Lab 3: Static Panel Models

Econometrics Beyond Ordinary Least Squares

Patrick Mokre

WS 2020/2021

Panel Data

Panel Data combines aspects of time series and cross-sectional econometric analysis.

We will have to deal with multi-dimensional group-specific effects.

Popular examples:

- Longitudinal Surveys
- Cross-Country Macro analysis
- Experiments rolled out in multiple waves (Why do researchers do that?)

OLS: Unbiasedness and Efficiency

Unbiasedness: $E[\hat{\beta}] = \beta \Leftarrow E[\epsilon_i \mid X_i] = E[\epsilon_i] = 0.$

Efficiency: $Var(\hat{\beta}) \leq Var(\beta^*) \quad \forall \beta^* \neq \hat{\beta} \Leftarrow V[\hat{\epsilon}_i] = \sigma^2 \quad \forall i \in N; Cov[\epsilon_i, \epsilon_j] = 0 \quad \forall i \neq j \in N.$

Under Gauss-Markov Conditions 1 through 4, $\hat{\beta}_{OLS} = (X'X)^{-1}X'y$ is the **best unbiased linear estimator (BLUE)**.

If $E[\epsilon_i \mid X_i] \neq 0$ or if the **covariance matrix** of ϵ has no uniform diagonal and only zero off-diagonal entries, we **cannot guarantee BLUE**.

$$\Sigma_\epsilon = \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix} \tag{1}$$

Panel Data: Big Questions

- Time Component: Did something happen that year?
- Serial Correlation: Cyclical Over- and undershooting?
- Dynamic Effects: last year's investment and this year's output?
- Group Effects: What if unobserved effects aren't completely independent?
- Advantages of Panels: Number of observations, dynamic effects, group effects
- Disadvantages of Panels: Small-sample properties of estimators, robusticity

Panel Data: Representation

You have observations (y_{it}, x_{it}) for individuals $i \in I$ and periods $t \in T$. Estimate the impact of x on y.

Most general formulation of a model (not much generality in meaning):

$$y_{it} = \alpha_{it} + x_{it}'\beta_{it} + \epsilon_{it} \tag{2}$$

- Strengths: Complex, Precise, allows for heterogeneity
- Weaknesses: No generalization

Panel Data: Pooled Errors

Simpler panel model:

$$y_{it} = \alpha + x_{it}'\beta + \epsilon_{it} \tag{3}$$

We cannot assume ϵ_{it} to be independent and identically distributed (**"i.i.d."**), and specifically not $\epsilon_{it} \sim N(0, \sigma^2)$, if i and t are meaningful.

In face of auto-correlation: Construct **Newey-West pooled errors** for groups $i \in g \in G$.

$$y_{it} = \alpha + x_{it}'\beta + \epsilon_{it} \tag{4}$$

$$\epsilon_{it} \sim N(0, \sigma_g) \tag{5}$$

$$\hat{V}(\hat{\beta}_k) = [X'X]^{-1}[\sum_c^C x_c'\hat{\epsilon}_c'\hat{\epsilon}_c x_c][X'X]^{-1}$$

Panel Data: Fixed and Random Effects

Fixed Effects: Cross-sectional units can have unobserved, **mean-shifting** effects, eg. social connections favoring promotions. This allows for **no common intercept term**.

Think of: Runners starting from somewhere behind the starting line.

$$y_{it} = \alpha_i + \beta' x_{it} + \epsilon_{it} \; ; \; \epsilon_{it} \mid IID(0, \sigma_\epsilon^2) \tag{6}$$

Random Effects: Alternatively, the effects can be treated as individual but drawn from a distribution $F(0, \sigma_\alpha^2)$ which is **independent from** x_{it}.

Think of: Wealth studies, but everyone plays the lottery.

$$y_{it} = \mu + \beta' x_{it} + \alpha_i + u_{it} \tag{7}$$

Fixed Effect Models

Least Squares Dummy Variable (LSDV) approach:

$$y_{it} = \sum_{j=1}^{N} \alpha_j d_{ij} + x_{it}'\beta + \epsilon_{it} \tag{8}$$

Mean Differencing:

$$\begin{aligned} \bar{y}_i &= \frac{1}{T}\sum_{t=1}^{T} y_{it} \\ \bar{y}_i &= \alpha_i + \bar{x}_i' + \bar{\epsilon}_i \\ y_{it} - \bar{y}_i &= (x_{it} - \bar{x})'\beta + (\epsilon_{it} - \bar{\epsilon}_i) \ \alpha_i \text{ drops out} \end{aligned} \tag{9}$$

First Differencing

$$\begin{aligned} y_{it} - y_{it-1} &= (\alpha_i - \alpha_i) + (x_{it} - x_{it-1})'\beta + (\epsilon_{it} - \epsilon_{it-1}) \\ \Delta y_{it} &= \Delta x_{it}'\beta + \Delta\epsilon_{it} \end{aligned} \tag{10}$$

Estimating Fixed Effects

Mean Differenced Models are also called **within estimations**:

- Efficiently estimated by **pooled OLS**.
- **Unbiased only if** covariates are strictly exogenous, ie. $E[(x_{it} - \bar{x}_i \epsilon_{it})] = 0$ and $E[x_{it}\epsilon_{is}] = 0 \quad \forall t \neq s \in T$
- Same estimators but different standard erros as in **LSDV**.
- Intercept can be retrieved.

$$\hat{\beta}_{FE} = \left(\sum_{i=1}^{N} \sum_{j=1}^{T} (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)' \right)^{-1} \sum_{i=1}^{N} \sum_{j=1}^{T} (x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i) \qquad (11)$$

$$\hat{\alpha}_i = \bar{y}_i - \hat{\beta}'_{FE}\bar{x}_i \qquad (12)$$

Note: In short and wide panels (small T, large N) the **intercept** cannot be efficiently retrieved.

Random Effects

RE allow for serial correlation of error terms and decompose errors in a time-dependent **effect error** α_i and a time-invariant **error** ϵ_{it}.

$$y_{it} = \mu + x_{it}'\beta + u_{it} \tag{13}$$

$$u_{it} = \alpha_i + \epsilon_{it} \tag{14}$$

$$E(\epsilon_{it} \mid x_{it}) = 0 \Rightarrow E(u_{it} \mid \alpha_i, x_{it}) = 0$$

Properties of u_{it}:

$$E(u_{it}^2) = \sigma_\alpha^2 + \sigma^2 + 2Cov(\alpha_i, u_{it}) = \sigma_\alpha^2 + \sigma^2 \tag{15}$$

$$E(u_{it}u_{is}) = E[(\alpha_i + u_{it})(\alpha_i + u_{is})] = \sigma_\alpha^2 \tag{16}$$

Estimating Random Effects

- μ and β can be consistently estimated by OLS.
- Standard errors **can not be consistently estimated by OLS**.
- **GLS** provides consistent and efficient estimation, for ι_T a $T \times 1$ vector of 1s and I_T the T-dimensional identity matrix.

$$V[\alpha_i \iota_T + \epsilon_{it}] \equiv \Omega = \sigma_\alpha^2 \iota_t \iota_T' + \sigma_\epsilon^2 I_T \tag{17}$$

$$\Omega^{-1} = \sigma_\epsilon^{-2} \left[I_T - \frac{\sigma_\alpha^2}{\sigma_\epsilon^2 + T\sigma_\alpha^2} \iota_T \iota_T' \right]$$

$$= \sigma_\epsilon^{-2} \left[\left(I_T - \frac{1}{T} \iota_T \iota_T' \right) + \psi \frac{1}{T} \iota_T \iota_T' \right]$$

$$\psi = \frac{\sigma_\alpha^2}{\sigma_\epsilon^2 + T\sigma_\alpha^2} \tag{18}$$

$$\hat{\beta}_{GLS} = \left(\sum_{i=1}^{N} \sum_{j=1}^{T} (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)' + \psi T \sum_{i=1}^{N} (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})' \right)^{-1}$$

$$\times \left(\sum_{i=1}^{N} \sum_{j=1}^{T} (x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i)' + \psi T \sum_{i=1}^{N} (\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y})' \right) \tag{19}$$

Estimating Random Effects: Between and Within Estimator

- The GLS estimator is a **weighted mean** of the between and within estimators with weights λ and $(1-\lambda)$.

$$\hat{\beta}_{GLS} = \lambda \hat{\beta}_B + (1-\lambda)\hat{\beta}_W \tag{20}$$

$$\hat{\beta}_B = \left(\sum_{i=1}^{N} (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})' \right)^{-1} \left(\sum_{i=1}^{N} (\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y})' \right)$$

$$\hat{\beta}_W = \left(\sum_{i=1}^{N} \sum_{j=1}^{T} (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)' \right)^{-1} \sum_{i=1}^{N} \sum_{j=1}^{T} (x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i)$$

Relationship between Panel Estimators

Relationship between FE, RE and Pooled OLS is given by effect heterogeneity σ_α^2.

- For maximum heterogeneity, FE converges to RE
- For minimum heterogeneity, RE converges to Pooled OLS
- FE does not allow for identification of time-invariant covariates (eg. years of schooling among workers).
- FE identifies differences **within** individuals $y_{it} - \bar{y}_i$, not between individuals $\bar{y}_i - \bar{y}_j$.
- RE does not allow for meaningful identification of effects
- RE identifies the between difference $\bar{y}_i - \bar{y}$.
- RE is **more efficient** for $\psi = \frac{\sigma_\alpha^2}{\sigma_\epsilon^2 + T\sigma_\alpha^2} > 1$.

Model Selection (Hausman Test)

The choice between FE and RE makes a substantial difference in estimating β.

The expected values in FE an RE are the same if x_{ij} and α_i are **uncorrelated** $\Rightarrow$ prefer FE to identify α_i, or RE for higher efficiency.

Hausman Test: $H_0 : E[x_{it}\alpha_i] = 0$; $H_A : E[x_{it}, \alpha_i] \neq 0$.

- $\hat{\beta}_{FE}$ is consistent and efficient under both H_0 and H_A.
- $\hat{\beta}_{RE}$ is consistent only under H_0.
- Only under H_0, $V[\hat{\beta}_{FE} + \hat{\beta}_{RE}] = V[\hat{\beta}_{FE}] + V[\hat{\beta}_{RE}]$.

Hausmann test statistic:

$$\xi_H = (\hat{\beta}_{FE} - \hat{\beta}_{RE})'[\hat{V}[\hat{\beta}_{FE}] - \hat{V}[\hat{\beta}_{RE}]]^{-1}(\hat{\beta}_{FE} - \hat{\beta}_{RE}) \tag{21}$$

$$\xi_H \sim \chi^2(K)$$