

# Time-invariant Variables in Panel Data: A Comment

Peter Egger\*

August 28, 2001

## Abstract

Frequently, estimates of time-invariant variables are obtained from a second-step OLS regression of the fixed effects parameters on the invariant variables. This paper argues that the second-step regression should be estimated by GLS rather than OLS.

**Key words:** Panel econometrics

**JEL classification:** C33

## 1 Introduction<sup>1</sup>

In many applications, panel data comprise interesting independent variables, which vary only in a single dimension (e.g. experience in wage regressions, distance or common borders in bilateral trade regressions, etc.). With panel data, there are three opportunities to derive parameter estimates of such variables. First, a random effects model (REM) obtains a direct estimate.

---

\*Austrian Institute of Economic Research, Arsenal, Objekt 20, P.O. Box 91, A-1103 Vienna, Austria; Tel. +43-1-7982601-475; Fax: +43-1-7989386.

<sup>1</sup>I should like to thank Badi Baltagi for helpful comments.

However, this is only consistent in the absence of any correlation between the observables and the panel-specific error term (i.e. with exogenous unobserved effects). Second, if the REM is inconsistent and the unobserved effects are endogenous (which is testable according to Hausman, 1978), one can follow Hausman & Taylor (1981) and apply instrumental variable techniques to overcome this problem (henceforth, I refer to this model as HTM). The consistency of the HTM and the validity of the available instruments are testable by the Hausman & Taylor (1981) over-identification test. Third, one can estimate a fixed effects (least-squares dummy variables, LSDV) model and estimate the parameters of interest from a regression of the estimated fixed effects parameters on the critical variables in a second step. Practitioners frequently prefer the latter approach over the others. It is easy to tackle, since the literature recommends to estimate the regressions in both steps by OLS (compare Hsiao, 1986). This comment concentrates on the latter approach and argues that the proper second-step estimation is GLS rather than OLS. However, it is also relevant for the estimation of the HTM.

The next section motivates the GLS approach. Section 3 gives an empirical example from international trade economics and estimates a standard so-called gravity model. The last section concludes.

## **2 Time-Invariant Variables and GLS in the Two-step Approach**

Without any loss of generality, my analysis concentrates on the two-way panel data model, where one dimension is time and the other is individual effects.

At least one variable in the model be time-invariant.<sup>2</sup> As mentioned above, the literature suggests to obtain an estimate of the time-invariant regressors by running an OLS-regression of the estimated fixed effects' (i.e. individual-specific) parameters from the first-step LSDV regression on the time-invariant regressors of interest in a second step. However, this ignores the "*estimated-parameters-as-dependent-variables*" nature of the problem, since it omits the distribution of the estimated fixed effects' coefficients. Following Saxonhouse (1977), second-step regressions of parameters as dependent variables require a GLS transformation of the data to account for the variance of the estimated parameters and the variance across the parameters. In this spirit, the transformation matrix to the second-step problem under consideration is:

$$\widehat{\Omega}_S = \widehat{\Sigma}_{\mu_i}^2 \otimes I_T + I_{NT} \otimes \widehat{\sigma}_\mu^2, \quad (1)$$

where  $\widehat{\Sigma}_{\mu_i}^2$  is a  $N \times N$  diagonal matrix with the estimated variances of the individual effects' parameters as its entries,  $I_T$  ( $I_{NT}$ ) is a  $T \times T$  ( $NT \times NT$ ) identity matrix,  $\widehat{\sigma}_\mu^2$  is the variance of the estimated fixed effects parameters, i.e. Nerlove's (1971) upper-bound estimate of the cross-sectional variance component (see Baltagi, 1995, for an overview),  $N$  is the number of individuals (cross-sections), and  $T$  is the number of time periods. A balanced data-set is assumed.<sup>3</sup> Both the fixed effects' coefficients from the first-step regression,  $\widehat{\mu}_i$  with  $\mu_i \sim IID(0, \sigma_\mu^2)$ , and the time-invariant variables of interest

---

<sup>2</sup>The arguments are equivalent for  $n$ -way panels, where at least one variable varies in only a single dimension (it needs not necessarily be time-invariant).

<sup>3</sup>However, the analysis can easily be extended to the case of an unbalanced data-set in the first-step regression.

$(z_i)$  must be transformed according to

$$\widehat{\mu}_i^* = \widehat{\Omega}_S^{-1/2} \widehat{\mu}_i; \quad z_i^* = \widehat{\Omega}_S^{-1/2} z_i, \quad (2)$$

where the data are stacked and sorted first by individuals and then by time.<sup>4</sup>

The typical elements of  $\widehat{\mu}_i^*$  and  $z_i^*$  are

$$\widehat{\mu}_i^* = \frac{\widehat{\mu}_i}{(\widehat{\sigma}_{\mu_i}^2 + \widehat{\sigma}_{\mu}^2)^{1/2}}; \quad z_i^* = \frac{z_i}{(\widehat{\sigma}_{\mu_i}^2 + \widehat{\sigma}_{\mu}^2)^{1/2}}. \quad (3)$$

Performing OLS on the transformed model yields the GLS parameter estimates ( $\widehat{\gamma}_z$ ) for the time-invariant variables

$$\widehat{\mu}_i^* = \widehat{\gamma}_z z_i^* + \widehat{\eta}_i, \quad (4)$$

where  $\widehat{\eta}_i$  is the error term, i.e. the estimated remaining individual-specific effect. Principally, one should also transform the data in the HTM's second regression, which obtains an estimate of  $\gamma_z$  from a two-stage least squares instrumental variable (IV) regression. Hausman & Taylor (1981) propose obtaining an estimate of the Within residuals (i.e.  $\widehat{d}_i = \widehat{\mu}_i + \widehat{\varepsilon}_i$ , with the latter as the LSDV residual averaged over time) from a Within regression. In the second stage,  $\widehat{d}_i$  is regressed on the time-invariant variables using the exogenous variables (uncorrelated with the unobserved effect,  $\mu_i$ ) as instruments. Accordingly, the estimated variance of  $\widehat{d}_i$  should be taken into account and the related IV regression should be performed on the data (including the instruments) weighted as suggested in (3).

---

<sup>4</sup>Hence, each fixed effects parameter is repeated  $T$  times.

### 3 An Example: The Gravity Model

The gravity model is a well-established empirical model in international trade economics. The traditional model explains bilateral trade flows (usually exports) by exporter GDP, importer GDP, exporter population, importer population, distance between the trading partners' capitals, a common language dummy (set to one, if two trading partners exhibit the same official language, otherwise zero) and a common border dummy (set to one, if two trading partners face common borders, otherwise zero), compare Bergstrand (1985), Hamilton & Winters (1992) or Baldwin (1994) for three prominent examples. The distance variable and the two dummies do not vary over time, and a LSDV regression in a two-way panel with time and bilateral fixed effects does not provide a first-step parameter estimate of these variables.

I estimate the model using data on bilateral trade flows among the northern EU economies (EU15 excluding Austria, Finland, Greece, Portugal, Spain and Sweden; Belgium and Luxembourg are treated as a single country) over the period 1987-1997. Data on real bilateral exports come from UNO, real GDP and population are from IMF (International Financial Statistics), distance is computed following Egger (2000). All continuous variables are in logs. The panel is balanced and contains 56 bilateral relationships.

> Table 1 <

Table 1 presents the regression results from the first-step two-way LSDV regression with fixed time and bilateral effects and the two alternatives for the second-step regression (OLS and GLS). The Hausman test reveals that a simple REM would not obtain consistent parameter estimates. All coeffi-

cients in the second-step regression exhibit the expected sign. As expected, the standard errors from the second-step OLS regression are higher than their (more appropriate) GLS counterparts. However, in an extreme case the OLS regression in the second step could obtain a significant sign whereas the GLS regression would not.<sup>5</sup>

## 4 Conclusions

Panel data frequently contain variables, which are invariant in one of the available dimensions. There are three possibilities to obtain parameter estimates of such variables. One of them is to run a fixed effects regression in a first step and regress the derived coefficients of the fixed effects on the critical variables (invariant in one dimension) in a second step. So far, the literature suggests to estimate the second-step model by OLS. This paper argues that the second step regression should be GLS in the spirit of Saxonhouse (1977) rather than OLS. The GLS regression takes the variances and the distribution of the estimated coefficients into account whereas simple OLS does not. Similarly, this concept could be applied to the second regression in Hausman & Taylor's (1981) model.

## 5 References

Baldwin, R.E. (1994), *Towards and Integrated Europe* (CEPR, London).

---

<sup>5</sup>In the present application, the coefficient of distance is significant at 5% in the OLS set-up and only at 10% in the GLS regression.

Baltagi, B.H. (1995), *Econometric Analysis of Panel Data* (Wiley, Chichester).

Bergstrand, J. H. (1985), The Gravity Equation in International Trade: Some Microeconomic Foundations and Empirical Evidence, *Review of Economics and Statistics* **67**, pp. 474-481.

Egger, P. (2000), A Note on the Proper Econometric Specification of the Gravity Equation, *Economics Letters* **66**, pp. 25-31.

Hamilton, C.B. and A.L. Winters (1992), Opening up International Trade with Eastern Europe, *Economic Policy* **14**, pp. 77-116.

Hausman, J.A. (1978), Specification Tests in Econometrics, *Econometrica* **46**, pp. 1251-1271.

Hausman, J.A., and W.E. Taylor (1981), Panel Data and Unobservable Individual Effects, *Econometrica* **49**, pp. 1377-1398.

Hsiao, C. (1986), *Analysis of Panel Data* (Cambridge University Press, Cambridge, Mass.).

Nerlove, M. (1971), A Note on Error Components Models, *Econometrica* **39**, 359-382.

Saxonhouse, G.R. (1977), Regressions From Samples Having Different Characteristics, *Review of Economics and Statistics* **59**, pp. 234-237.

**Table 1: First-step and Second-step Gravity Model Regression Results**

First step <sup>a)</sup> : Explaining variables	Fixed effects	Second step <sup>b)</sup> : Explaining variables	OLS	GLS
Exporter GDP	1.3680 *** (0.0802)	Distance	-0.0689 ** (0.0339)	-0.0657 *) (0.0340)
Importer GDP	0.9732 *** (0.0802)	Common language	0.2610 (0.6829)	0.2624 (0.6823)
Exporter population	-3.7025 *** (0.6754)	Common border	1.2228 *** (0.4591)	1.2051 *** (0.4590)
Importer population	1.7513 *** (0.6754)			
Observations	616	Observations	616	616
Cross-sections	56	Cross-sections	56	56
Adjusted R <sup>2</sup>	0.9962	Adjusted R <sup>2</sup>	0.0151	0.0145
Time effects: F(10,546)	2.64			
Bilateral effects: F(55,546)	727.83			
Hausman test: $\chi^2(14)$	32.67			

a) All variables in logs. Dependent variable is log of real bilateral exports. - b) Dependent variable is fixed effects parameters from first stage regression. Distance is defined as the log of miles between capitals (compare Egger, 2000).

Standard errors in parentheses. \*\*\*) significant at 1%; \*\*) significant at 5%; \*) significant at 10%.