observations is: $SST_B = SSM_B + SSE_B$. The sum of squared residuals SSE_B provides a biased and inconsistent estimator for $var(\alpha_i + \varepsilon_{i.}) = \sigma^2_{\alpha_i} + (1/T)\sigma^2_{\varepsilon}$ when $E(\alpha_i \mid \mathbf{X}_{it}, \mathbf{Z}_i) \neq 0$ and unbiased and consistent when $E(\alpha_i \mid \mathbf{X}_{it}, \mathbf{Z}_i) = 0$. Whenever we have consistent estimators for both β and γ, a consistent estimator for σ^2_{α} can be obtained with:

$$MSE_B = \widehat{\sigma^2_{\alpha}} + \frac{\widehat{\sigma^2_{\varepsilon}}}{T} = \frac{SSE_B}{N-k-g-1}. \tag{2.10}$$

Estimators that correspond to weighted average of the within and the between estimators are obtained as follows. Summing the within equation and the between equation multiplied by a parameter θ and repeated T times, which amounts to transform equation (3) applying the linear operator $\mathbf{W}+\theta\mathbf{B}$, we obtain:

$$y_{it} - y_{i.} + \theta y_{i.} = (\mathbf{W}+\theta\mathbf{B})\mathbf{X}_{it}\beta + \theta\mathbf{Z}_{i.}(\gamma+\phi) + \theta\mathbf{X}_{i.}\pi + \theta\alpha_i^M + (\mathbf{W}+\theta\mathbf{B})\varepsilon_{it} \tag{2.11}$$

When $\pi \neq 0$, that is when $E(\alpha_i \mid \mathbf{X}_{it}) \neq 0$, and when $E(\alpha_i \mid \mathbf{Z}_i) = 0 = \phi$, Mundlak [1978] and Krishnakumar [2006] prove that the best linear unbiased estimator is obtained for a value of θ corresponding to the generalized least square model (GLS)

or random effect estimator, which leads to the following estimates:

$$\widehat{\beta}_{GLS} = \widehat{\beta}_W, \widehat{\pi}_{GLS} = \widehat{\beta}_B - \widehat{\beta}_W \text{ and } \widehat{\gamma}_{GLS} = \widehat{\gamma}_B, \tag{2.12}$$

$$var\left(\widehat{\pi}_{GLS}\right) = var\left(\widehat{\beta}_B\right) + var\left(\widehat{\beta}_W\right) \text{ and } var\left(\widehat{\gamma}_{GLS}\right) = var\left(\widehat{\gamma}_B\right) \tag{2.13}$$

A demonstration for the equalities on estimated variance-covariance matrix of estimated parameters is also found in Baltagi (2009, p.152-154). More precisely, Mundlak (1978) shows that (i) if α_i is correlated with *every* column of $\mathbf{X}_{i.}$ ($\pi \neq 0$, with $k_1 = 0$ and $k_2 = k$), the Gauss-Markov estimator for β is the within groups estimator $\widehat{\beta}_W$, and (ii) if α_i is uncorrelated with *every* column of $\mathbf{X}_{i.}$ ($\pi = 0$ in the true model, with $k_1 = k$ and $k_2 = 0$), the Gauss-Markov estimator for β is the "usual" GLS estimator $\widehat{\beta}_{GLS}$, (iii) Using the "usual" GLS estimate assuming the restriction $\pi = 0$ when the true model is such that $\pi \neq 0$ is denoted "'restricted GLS" ($RGLS$) and leads to biased estimate $\widehat{\beta}_{RGLS} \neq \widehat{\beta}_{GLS} = \widehat{\beta}_W$, because of the k omitted variables $\mathbf{X}_{i.}$ bias. The estimator $\widehat{\pi}_{GLS}$ is equal to the difference between the between estimator and the within estimator $\widehat{\beta}_{1,B} - \widehat{\beta}_{1,W}$.

For the time-invariant variables, the estimator $\widehat{\gamma}_{GLS}$ is exactly the between estimator $\widehat{\gamma}_B$. The standard error estimator, $\widehat{\sigma}_{\widehat{\gamma}_{GLS}}$, of the estimated parameter using the Mundlak GLS model is exactly the same as the one of the between regression (see appendix 1). This estimator – even though the data includes NT observations – takes into account that for time-invariant explanatory variables only N observations should

be used. It has $N-k-g-1$ degrees of freedom:

$$\widehat{\sigma}_{\widehat{\gamma}_{GLS}} = \widehat{\sigma}_{\widehat{\gamma}_B} = \sqrt{\frac{SSE_B}{N-k-g-1}} \frac{1}{\sqrt{CSS(z_j)}\sqrt{1-R_A^2(z_j)}}, \text{for } j = 1, ..., g \qquad (2.14)$$

where CSS is the sum of squares corrected for the mean and $1-R_A^2(z_j)$ is the "tolerance" (or the inverse of the variance inflation factor) and where $R_A^2(z_j)$ is the coefficient of determination of the following auxiliary regression where the explanatory variable z_i is explained by all *other* explanatory variables of regression (3) in the between dimension (Stewart 1987):

$$z_j = \mathbf{X}_{i.}\pi' + \mathbf{Z}_{i,-j}\gamma' + \eta_j^{'}$$

where $\eta_j^{'}$ are disturbances, where $\pi'$and γ' are coefficients, and where $\mathbf{Z}_{i,-j}$ is the matrix of time-invariant variables excluding the column of observations related to the variable z_j. $R_A^2(z_i)$ measures the effect of other explanatory variables on the estimated standard error of the estimated parameter of a given variable z_i. When $R_A^2(z_i)$ is close to unity, there is a potential problem of near-multicollinearity. The root mean squared error of the between regression (including the degrees of freedom in the denominator) is $\sqrt{\frac{SSE_B}{N-k-g-1}}$. The convergence of the estimator is obtained by increasing N to infinity, for T fixed. An increase of T does not lead the between estimator to converge. Inference on time-invariant explanatory variables do not depend on the number of observations in the time dimension T.

2.2. A Pre-test Estimator

The pre-test estimator consists of a first step using the Mundlak-Krishnakumar estimator (equation (2.3) using GLS method) and then perform specification tests of null hypothesis $H_0 : E(\alpha_i \mid \mathbf{X}_{mit}) = \pi_m = 0$ against the alternative $H_0 : E(\alpha_i \mid \mathbf{X}_{mit}) = \pi_m \neq 0$ for each of the time-varying explanatory variables $0 \leq m \leq k$. If H_0 is rejected for exactly k_2 variables, these variables are included in the subset $\mathbf{X}_{2it}$ assumed to be endogenous due to their correlation with α_i but not with ε_{it}, while the remaining $k_1$variables are included in the subset $\mathbf{X}_{1it}$. *Hence, no prior information on how to split* $\mathbf{X}_{it}$ *is required for this pre-test estimator.* In practice, researchers priors on how to split $\mathbf{X}_{it}$ and $\mathbf{Z}_i$ based on theory may be misleading because of unobservable and omitted time-invariant variables, measurement errors, and so on.

The second step consists of using an *unrestricted* Hausman Taylor (1981) estimator using $\mathbf{X}_{1it}$ as instruments for $\mathbf{Z}_2$ endogenous time invariant variables, but *keeping the time invariant variables* $\mathbf{X}_{2i.}$. which are not used as instruments, in order to correct for their endogeneity (this is not usually done with the Hausman and Taylor estimator).

More precisely, if the specification tests rejects H_0 for $k_2 \geq 1$ variables, it may still be possible to obtain consistent estimates of both β and γ in a second stage. Let

$$\widehat{d}_i = y_{i.} - \mathbf{X}_{i.}\widehat{\beta}_W = \left(\mathbf{B} - \mathbf{X}_{i.}\left(\mathbf{X}'_{it}\mathbf{W}\mathbf{X}_{it}\right)^{-1}\mathbf{X}'_{it}\mathbf{W}\right) y_{it}$$

be the NT vector of group means estimated from the within-groups residuals. Ex-

panding this expression using equation 2.3 including only $\mathbf{X}_{2i.}$leads to:

$$\widehat{d}_i = \mathbf{Z}_{1i.}\gamma_1 + \mathbf{Z}_{2i.}\left(\gamma_2 + \phi_2\right) + \mathbf{X}_{2i.}\pi_2 + \alpha_i^M + \left(\mathbf{B} - \mathbf{X}_{i.}\left(\mathbf{X}_{it}'\mathbf{W}\mathbf{X}_{it}\right)^{-1}\mathbf{X}_{it}'\mathbf{W}\right)\varepsilon_{it}. \quad (2.15)$$

Treating the last two terms as an unobservable mean zero disturbance, consider estimating γ from the above equation using N observations. If α_i is correlated with the columns of $\mathbf{Z}_{2i}$, $E\left(\alpha_i \mid \mathbf{Z}_{2i}\right) = \phi_2 \neq 0$, according to prior information, both OLS and GLS will be inconsistent estimates for γ. Consistent estimation is possible, however, if the columns of $\mathbf{X}_{1it}$, uncorrelated with α_i according to the non rejection of the null hypothesis of preliminary tests, provide sufficient instruments for the columns of $\mathbf{Z}_{i.}$ in equation (4.1). A necessary condition for identification of γ_2 and ϕ_2 is that $k_1 \geq g_2$: there are at least as many exogenous time-varying variables as there are endogenous time-invariant variables (proposition 3.2 in Hausman and Taylor (1981)). When the condition $k_1 \geq g_2$ is fulfilled, one may proceed to a second step for estimating γ knowing that the first step provided efficient estimates of β.

The two stage least squares (2SLS) estimator for γ in equation (4.1) is:

$$\widehat{\gamma}_{II} = \left(\left[\mathbf{Z}_i', \mathbf{X}_{2i.}\right]\mathbf{P}_A\left[\mathbf{Z}_i', \mathbf{X}_{2i.}\right]\right)^{-1}\left[\mathbf{Z}_i', \mathbf{X}_{2i.}\right]'\mathbf{P}_A\widehat{d}_i \quad (2.16)$$

where $\mathbf{A} = \left[\mathbf{X}_{1it}, \mathbf{Z}_{1i}\right]$ and $\mathbf{P}_A$ is the orthogonal projection operator onto its column space. The sampling error is given by

$$\widehat{\gamma}_{II}-\gamma = \left(\left[\mathbf{Z}_i', \mathbf{X}_{2i.}\right] \mathbf{P}_A \left[\mathbf{Z}_i', \mathbf{X}_{2i.}\right]\right)^{-1} \left[\mathbf{Z}_i', \mathbf{X}_{2i.}\right]' \mathbf{P}_A \left(\alpha_i^M + \left(\mathbf{B} - \mathbf{X}_{i.} \left(\mathbf{X}_{it}' \mathbf{W} \mathbf{X}_{it}\right)^{-1} \mathbf{X}_{it}' \mathbf{W}\right) \varepsilon_{it}\right)$$

and under the usual assumptions governing $\mathbf{X}_{it}$ and $\mathbf{Z}_i$, the 2SLS estimator is consistent for γ, since for fixed T, $\text{plim}_{N\to\infty} \frac{1}{N} \mathbf{A}' \alpha_i = 0$ and $\text{plim}_{N\to\infty} \frac{1}{N} \mathbf{X}_{it}' \varepsilon_i = 0$. Having consistent estimates of β and, under the condition $k_1 \geq g_2$, γ, we can construct consistent estimators for the variance components. A consistent estimate of σ_ε^2 can be derived from the within-group residuals in the first step $\widehat{\sigma_\varepsilon^2} = MSE_W$. Whenever we have consistent estimators for both β and γ, a consistent estimator for σ_α^2 can be obtained. Let $s^2 = (1/N) \left(Y_{i.} - X_{i.}\widehat{\beta}_W - Z_i\widehat{\gamma}_{II} - \mathbf{X}_{2i.}\widehat{\pi}_{2,II}\right)' \left(Y_{i.} - X_{i.}\widehat{\beta}_W - Z_i\widehat{\gamma}_{II} - \mathbf{X}_{2i.}\widehat{\pi}_{2,II}\right)$: then

$$\operatorname*{plim}_{N\to\infty} s^2 = \operatorname*{plim}_{N\to\infty} \frac{1}{N} \left(\alpha_i + \varepsilon_i\right)' \left(\alpha_i + \varepsilon_i\right) = \sigma_\alpha^2 + \frac{1}{T}\sigma_\varepsilon^2$$

so that $s_a^2 = s^2 - (1/T)\, s_\varepsilon^2$ is consistent for s_a^2.

In the particular case when one assumes that α_i is uncorrelated with the columns of $\mathbf{Z}_i$: $E\left(\alpha_i \mid \mathbf{Z}_i\right) = 0 = \phi$, and when all time varying variables reject the null hypothesis H_0 $(k_1 = 0)$ using N observations to estimate the above equation in a second stage II, both OLS and GLS will be consistent estimates of γ identical to the between estimates ($\widehat{\gamma}_{II,OLS} = \widehat{\gamma}_B$, $\widehat{\sigma}_{\widehat{\gamma}_{II,OLS}} = \widehat{\sigma}_{\widehat{\gamma}_B}$) and with $\widehat{\pi}_{II,OLS} = \widehat{\beta}_B - \widehat{\beta}_W$.

Baltagi, Bresson and Pirotte (BBP) (2003) pre-test estimator reverts to the usual

random effects estimator if the standard Hausman test based on the "within-groups" versus the random effects estimators is not rejected. It reverts to the HT estimator if the choice of strictly exogenous regressors is not rejected by a second Hausman test based on the difference between the "within-groups" and HT estimators. Otherwise, this pre-test estimator reverts to the "within-groups" estimator.

The pre-test estimator differs from BBP (2003) in that (1) it assumes Mundlak auxiliary regression, (2) it tests the endogeneity of each time-varying variables instead of relying on prior information for splitting $\mathbf{X}_{it} = [\mathbf{X}_1, \mathbf{X}_2]$, (3) it includes $\mathbf{X}_{2i.}\pi_2$ which may change widely the estimates of γ in the second stage, (4) when $0 \leq k_1 < g_2$ (not enough exogenous $\mathbf{X}_{1it}$ to instrument endogenous $\mathbf{Z}_{2i}$) or respectively when $g_2 = 0$ and $k_1 = k$ (no endogenous $\mathbf{Z}_{2i}$, and all time varying variables are endogenous $\mathbf{X}_{1it}$) it reports a biased (respectively unbiased) estimators for γ, instead of no estimates at all.

Because classical statistical hypothesis testing implies non-zero probabilities of type I error (p-value) and type II error (one minus the power of the test), a pre-test estimator cannot perform as well as an estimator where the researcher exactly knows the "true" model based on "true prior information" before testing. This only occurs when the number of observations tends to infinity, so that the probabilities of both types of errors tends to zero.

Mundlak (1978) considers the following estimators for his theoretical model: pooled OLS ($\theta = 1$), between, within ($\theta = 0$), Generalized least square (GLS) or random

effects estimator (for an estimated value of θ), restricted generalized least square with the restriction $\pi = 0$ (that is applying the usual random effect model whereas the true model is such that: $\pi \neq 0$), restricted pooled OLS with the restriction $\pi = 0$. The reasons for considering restricted estimators are twofold. Firstly, these restrictions are likely to decrease the variance of the estimators although these restricted estimators are generally biased. There is therefore a trade off between bias and variance and the choice of an estimator depends on the weights to be assigned to the two components. Secondly, the Krishnakumar-Mundlak is not currently used for doing inference on time invariant variables. *It means that some currently published papers where the data generating process could be approximated by the Mundlak theoretical model report Mundlak's restricted estimators.* It is interesting to evaluate the bias of the estimated parameters and of the estimated standard errors of the time invariant variables for the restricted models. We therefore do a similar investigation as Mundlak (1978) did on retricted estimators for time varying variables, and we focus on the estimators related to time-invariant variables when our pre-test is estimator is consistent.

3. A Return to Schooling Illustration of the Unrestricted Model

3.1. Pre-test Estimator

The pre-test estimator leads to dramatic changes with respect to alternative estimators of time invariant variables in panel data. This is demonstrated empirically for a return to schooling example based on a panel of 595 individuals observed over the

period 1976-1982 drawn from the Panel Study of Income Dynamics (PSID) (Baltagi and Khanti-Akom (1990), Cornwell and Rupert (1988)). First, the tests based on Mundlak-Krishnakumar specification lead to alternative choices of the instruments than the ones chosen by Baltagi and Khanti-Akom (1990) and Cornwell and Rupert (1988). Second, when the average over time of endogenous time varying variables are not omitted in the Hausman-Taylor estimates, the estimates of the return to schooling are much lower than the ones found by Baltagi Khanti-Akom [1990] and Cornwell and Rupert [1988].

In tables 1 and 2, the dependent variable log(wage) is explained by nine time-varying variables ($\mathbf{X}_{it}$'s) and three time-invariant variables ($\mathbf{Z}_i$'s). Baltagi and Khanti-Akom (1990) and Cornwell and Rupert (1988) assumed that four of the $\mathbf{X}_{it}$ variables and two of the $\mathbf{Z}_i$ variables are uncorrelated with the individual effects. These are denoted by $\mathbf{X}_1$ and $\mathbf{Z}_1$, respectively. They are listed at the bottom of table 2. The remaining $\mathbf{X}_{it}$ and $\mathbf{Z}_i$ are correlated with the individual effects. They are denoted by $\mathbf{X}_2$ and $\mathbf{Z}_2$, respectively.

Table 1 includes estimates of the unrestricted model using six estimators: within, between, Mundlak-Krishnakumar, pre-test unrestricted Hausman-Taylor, repeated between (RB), ordinary least square (OLS) and FEVD. Table 2 includes six estimates of the restricted model: step II restricted between, step II repeated restricted between, GLS, OLS, FEVD and HT($\mathbf{X}_1^{**}$), where the set of exogenous time varying variables

$\mathbf{X}_1^{**}$ is the one chosen by Baltagi and Khanti-Akom (1990). [1]

For the unrestricted model, we observe the following. The estimates of the all parameters are identical for all estimators, except for the average over time of a time-varying variable in the Between or Repeated Between estimators which exclude the time- and individual-varying variable. For the time- and individual-varying variable, Mundlak-Krishnakumar, LSDV and FEVD estimators exhibit the same estimated standard deviation and value of the t-statistics, which differs from OLS. But, for the time-invariant variable, the estimated standard deviation and the values of the t-statistic vary dramatically. For the Between and the Mundlak-Krishnakumar estimators, the estimated standard deviations of time invariant variables are equal, whereas the estimated standard deviations for the repeated between estimate are 2.67 times smaller, the ones for the OLS estimate are 2.33 times smaller, and the ones for the FEVD estimate are 4.63 smaller. It seems, and this is confirmed later on by our theoretical analysis, that the standard deviation of the OLS and FEVD is systematically smaller and the value of the t-statistic systematically larger than for the Mundlak estimator. If this is the case, inference with the OLS and FEVD estimator will less often reject the null-hypothesis that γ is equal to zero, and more often conclude that the time-invariant variable has an effect on the endogenous variable statistically different from zero.

In table 1, the within estimator, which is consistent, serves as a benchmark for the

[1] Five of these thirteen estimates are also reported in Baltagi [2009], p.27-29: Within, Between, Restricted GLS, Restricted OLS and p.157, Restricted Hausman Taylor HT($\mathbf{X}_1^{**}$).

Hausman (1978) test. The experience variable (EXP) is a time trend which differs only in levels for each individuals, i.e. $EXP_{it} = EXP_{i.} + t$ where $EXP_{i.}$ is the average number of years of experience in the sample of years 1976-1982 for each individual i and t is a time trend (= -3 for 1976,...,= $+3$ for 1982). This individual time trend accounts for 65% of the within variance of the dependent variable log(wage), along with a correlation coefficient of 0.8. Once this variable is included in the model, the partial R2 contribution of the eight other within transformed variables is very small (the $R2$ increases by only 0.74%). The estimated coefficients of blue collar occupation (OCC), of marital status (MS), of living in the south of the U.S.A. (SOUTH), of weeks worked (WKS), of industry (IND), of the standard metropolitan statistical area (SMSA), of belonging to an union (UNION) and of EXP^2 are statistically significant.

The between estimator includes estimates of the three time invariant variables: female workers (FEM), black worker (BLK), and years of schooling (ED). Degrees of freedom are 582. All the estimated coefficients are statistically significant. The 95% confidence interval for the return to school is $[0.045, 0.057]$. The standard error estimator, $\widehat{\sigma}_{\widehat{\gamma}_{GLS}}$, of the estimated parameter using the Mundlak-Krishnakumar GLS model is exactly the same as the one of the between regression (table 1, column 2):

$$\widehat{\sigma}_{\widehat{\gamma}_{GLS}} = \widehat{\sigma}_{\widehat{\gamma}_B} = \frac{\sqrt{\frac{SSE_B}{N-k-g-1}}}{\sqrt{CSS(z_i)}\sqrt{1-R_A^2\left(z_j\right)}} = \frac{\sqrt{\frac{42.07257}{595-13}}}{\sqrt{4623.77479}}\frac{1}{\sqrt{0.50673}} = 0.00555$$

where $1 - R_A^2 = 0.5067$ is the tolerance and CSS is the sum of squares corrected for the mean of the time invariant explanatory variable z_i.

The Mundlak-Krishnakumar estimator is able to deal with a non linear model (this one includes EXP^2) as long as the model remains linear with respect to the parameters. It provides t-tests for parameters of the average over time of the nine time-varying variables, indexed by m ($H_0 : \widehat{\pi}_{GLS,m} = \widehat{\beta}_B - \widehat{\beta}_W$). At the 5% threshold, these tests accept the null hypothesis for the parameters of the variables $\mathbf{X}_1^* = (EXP^2$, SOUTH, IND) which is the choice of exogenous variable of the pre-test estimator. It is remarkable that EXP is strongly endogenous in the sample whereas EXP^2 is exogenous in the sample. Cornwell and Rupert [1988] assumed that $\mathbf{X}_1 =$ (SOUTH, WKS, SMSA, MS) are exogenous. Baltagi and Khanti-Akom (1990) assumed that $\mathbf{X}_1^{**} =$ (SOUTH, IND, OCC, SMSA) are exogenous. However, the t-statistic in the Mundlak parameter is equal to $t = 4.01$ for OCC and to $t = 6.77$ for SMSA. The choice of exogenous time-invariant variables is based on prior assumption: $\mathbf{Z}_1 =$ (FEM, BLK). Years of education $\mathbf{Z}_2 =$ (ED) is assumed to be correlated with the individual effect.

In column 4, the pre-test unrestricted Hausman Taylor estimator using $\mathbf{X}_1^*$ as instruments of $\mathbf{Z}_2 =$ (ED) and including the average over time of endogenous time varying variables $\mathbf{X}_2^* =$ (EXP, WKS, SOUTH, SMSA, MS, OCC, UNION) leads to larger estimates for $SOUTH_{it}$ and for IND_{it} than with the consistent within estimator. The return to schooling estimates (0.049) is nearly the same than the Between estimator (0.051). However, it is no longer statistically significant, with a 95% confidence interval equal to $[-0.05, 0.15]$, to be compared with the Between estimator

$[0.045, 0.057]$.

This lack of precision is partly due to the fact that $\mathbf{X}_1^*$ consists of three weak instruments explaining altogether only 12% of the variance of years of education. The pre-test estimator leads to choose highly exogenous instruments X_1. If the variables Z_2 are strongly correlated with unobservable heterogeneity α_i, then the exogenous instruments X_1 will be weakly correlated with Z_2. Because the pre-test estimator leads to choose weak instruments, the Mundlak estimator is likely to be as reliable as the Hausman-Taylor estimator for this data set.

3.2. Repeated Between Estimator

We present this estimator because repeating T times the time invariant observations turns out to be the major component of the increase of the t-statistic of time-invariant variables using panel data for several other estimators. However, Kelejian and Stephan (1983) argue *"the effect of a random component can only be averaged out if the sample increases in the direction of that random (time or individual) component"*. By contrast, Oaxaca and Gleiser (2003) assume that the consistency of the estimator of a parameter of a time-invariant explanatory variable *"depends on the time series observations approaching infinity"*. The estimated parameters of time-invariant variables are the same with the repeated between and the between estimator ($\gamma^B = \gamma^{RB}$). The between estimator of the standard error with observations repeated T times is equal

to (the subscript for this estimator is RPB):

$$\widehat{\sigma}_{\widehat{\gamma}}^{RB} = \frac{\sqrt{\frac{T \cdot SSE_B}{NT-k-g-1}}}{\sqrt{T \cdot CSS(z_j)}} \frac{1}{\sqrt{1-R_A^2(z_j)}} = \sqrt{\frac{N-k-g-1}{NT-k-g-1}} \widehat{\sigma}_{\widehat{\gamma}}^{B} = \frac{\widehat{\sigma}_{\widehat{\gamma}}^{B}}{2.671}.$$

Inference uses $NT-k-g-1$ degrees of freedom instead of $N-k-g-1$ in the between estimator. The coefficient of determination of the auxiliary regression $R_A^2(z_j)$ does not change in the between or repeated between samples. The estimated standard error of the estimated parameter of the time invariant variables is divided by $\sqrt{\frac{4165-13}{595-13}} = 2.67$ in table 1. As the parameter estimate is the same as γ^B, the repeated between t^{RB}-statistic amounts to multiply the between t^B statistics by the following factor:

$$t^{RB} = \frac{\widehat{\gamma^B}}{\widehat{\sigma}_{\widehat{\gamma^B}}^{RB}} = \sqrt{\frac{NT-k-g-1}{N-k-g-1}} \frac{\widehat{\gamma^B}}{\widehat{\sigma}_{\widehat{\gamma}}^{B}} = \sqrt{\frac{NT-k-g-1}{N-k-g-1}} t^B$$

When N is large, the t-statistic of the repeated between model is multiplied by around $\sqrt{T}$ (say by 2 when $T = 4$ and by 5 when $T = 25$) with respect to the between model.

3.3. Pooled OLS Unrestricted Estimator

The Pooled OLS ($\theta = 1$) estimator ignores the random effects. It leads to the same parameter estimates as the Mundlak estimator, but not the same standard error estimates. The reason why we analyse the Pooled OLS estimator for time-invariant explanatory variables is that it is still used by some applied researchers. When test-

ing does not reject the existence of random effects and when the Hausman test leads to reject the random effect models with respect to the "correlated random effects" model, some applied researchers present Pooled OLS estimates with time invariant explanatory variables, on the ground that the required within estimates wipes out time-invariant variables. For example, Oaxaca and Geisler [2003] evaluate the consistency of the OLS estimator of a parameter of a time-invariant explanatory variable.

The OLS estimator uses the irrelevant $NT-k-g-1$ degrees of freedom for time-invariant variables. The estimated standard error is larger than for the repeated between estimator, because the $RMSE$ of the OLS estimator is larger since it includes the sum of squares of the error of the within model: $SSE_{OLS} = T \cdot SSE_B + SSE_W$ (the last equality presents table 1 values):

$$\begin{aligned}
\widehat{\sigma}_{\widehat{\gamma}}^{OLS} &= \frac{\sqrt{\frac{T \cdot SSE_B + SSE_W}{NT-k-g-1}}}{\sqrt{T \cdot CSS(z_i)}} \frac{1}{\sqrt{1-R_A^2(z_j)}} \\
&= \sqrt{1+\frac{SSE_W}{T \cdot SSE_B}} \sqrt{\frac{N-k-g-1}{NT-k-g-1}} \widehat{\sigma}_{\widehat{\gamma}}^{B}. \\
&= \sqrt{1+\frac{84.11480}{7 \cdot 42.07257}} = 1.1338 \cdot \frac{\widehat{\sigma}_{\widehat{\gamma}}^{B}}{2.671} = \frac{\widehat{\sigma}_{\widehat{\gamma}}^{B}}{2.33}.
\end{aligned}$$

$R_A^2(z_j)$ is the determination coefficient of the auxiliary regression. It does not change with respect to the between, repeated between and OLS estimators. In table 1 example, as $T=7$ is small, the OLS t^{OLS} statistics (equal to 2.33 times the between t^B statistics) is close to the repeated between t^{RB} statistics (equal to 2.67 times the between t^B statistics). However, as the time-invariant variables do not explain any

variance of the within transformed explained variable, taking into account SSE_W is irrelevant for doing inference on time-invariant variables.T

4. The Omitted Variable Bias of the Restricted Models

4.1. Restricted Random Effect Estimator

Both, the restricted random effects and the Mundlak-Krishnakumar estimator use the same weight $\widehat{\theta}$ for computing quasi demeaned variables (the second equality refers to our example) using the Swamy and Arora method (1972):

$$y_{it} - y_{i.} + \widehat{\theta} y_{i.} \text{ with } \widehat{\theta} = \frac{RMSE_W}{\sqrt{T} \cdot RMSE_B} = \frac{0.14071}{\sqrt{7} \cdot 0.26887} = 0.1978.$$

The between regression includes the same variables (the average-over-time of all explanatory variables. Hence, the root mean squared error of the between estimator $RMSE_B$ is the same in both models. The restricted random effect model faces an omitted variable bias when one rejects the null hypothesis $\widehat{\beta}_B = \widehat{\beta}_W$, for at least one time-varying variable. In the wages example, the bias with the random effect model is (γ_{GLS} denotes the Mundlak GLS estimator whereas γ_{RGLS} denotes the Mundlak restricted GLS estimator):

$$\widehat{\gamma}_{RGLS} = \widehat{\gamma}_B + \sum_{j=1}^{j=k} \left(\widehat{\beta}_B - \widehat{\beta}_W \right) \widehat{\beta}_{\widehat{\theta} x_{i.} / \widehat{\theta} z_{ii}} = 0.06966 + 0.03044 = 0.10$$

with $\beta_{x_{i.}/z_i}$ given by the auxiliary regression using OLS on quasi demeaned variables (table 2, column 1):

$$\widehat{\theta}x_{i.} = \beta_{x_{i.}/z_i}\widehat{\theta}z_i + \beta_{x_{i.}/x_{it}}\left(\widehat{\theta}x_{i.} + x_{it} - x_{i.}\right) + \widehat{\theta}\beta_0 + \varepsilon_{it}.$$

The omitted variable bias on the parameters of the time-invariant variables is large for ED (the return of education parameter nearly doubled: it is multiplied by $1.93\beta^B$) and it is smaller for BLK ($1.33\beta^B$)and FEM ($1.06\beta^B$). The omitted variable bias on the t-statistic is $1.87t^B$ for ED, $1.03t^B$ for BLK, $1.14t^B$ for FEM.

4.2. Restricted Hausman-Taylor Estimator

In table 2, the column HT($\mathbf{X}_1^{**}$) presents Baltagi and Khanti-Akom (1990) estimates using the standard Hausman Taylor procedure, with the set of instruments $\mathbf{X}_1^{**}$ and omitting the average-over-time of the endogenous time-varying variables $\mathbf{X}_2^{**}$. The return to schooling is 0.137 with a confidence interval $[0.116, 0.158]$.

The rise of the return to schooling with respect to the between estimate is only partly due to "instrumental variables" correcting the endogeneity of schooling with the individual effect. Around half of the rise is given by the omission of time-invariant variables which corrects the endogeneity of time varying regressors with the individual effect. The other component of the rise of the parameter is related to the *strong* instrument "blue collar occupation" which is strongly correlated with the number of years of education (its correlation coefficient around 0.45). This is likely to drive the

results emphasized by Boumahdi and Thomas [2006] for their measure of instruments relevance on this data set. They find that the gain in efficiency across the various sets of instruments seems to be in the education variable. However, "blue collar occupation" is endogenous, with a relatively large difference between its estimated parameters $\widehat{\beta}_B - \widehat{\beta}_W$ with a relatively large t statistics ($t = 4.57$) in the Mundlak equation. Unobservable ability affect primarily the number of years of education, which determine the outcome of a "blue collar" occupation.

Let us finally remark that Angrist and Pischke (2009, p.64-68) do not recommend to include the time varying variable OCC_{it} (occupation) when investigating the causal link between education and wages, because it is an outcome variable which occurs at a later stage than education.

4.3. Two-stage Restricted Between

A common practice consists of a two-stage *restricted* between (denoted $II - RB$) estimator of time-invariant variables, with the restriction $\widehat{\beta} = \widehat{\beta}_W$ for the average-over-time of time-varying variables in a between regression. Let

$$\widehat{d}_i = y_{i.} - \mathbf{X}_{i.}\widehat{\beta}_W = \left(\mathbf{B} - \mathbf{X}_{i.}\left(\mathbf{X}'_{it}\mathbf{W}\mathbf{X}_{it}\right)^{-1}\mathbf{X}'_{it}\mathbf{W}\right) y_{it}$$

be the N vector of group means estimated from the within-groups residuals. Expanding this expression leads to:

$$\widehat{d}_i = \mathbf{Z}_{1i.}\gamma_1 + \mathbf{Z}_{2i.}\left(\gamma_2 + \phi_2\right) + \alpha_i^M + \left(\mathbf{B} - \mathbf{X}_{i.}\left(\mathbf{X}'_{it}\mathbf{W}\mathbf{X}_{it}\right)^{-1}\mathbf{X}'_{it}\mathbf{W}\right)\varepsilon_{it} + \mathbf{X}_{i.}\left(\widehat{\beta}_B - \widehat{\beta}_W\right) \tag{4.1}$$

Treating the last two terms as an unobservable mean zero disturbance, consider estimating γ from the above equation using N observations, without taking into account that the disturbances of the restricted between includes this omitted term: $\mathbf{X}_{i.}\left(\widehat{\beta}_B - \widehat{\beta}_W\right)$. There is then an omitted variable bias on the parameter of the time-invariant variable:

$$\widehat{\gamma}_{II-RB} = \widehat{\gamma}_B + \sum_{j=1}^{j=k}\left(\widehat{\beta}_B - \widehat{\beta}_W\right)\widehat{\beta}_{x_{j.}/z_i}$$

with $\beta_{x_{i.}/z_i}$ estimated using the following auxiliary regressions in the between dimension:

$$x_{i.} = \beta_{x_{i.}/z_i} z_i + \beta_{0,x_{i.}/z_i} + \varepsilon_{i,x_{i.}/z_i}.$$

In this two stage restricted between estimator, the omitted variable bias on the estimated standard error of the estimated parameter of the time-invariant variables leads to differences with respect to the between estimator:

$$\begin{aligned}
\widehat{\sigma}^2_{\widehat{\gamma}_{II-RB}} &= \sqrt{\frac{SSE_{II-RB}}{N-g-1}}\frac{1}{\sqrt{CSS(z_i)}\sqrt{1-R^2_{A,RB}\left(z_j\right)}} \neq \\
\widehat{\sigma}_{\widehat{\gamma}_B} &= \sqrt{\frac{SSE_B}{N-k-g-1}}\frac{1}{\sqrt{CSS(z_i)}\sqrt{1-R^2_A\left(z_j\right)}}
\end{aligned}$$

First, the sum of squares of errors SSE_{II-RB} is larger than the SSE_B because the model constrains the parameters of the time-varying variables to their within estimate which may not minimize the between sum of squares of errors. However, this increase of the estimated standard error may be offset for two reasons due to the fact the averages of k time-varying variables are on the left hand side of the equation. First, the degrees of freedom increase by k. This decreases the root mean squared error. Second, the variance inflation factor decreases because $R^2_{A,RB}(z_j) < R^2_A(z_j)$. $R^2_A(z_j)$ is the coefficient of determination of an auxiliary regression where the time-invariant variable is correlated with the other $g-1$ time-invariant explanatory variables on the right hand side of the equation. In the between estimator, $R^2_A(z_j)$ is the coefficient of determination of an auxiliary regression where the time-invariant variable is correlated with the other $g-1$ time-invariant explanatory variables and k averages of time-varying explanatory variables. As the number of explanatory variables increases by k, one has $R^2_{A,RB}(z_j) < R^2_A(z_j)$.

With respect to the estimated parameter of the non restricted between, the estimated parameter of the restricted between are multiplied by 0.89 for BLK, by 1.23 for ED and by 1.42 for FEM. With respect to the t-statistics of the non restricted between, the t-statistics of the restricted between are multiplied by 0.77 for BLK, by 1.47 for ED and by 1.86 for FEM.

Oaxaca and Geisler (2003) propose an alternative correction of the omitted variable bias of the two-stage restricted between approach than the Mundlak-Krishnakumar

estimator. They estimate the *repeated* restricted between with a two stage generalized least square estimator using a covariance matrix that takes into account that the disturbances of the restricted between includes this omitted term: $\mathbf{X}_{i.}\left(\widehat{\beta}_B - \widehat{\beta}_W\right)$. The first drawback of the Oaxaca and Geisler (2003) estimator is that the second step between should not be estimated using observations repeated T times. The second drawback is that it is simpler to include explicitly the variables $\mathbf{X}_{i.}$ in the Mundlak-Krishnakumar estimator, than to compute a two stages GLS estimator with a specific covariance matrix.

4.4. Three-stage FEVD Unrestricted and Restricted Estimator

The FEVD estimator adds a third stage to the previous two stage estimator. It is assumed that time-invariant variables are not correlated with the random individual effects: $E(\alpha_i \mid \mathbf{Z}_i) = \phi = 0$ and three sets of alternative assumptions can be dealt with ($E(\alpha_i \mid \mathbf{X}_{it}) = \pi = \mathbf{0}$, or $E(\alpha_i \mid \mathbf{X}_{2it}) = \pi_2 \neq \mathbf{0}$ or $E(\alpha_i \mid \mathbf{X}_{it}) = \pi \neq \mathbf{0}$ (all time-varying variables are endogenous). Plümper and Troeger (2007) use a restricted FEVD estimator with a second stage and a third stage omitting variables $\mathbf{X}_{2i.}$ or $\mathbf{X}_{i.}$ which is consistent with the assumption $E(\alpha_i \mid \mathbf{X}_{it}) = \pi = \mathbf{0}$. Note that if α_i is random instead of being fixed, the "usual" GLS estimator is consistent, which is not the case of the first stage "within-groups" estimator for the FEVD.

Let us denote $\widehat{\varepsilon}_{i,B}$ the between residuals of stage II restricted between:

$$\widehat{\varepsilon}_{i,B} = \widehat{\alpha}_i^M + \left(\mathbf{B} - \mathbf{X}_{i.}\left(\mathbf{X}'_{it}\mathbf{W}\mathbf{X}_{it}\right)^{-1}\mathbf{X}'_{it}\mathbf{W}\right)\widehat{\varepsilon}_{it} + \mathbf{X}_{i.}\left(\widehat{\beta}_B - \widehat{\beta}_W\right) = y_{i.} - \mathbf{X}_{i.}\widehat{\beta}_W - \mathbf{Z}_{i.}\widehat{\gamma}_{II-RB}$$

The third stage of the FEVD estimator is an OLS regression which includes the residual of the second stage repeated between regression denoted $\widehat{\varepsilon}_{i,B}$ with a parameter δ estimated or constrained to one.

$$y_{it} = \mathbf{X}_{it}\beta + \mathbf{Z}_i\gamma + \mathbf{X}_{i.}\pi + \widehat{\varepsilon}_{i,B} \cdot \delta + \eta_{it,III}.$$

Expanding this expression yields:

$$y_{it} - y_{i.} = \mathbf{W}y_{it} = (\mathbf{X}_{it} - \mathbf{X}_{i.})\,\beta - \mathbf{X}_{i.}\left(\beta - \delta\widehat{\beta}_W\right) + \mathbf{Z}_i\,(\gamma - \delta\widehat{\gamma}_{II-RB}) + \eta_{it,III} \quad (4.2)$$

$\mathbf{W}y_{it}$ is orthogonal in the sample to all time-invariant variables, so that OLS estimates are: $\widehat{\gamma} = \widehat{\delta}\widehat{\gamma}_B$, $\widehat{\pi} = \widehat{\delta}\left(\widehat{\beta}_B - \widehat{\beta}_W\right)$, $\widehat{\beta} = \widehat{\delta}\beta_W$, $\widehat{\beta} = \widehat{\beta}_W$, $\widehat{\delta} = 1$, and $\widehat{\eta}_{it,III} = \widehat{\varepsilon}_{it} - \widehat{\varepsilon}_{i.}$. The residuals of the third step regression are exactly the within-groups regression residuals. The only change of the three stage FEVD estimator with respect to the two stage estimator is related to the estimated standard error of the parameters γ of time invariant variables $\mathbf{Z}_i$. The FEVD estimator of the standard error of the estimated parameter of a time-invariant variable amounts to substitute the mean squared error in the between dimension ($MSE_B = SSE_B/\left(N - k - g - 1\right)$) by the mean squared error in the within dimension ($MSE_W = SSE_W/\left(NT - N - k\right)$). The estimated parameters $\widehat{\gamma}_{II-RB}$ are related to the projection in the between subspace of observations, but their estimated standard errors are related to the projection in the within subspace of observations, which is orthogonal to the between subspace. This

contradicts the Frisch-Waugh-Lovell theorem: the same orthogonal projection matrix has to be used in the parameter estimate and in the estimator of its variance.

The Frisch-Waugh-Lovell decomposition of the Between and the Within dimension holds with any of these assumptions: effects are fixed or random, or the time-varying variables are correlated or not with the random effect do not matter at all. As the residuals in the between dimension are excluded from the computation of the variance of the parameters, the potential correlations of the time-invariant variable with the individual random effect (which are in the between dimension) are excluded from the computation of the FEVD variance of the parameters. As well, adding a large number of time-invariant variables in the regression, in particular when they are near-multicollinear with the time-invariant variable of interest, changes the estimated parameter, but does not increase the FEVD estimated standard error of the estimated parameter of the time-invariant variable of interest. So the FEVD wipes out near-multicollinearity problems in the time-invariant dimension, which is also "practical" for reducing estimated standard errors.

This analysis explains why Kristensen and Wawro 2007 (footnote p.22) found that the FEVD estimated standard errors were relatively "too small" using Monte Carlo simulations with respect to other estimators. The FEVD estimator of the standard error is (the last equality refers to the unrestricted FEVD case, table 1 column 6):

$$\widehat{\sigma}_{\widehat{\gamma}_{FEVD}} = \frac{\sqrt{\frac{SSE_W}{NT-N-k}}}{\sqrt{T \cdot CSS(z_i)}} \frac{1}{\sqrt{1-R_A^2(z_j)}} = \sqrt{\frac{N-k-g-1}{NT-N-k}} \sqrt{\frac{SSE_W}{T \cdot SSE_B}} \widehat{\sigma}_{\widehat{\gamma}}^B$$

$$= \sqrt{\frac{595-13}{4165-595-9}}\sqrt{\frac{84.11480}{7\cdot 42.07257}}\,\widehat{\sigma}^{B}_{\widehat{\gamma}} = \frac{1}{2.4736}\cdot\frac{1}{1.8712}\cdot\widehat{\sigma}^{B}_{\widehat{\gamma}} = \frac{\widehat{\sigma}^{B}_{\widehat{\gamma}}}{4.6286}$$

The FEVD estimated standard error of a time-invariant variable is usually biased downwards for two reasons:

- It uses $NT-N-k$ degrees of freedom (with N the number of individuals, k the number of time-varying explanatory variables, with T the number of periods) instead of $N-k-g-1$ degrees of freedom.

- It multiplies the repeated between estimator of the standard error by a positive factor $\sqrt{\frac{SSE_W}{T\cdot SSE_B}}$ which can be much smaller than one when T increases.

The combination of the potential omitted variable bias of the parameter estimate of the time invariant variable ED in the restricted between (step II) and of the bias of the estimated standard error implies a t-statistics equal to $t_{RFEVD} = 5.69\cdot t_B$ times the t-statistics of the Mundlak model for the restricted FEVD (table 2, column 6). For the unrestricted FEVD (table 1, column 6), only the bias of the estimated standard error matters, so that the $t_{FEVD} = 4.62\cdot t_B$ times the t-statistics of the Mundlak model for ED. For the time invariant BLK, $t_{RFEVD} = 3.6\cdot t_B$ and for FEM, $t_{RFEVD} = 8.6\cdot t_B$. Note that when T corresponds to an aggregate annual time series ($T=25, \sqrt{T}=5$), the increase of t-statistics is effect is likely to double with respect to the wages example ($T=7, \sqrt{T}=2.6$). Mitze [2009] also finds that the FEVD estimator tends to have a smaller root mean square error (rmse) than both Hausman-Taylor models with perfect and imperfect knowledge about the underlying correlation

between right hand side variables and residual term. Nonetheless, concluding that FEVD is "more efficient" is misleading, because the computation of the estimated standard errors of estimated parameters contradicts Frisch-Waugh-Lovell theorem.

5. Conclusion

Our theoretical and empirical investigation on inference on time-invariant variables shows that a pre-test estimator based upon the Mundlak-Krisnakumar and a modified Hausman-Taylor estimator should be used. Furthermore, the procedures are already programmed and available in all econometric softwares. The first stage consists of the usual random effects GLS estimator including all the variables $\mathbf{X}_{i.}$. The second stage uses instrumental variables $\mathbf{X}_{1i.}$ keeping the $\mathbf{X}_{2i.}$ as explanatory variables, using instrumental variables estimator or the Hausman-Taylor estimator procedures.

An example shows first that a time-invariant variable is not statistically significant for some estimators and highly statistically significant with other estimators.

The Mundlak-Krishnakumar regression reports within estimates and between estimates, with tests of the null hypothesis $\widehat{\beta}_B - \widehat{\beta}_W = 0$ for each time-varying explanatory variable. In the case where at least one (but not all) $x_{i.}$ have estimated parameter $\widehat{\beta}_B - \widehat{\beta}_W$ small and not significantly different from zero, the Mundlak estimator suggests which of the $x_{i.}$ is exogenous. Then, these exogenous $x_{i.}$ are the ones to be used as instrumental variables in a variant of the Hausman and Taylor [1981] instrumental variables estimator, including the average over time of endogenous time-varying vari-

ables. Using this pre-test estimator, one does not rely on subjective priors for deciding which time-varying variables are or are not endogenous in the Hausman and Taylor estimator.

One may try further Chamberlain (1984) estimators where the correlation of explanatory variables with the random individual effects is not only contemporaneous, but can also be related to leads and lags of explanatory variables.

It is not necessarily the failure of a model that the outcome of the tests is: $\left|\widehat{\beta}_B\right| > \widehat{\beta}_W = 0$. Some variables have zero or small correlation coefficients in the within subspace, and large correlation coefficients in the between dimension, just because it is the statistical information included in these rarely changing or time-invariant variables during the period of observations in available data sets. If we turn to the case of rarely changing variables instead of time-invariant variables, these variables are characterized by small within correlation coefficients with the explained variable and with all other explanatory variables. By contrast, they may have high between correlation coefficients with the explained variable and with all other explanatory variables. The variables are typically such that the null hypothesis $\widehat{\beta}_B - \widehat{\beta}_W = 0$ is rejected. The Mundlak estimator provides directly the t-test related to this null hypothesis. It is of particular interest for those variables, which are likely to be endogenous and significant variables in the Mundlak approach. If ever they do not have large between correlation coefficients, then the joint hypothesis $\widehat{\beta}_B = \widehat{\beta}_W = 0$ may not be rejected.

References

[1] Angrist J.D. and Pischke J.-S. [2009]. *Mostly Harmless Econometrics.* Princeton University Press, Princeton.

[2] Baltagi B. H. [2008]. *Econometric Analysis of Panel Data.* Fourth Edition, Wiley, Chichester.

[3] Baltagi B. H. [2009]. *A Companion to Econometric Analysis of Panel Data.* Wiley, Chichester.

[4] Baltagi B. H. and S. Khanti-Akom [1990]. On efficient estimation with panel data: an empirical comparison of instrumental variables estimators, *Journal of Applied Econometrics,* 5 pp. 401-406.

[5] Baltagi B. H., Bresson G. and Pirotte A. [2003]. Fixed Effects, Random Effects or Hausman-Taylor? A Pre-test Estimator. *Economics Letters.* 79 pp. 361-369.

[6] Beck N. and Katz J. [2001]. Throwing out the baby with the bath water: a comment on Green, Kim and Yoon. *International Organization* 55(2) pp. 487-495.

[7] Blaydes L. [2006]. Rewarding Impatience: A response to Goodrich. *International Organization*, 60, pp. 515-525.

[8] Boumahdi R. and Thomas A. [2006]. Instrument relevance and efficient estimation with panel data. *Economics Letters*, 93, pp. 305-310.

[9] Breusch T., Ward M.B., Nguyen H. and Kompas T. [2010]. On the fixed effects vector decomposition. Munich Personnal Repec Archive Paper 21452.

[10] Card D. [1999]. The Causal Effect of Education on Earnings. in Handbook of Labor Economics. Volume 3A. ed. by O. Ashenfelter and D. Card. North Holland. Amsterdam and New York.

[11] Cameron A.C. and Trivedi P.K. [2009]. *Microeconometrics Using Stata*. Stata Press, College Station, Texas.

[12] Chamberlain G. [1982]. Multivariate Regression Models for Panel Data. *Econometrica.* 18. pp. 5-46.

[13] Frisch R. and Waugh F.V. [1933]. Partial time regression as compared with individual trends. *Econometrica*, 1, pp. 387-401.

[14] Goodrich B. [2006]. A Comment on "Rewarding Impatience". *International Organization*, 60, pp. 499-513.

[15] Green D.P., Kim S.Y. and Yoon D.H. [2001]. Dirty Pool. *International Organization* 55(2) pp. 441-468.

[16] Greene W. [2008]. *Econometric Analysis.* Sixth edition. Pearson International Edition, New Jersey.

[17] Greene W. [2010]. Fixed Effects Vector Decomposition: A Magical Solution to the Problem of Time Invariant Variables in Fixed Effects Models? Working Paper, New York University.

[18] Hausman J.A. [1978]. Specification Tests in Econometrics. *Econometrica.* 46. pp. 1251-1271.

[19] Hausman J.A. and Taylor W.E. [1981]. Panel data and unobservable individual effects. *Econometrica.* 49. pp. 1377-1398.

[20] Hsiao C. [2003]. *Analysis of Panel Data.* 2nd edition. Cambridge University Press. Cambridge.

[21] Johnston J. and DiNardo J. [1997]. *Econometric Methods.* 4th edition. McGraw Hill.

[22] Kelejian H.K. and Stephan S.W. [1983]. Inference in Random Coefficient Panel Data Models: a Correction and Clarification of the Literature. *International Economic Review* 24(1) pp. 249-254.

[23] Krishnakumar J. [2006]. Time Invariant Variables and Panel Data Models: A Generalised Frisch-Waugh Theorem and its Implications. in B. Baltagi (ed.), *Panel Data Econometrics: Theoretical Contributions and Empirical Applications,* published in the series "Contributions to Economic Analysis", North Holland (Elsevier Science), Amsterdam, Chapter 5, 119-132.

[24] Kristensen and Wawro [2007]. On the Use of Fixed Effects Estimators for Time Series Cross Section Data. Princeton University.

[25] Lovell M.C. [1963]. Seasonal adjustment of economic time series and multiple regression analysis. *Journal of the American Statistical Association.* 58 pp. 993-1010.

[26] Mitze T. [2009]. "Endogeneity in Panel Data Models with Time-Varying and Time-Fixed Regressors: To IV or not IV?," Ruhr Economic Papers 0083, Rheinisch-Westfälisches Institut für Wirtschaftsforschung, Ruhr-Universität Bochum, Universität Dortmund, Universität Duisburg-Essen.

[27] Mundlak Y. [1978]. On the pooling of time series and cross section data. *Econometrica.* 46. pp. 69-85.

[28] Oaxaca R.L. and Geisler I. [2003]. Fixed effects models with time invariant variables: a theoretical note. *Economics Letters.* 80 pp. 373-377.

[29] Plumper D. and Tröger V. [2007]. Efficient Estimation of Time Invariant and Rarely Changing Variables in Finite Sample Panel Analyses with Unit Fixed Effects. *Political Analysis.* 15(2). pp. 124-139.

[30] Stewart G.W. [1987]. Collinearity and Least Squares Regression, *Statistical Science* 2(1). pp. 68-100.

[31] Swamy P.A.V.B and Arora S.S. [1972]. The exact finite sample properties of the estimators of coefficients in the error components regression models. *Econometrica.* 40. pp. 261-275.

Table 1. Dependent variable: log wage: Unrestricted Models.

		Within	Between	GLS Mundlak	Pre-Test HT(X1*)	Repeated Between	OLS	FEVD
Constant		-	5.12 (0.203)	5.12 (0.203)	5.18 (0.861)	5.12 (0.076)	5.12 (0.08)	-
EXP	(*it*)	0.113 (0.003)	-	0.113 (0.003)	0.114 (.00226)	-	0.113 (0.005)	0.113 (0.003)
EXP2	(*it*)	-0.00042 (0.00005)	-	-0.00042 (0.00005)	**-0.000449 (0.0000485)**	-	-0.00042 (0.00011)	-0.00042 (0.00005)
WKS	(*it*)	0.00084 (0.0006)	-	0.00084 (0.0006)	0.000802 (0.0006)	-	0.00084 (0.0011)	0.00084 (0.0006)
SOUTH	(*it*)	-0.0019 (0.035)	-	-0.0019 (0.035)	**-0.0388 (0.0272)**	-	-0.0019 (0.0680)	-0.0019 (0.035)
SMSA	(*it*)	-0.0425 (0.0194)	-	-0.0425 (0.0194)	-0.0405 (0.0193)	-	-0.0425 (0.0385)	-0.0425 (0.0194)
MS	(*it*)	-0.0297 (0.019)	-	-0.0297 (0.019)	-.0307 (.0189)	-	-0.0297 (0.037)	-0.0297 (0.019)
OCC	(*it*)	-0.0215 (0.014)	-	-0.0215 (0.014)	-0.0222 (0.0137)	-	-0.0215 (0.027)	-0.0215 (0.014)
IND	(*it*)	0.0192 (0.0154)	-	0.0192 (0.0154)	**0.0297 (0.0144)**	-	0.0192 (0.031)	0.0192 (0.0154)
UNION	(*it*)	0.0328 (0.0149)	-	0.0328 (0.0149)	0.0325 (0.0148)	-	0.0328 (0.0296)	0.0328 (0.0149)
EXP(*i*)	(*i.*)	-	0.0319 (0.0048)	-0.0813 (0.0054)	-0.0875 (.00272)	0.0319 (0.0018)	-0.0813 (0.0053)	-0.0813 (0.0012)
EXP2(i)	(*i.*)	-	-0.00057 (0.00011)	**-0.00015 (0.00012)**	-	-0.00057 (0.00004)	-0.00015 (0.00012)	-0.00015 (0.000026)
WKS(i)	(*i.*)	-	0.0092 (0.0036)	0.00835 (0.0037)	0.00874 (0.00448)	0.0092 (0.0013)	0.00835 (0.0019)	0.00835 (0.0080)
SOUTH(i)	(*i.*)	-	-0.057 (0.026)	**-0.055 (0.043)**	-	-0.057 (0.0097)	-0.055 (0.069)	-0.055 (0.0091)
SMSA(i)	(*i.*)	-	0.176 (0.026)	0.218 (0.032)	0.218 (0.0386)	0.176 (0.0096)	0.218 (0.040)	0.218 (0.0069)
MS(i)	(*i.*)	-	0.115 (0.048)	0.145 (0.051)	0.151 (0.0542)	0.115 (0.0168)	0.145 (0.042)	0.145 (0.011)
OCC(i)	(*i.*)	-	-0.168 (0.034)	-0.146 (0.0365)	-0.152 (0.169)	-0.168 (0.0126)	-0.146 (0.031)	-0.146 (0.0079)
IND(i)	(*i.*)	-	0.058 (0.026)	**0.0387 (0.0298)**	-	0.058 (0.0095)	0.0387 (0.032)	0.0387 (0.0064)
UNION(i)	(*i.*)	-	0.109 (0.029)	0.0763 (0.0328)	0.0835 (0.0381)	0.109 (0.0109)	0.0763 (0.0321)	0.0763 (0.0071)
FEM(i)	(*i*)	-	-0.317 (0.055)	-0.317 (0.055)	-0.319 (0.0601)	-0.317 (0.0205)	-0.317 (0.023)	-0.317 (0.012)
BLK(i)	(*i*)	-	-0.158 (0.045)	-0.158 (0.045)	-0.165 (0.0499)	-0.158 (0.0168)	-0.158 (0.019**)**	-0.158 (0.0097)
ED(i)	(*i*)	-	0.0515 (0.00555	0.0515 (0.00555)	**0.0489 (0.0469)**	0.0515 (0.00208)	0.0515 (0.00235)	0.0515 (0.0012)
D.of F.		3561	582	3561 and 582	3561 and 582	4152	4143	3561
RMSE		0.141	0.269	0.152		0.359	0.152	0.152
R2		0.907	0.544	0.645		0.544	0.645	0.645

Z1=(FEM, BLK), X1*=(SOUTH, IND, EXP2) using pre-test selection.

Table 2. Dependent variable: log wage: Restricted Models.

		R-Between (step II)	R-GLS	R-HT (X1**	Repeated R-Between	R-OLS	R-FEVD
Constant		5.92 (0.061)	4.264 (0.098)	2.913 (0.283)	5.92 (0.023)	5.25 (0.07)	-
EXP	(*it*)	-	0.082 (0.003)	0.113 (0.019)	-	0.040 (0.002)	0.113 (0.003)
EXP2	(*it*)	-	-0.0008 (0.00006)	-0.000419 (0.000055)	-	-0.0007 (0.00005)	-0.00042 (0.00005)
WKS	(*it*)	-	0.00084 (0.0008)	0.00084 (0.0006)	-	0.0042 (0.0011)	0.00084 (0.0006)
SOUTH	(*it*)	-	-0.0017 (0.027)	-0.0074 (0.032)	-	-0.0556 (0.012)	-0.0019 (0.035)
SMSA	(*it*)	-	-0.014 (0.020)	-0.0418 (0.0189)	-	0.151 (0.120)	-0.0425 (0.0194)
MS	(*it*)	-	-0.075 (0.023)	-0.0298 (0.019)	-	0.048 (0.020)	-0.0297 (0.019)
OCC	(*it*)	-	-0.050 (0.017)	-0.0207 (0.014)	-	-0.140 (0.014)	-0.0215 (0.014)
IND	(*it*)	-	0.004 (0.017)	0.0136 (0.0152)	-	0.046 (0.011)	0.0192 (0.0154)
UNION	(*it*)	-	0.063 (0.017)	0.0328 (0.0149)	-	0.092 (0.013)	0.0328 (0.0149)
FEM(i)	(*i*)	-0.449 (0.041)	-0.339 (0.051)	-0.131 (0.127)	-0.449 (0.016)	-0.367 (0.025)	-0.449 (0.009)
BLK(i)	(*i*)	-0.141 (0.051)	-0.210 (0.058)	-0.285 (0.155)	-0.141 (0.019)	-0.167 (0.022)	-0.141 (0.011)
ED(i)	(*i*)	0.0635 (0.0046)	0.100 (0.006)	0.137 (0.021)	0.0635 (0.0017)	0.0567 (0.0026)	0.0635 (0.001)
D.of F.		591	4152	3561 and 59	4161	4152	3561
RMSE		0.313	0.1896		0.313	0.349	
R2		0.366	0.428		0.366	0.428	

Z1=(FEM, BLK), X1**=(SOUTH, IND, OCC, SMSA), with t statistics for Mundlak estimation in table 1: OCC(i): **t=-4.01**, SMSA(i): **t=6.77**. R-OLS in Baltagi (2009), table 2.6, p.27. R-GLS in Baltagi (2009) table 2.8, p.28, R-HT (X1**) in Baltagi (2009), table 7.2, p.157.

Table 3: t-statistics of time invariant variables, as a proportion of the t-statistics of the between estimator.

		Between	GLS Mundlak	Pre-Test HT(X1*)	Repeated Between	OLS	FEVD
FEM(i)	t-statistics	-5,8	-5,8	-5,3	-15,5	-13,8	-26,4
	t/t(between)	**1,0**	**1,0**	**0,9**	**2,7**	**2,4**	**4,6**
BLK(i)	t-statistics	-3,5	-3,5	-3,3	-9,4	-8,3	-16,3
	t/t(between)	**1,0**	**1,0**	**0,9**	**2,7**	**2,4**	**4,6**
ED(i)	t-statistics	9,3	9,3	1,0	24,8	21,9	42,9
	t/t(between)	**1,0**	**1,0**	**0,1**	**2,7**	**2,4**	**4,6**
		R-Between	R-GLS	R-HT X1**	RRBetween	R-OLS	R-FEVD
FEM(i)	t-statistics	-11,0	-6,6	-1,0	-28,1	-14,7	-49,9
	t/t(between)	**1,9**	**1,2**	**0,2**	**4,9**	**2,5**	**8,7**
BLK(i)	t-statistics	-2,8	-3,6	-1,8	-7,4	-7,6	-12,8
	t/t(between)	**0,8**	**1,0**	**0,5**	**2,1**	**2,2**	**3,7**
ED(i)	t-statistics	13,8	16,7	6,5	37,4	21,8	63,5
	t/t(between)	**1,5**	**1,8**	**0,7**	**4,0**	**2,4**	**6,8**

Not for publication:

Table 4: Contribution to R2 in **Within** regressions with forward selection.

Variable	Rank	Partial R2	R2
EXP	**1**	**0.6505**	**0.6505**
EXP2	2	0.0060	0.6564
SMSA	3	0.0005	0.6569
UNION	4	0.0004	0.6574
MS	5	0.0002	0.6576
OCC	6	0.0002	0.6578
WKS	7	0.0002	0.6580
IND	8	0.0001	0.6581
SOUTH	9	0.0000	0.6581

Table 5: Contribution to R2 in **Between** regressions with forward selection.

Variable	Rank	Partial R2	R2
mED	1	**0.2129**	0.2129
mFEM	2	**0.1445**	0.3575
mSMSA	3	**0.0581**	0.4156
mEXP	**4**	**0.0420**	0.4576
mEXPSQ	**5**	**0.0298**	0.4874
mBLK	**6**	**0.0156**	0.5030
mOCC	7	0.0101	0.5131
mUNION	8	0.0118	0.5249
mWKS	9	0.0062	0.5311
mIND	10	0.0051	0.5362
mMS	11	0.0043	0.5405
mSOUTH	12	0.0038	0.5443