London School of Economics
Department of Economics

Vassilis Hajivassiliou
Winter Term 2024

Further Topics in Econometrics (Ec485/Ec518) Problem Set #1 — Panel Data Models

1. Consider the "dummy variable" linear regression model:

$$y_i = i_T \alpha_i + X_i \beta + \epsilon_i$$

for a balanced panel data set with $i = 1, \cdots, N$ cross-sectional units, each observed for T time periods. y_i and ϵ_i are $T \times 1$ vectors, X_i a $T \times k$ matrix, β a $k \times 1$ vector of unknown slope parameters, and there is a different intercept α_i for each unit i. Define N dummy variable vectors (of dimension $NT \times 1$) indicating the different units, e.g., d_i is an $NT \times 1$ vector with typical element

$$d_{it} = \begin{cases} 1 & \text{if observation } it \text{ refers to individual unit } i \\ 0 & \text{otherwise} \end{cases}$$

Stacking the observations for all N units in the standard way and defining the vector α $(N \times 1)$ conformably, gives the matrix formulation:

$$y = D\alpha + X\beta + \epsilon.$$

Define the usual projection matrix $M_d = I - D(D'D)^{-1}D'$. Use standard partitioned-regression results to show that:

(a) The OLS coefficient vector $\hat{\beta}$, known as the "fixed-effects" estimator, can be obtained by regressing $\{y_{it} - \bar{y}_{i.}\}$ on $\{x_{it} - \bar{x}_{i.}\}$, where $y_{i.}$ is the mean of the T observations of i for the y variable, and $x_{i.}$ is the $k \times 1$ vector of means of the x variables over the T observations of i.

(b) The OLS estimates for the N intercepts are:

$$\hat{\alpha}_i = \bar{y}_{i.} - \bar{x}'_{i.}\hat{\beta}.$$

(c) The disturbance variance estimator is:

$$s^2 = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - \hat{\alpha}_i - x'_{it}\hat{\beta})}{NT - N - k}$$

How does this expression differ from the one obtained by regressing $y_{it} - \bar{y}_{i.}$ on $x_{it} - \bar{x}_{i.}$?

2. Consider the linear regression model:

$$y_{it} = X_{it}\beta + \epsilon_{it}, \quad i = 1, \cdots, N, \quad t = 1, \cdots, T_i.$$

for an *unbalanced* panel-data set N cross-sectional units, observed for possibly different numbers T_i of time periods. The regressors are fixed in repeated samples.
The disturbance term is believed to have the *one-factor random-effects* structure:
$\epsilon_{it} = \alpha_i + \nu_{it}$ with α_i independent of ν_{jt} for any i, j, t, $\alpha_i \sim (0, \sigma_\alpha^2)$ i.i.d. over i, and $\nu_{it} \sim (0, \sigma_\nu^2)$ i.i.d. over both i and t. Define

$$\lambda_i \equiv 1 - \sqrt{\frac{\sigma_\nu^2}{T_i\sigma_\alpha^2 + \sigma_\nu^2}}$$

   (a) Show that the transformed error term: $\epsilon_{it}^* \equiv \epsilon_{it} - \lambda_i\bar{\epsilon}_{i\cdot}$ where $\epsilon_{it} \equiv \frac{1}{T_i}\sum_t \epsilon_{it}$, satisfies the Gauss-Markov conditions. Specifically, you should show that ϵ_{it}^* is homoskedastic and serially uncorrelated.

   (b) How would you obtain a consistent estimator for θ_i which you would need to define the feasible GLS estimator?

3. For a balanced panel data set, recall the transformations: $\{z_{it} - z_{i\cdot}\}$ "$Within$" $\{z_{i\cdot}\}$ "$Between$" $\{z_{it} - \theta z_{i\cdot}\}$ "GLS" Running OLS on the "within-", "between-", and "GLS-" transformed models defines the $\hat{\beta}_W$, $\hat{\beta}_B$ and $\hat{\beta}_{GLS}$ respectively. It can be shown (see Greene, sections 14.3–14.4) that $\hat{\beta}_{GLS}$ is a matrix-weighted average of $\hat{\beta}_W$ and $\hat{\beta}_B$. Specifically,

$$\hat{\beta}_{GLS} = F^W\hat{\beta}_W + (I - F^W)\hat{\beta}_B,$$

where $F^W \equiv [S_{XX}^W + (1-\theta)^2 S_{XX}^B]^{-1} S_{XX}^W$, θ was defined above, and $S_{XX}^{W,B}$ are sample-moment matrices of the X variables from the W, B transformations respectively.
Define three alternative Wu-Hausman statistics based on the three difference vectors:

$$\hat{d}_1 = \hat{\beta}_B - \hat{\beta}_W, \quad \hat{d}_2 = \hat{\beta}_{GLS} - \hat{\beta}_W, \quad \hat{d}_3 = \hat{\beta}_{GLS} - \hat{\beta}_B.$$

   (a) Show that if θ is known exactly (i.e., does not need to be estimated) the three Wu-Hausman tests will be *algebraically* equivalent.

   (b) What types of hypotheses can these statistics be used to test? When would these test procedures have high power?

   (c) Define a fourth Wu-Hausman statistic based on the difference vector: $\hat{d}_4 = \hat{\beta}_{GLS} - \hat{\beta}_{OLS}$ where $\hat{\beta}_{OLS}$ is the OLS estimator from the untransformed data.

i. Explain how you would calculate the variance-covariance matrix of $\hat{d}_4$.
NB: You do *not* need to calculate the precise expression — simply explain what the issues are.

ii. Would such a test have good power properties?

4. Consider the *linear dynamic balanced panel data model*:

$$y_{it} = \delta y_{i,t-1} + x'_{it}\beta + z'_i\gamma + \alpha_i + \nu_{it}, \qquad i = 1, \cdots, N, \quad t = 1, \cdots, T$$

with: k_x time-varying regressors, k_z time-invariant regressors, a_i an unobservable error i.i.d. over i, with unconditional zero mean and variance $\sigma^2_\alpha < \infty$, ν_{it} an error independent of all as and i.i.d. over both i and t with unconditional mean zero and variance $\sigma^2_\nu < \infty$.

(a) Discuss random effects estimation through Maximum Likelihood (ML) and Instrumental Variables (IV) methods along the lines of Bargava and Sargan.

(b) Discuss fixed effects estimation through first differencing and IV methods along the lines of Arellano and Bond. Explain why the regular $y_{it} - \bar{y}_{i\bullet}$ transformation is not useful for this model.

(c) Briefly compare methods (a) and (b) when the following additional complications are present in the linear dynamic panel data model:

i. One of the x_{it} regressors is correlated with ν_{it}.

ii. All of the x_{it} regressors are correlated with α_i.

iii. One of the x_{it} regressors is measured with error, ξ_{it}.

iv. One of the z_i regressors is measured with error, ζ_i.

5. Consider the dynamic linear regression model for balanced data:

$$y_{it} = \delta y_{i,t-1} + x'_{it}\beta + z'_i\gamma + \epsilon_{it} \quad , \quad i = 1, \cdots, N \ , \quad t = 1, \cdots, T$$

where ϵ_{it} follows the one factor error components model: $\epsilon_{it} = \alpha_i + \nu_{it}$ with α_i modelling individual unobserved persistent heterogeneity.

(a) Describe two estimation approaches for this model: the first should rely on the "Fixed Effects" principle of eliminating the unobserved persistent heterogeneity term α_i and carrying out estimation conditional on it.The second should rely on the "Random Effects" principle of deriving the (possibly optimal) estimator that considers either the full p.d.f.. or the first two moments of the disturbance vector $(\epsilon_{11}, \cdots, \epsilon_{1T}, \cdots, \epsilon_{i1}, \cdots, \epsilon_{iT}, \cdots, \epsilon_{N1}, \cdots, \epsilon_{NT})'$, i.e., $pdf(\epsilon|X, Z)$ or $E(\epsilon|X, Z)$ and $VCov(\epsilon|X, Z)$.

You should discuss the properties of the two estimation approaches under the following three scenarios about the ν_{it} error term:

i. $\nu_{it} \sim N(0, \sigma^2_\nu)$ i.i.d. over both i and t;

ii. $\nu_{it} = \xi_{it} + \lambda\xi_{i,t-1}$ with $\xi_{it} \sim N(0, \sigma^2_\xi)$ i.i.d. over both i and t;.

iii. $\nu_{it} = \rho\nu_{i,t-1} + \xi_{it}$ with $|\rho| < 1$ and $\xi_{it} \sim N(0, \sigma^2_\xi)$ i.i.d. over both i and t.

(b) Now assume the simplest $\alpha_i + \nu_{it}$ structure and consider how the two estimation approaches you described above will need to be modified to analyze the alternative models:

$$y_{it} = g(x_{it}, \beta, z_i, \gamma) + \delta y_{i,t-1} + \epsilon_{it} \qquad \text{(Model 1)}$$

where the non-linear function $g(.)$ is known up to parameter vectors β and γ;

and

$$y_{it} = h(x_{it}, \beta, z_i, \gamma, \delta y_{i,t-1}) + \epsilon_{it} \qquad \text{(Model 2)}$$

and where the non-linear function $h(.)$ is known up to parameter vectors β and γ and parameter δ.

(c) Finally suppose that in part (b), the δ parameter equals 0. What happens to Models 1 and 2 in such case? Discuss estimation when (i) all regressors are measured without error; and (ii) when one or more regressor(s) contain(s) errors of measurement. In such case (ii), does it make a difference whether the mismeasured regressors are among the Xs or the Zs?