Applied Data Analysis School

Panel Data Models

Anabela Carneiro[1] Miguel Portela[2]

[1]Universidade do Porto
[2]Universidade do Minho

November 17 & 19, 2020

Panel Data Models

- Outline
  1. Introduction to Panel Data Analysis
  2. Pooled OLS Model
  3. Random Effects Model
  4. Fixed Effects Model
  5. Test for the Presence of Fixed Effects
  6. Test for Random Effects: the Hausman Test
  7. Comparison of Estimators

Introduction

- Econometric models aim to establish causal relationships.
- Endogeneity may confound the evaluation of the effects of interest.
- Laboratory or randomized field experiments are rare in economics and other social sciences research.
- Hence, researchers depend on observational data to make causal claims.
- One of the most important group of models that mitigates endogeneity bias is known as fixed effects models (FE).
- Fixed effects models allow you to deal with endogeneity due to omitted variable.

Introduction

- A FE approach relies on panel data.
- The term panel data (or longitudinal data) is used for a wide variety of situations in econometrics.
- It refers to any data set with repeated observations over time for the same individuals (workers, households, firms, industries, regions, or countries).
- *Panel data* typically refers to situations in which the cross-sectional dimension (N) is large relative to the time dimension (T).

Panel Data

Example

Employer-employee level data: wages, schooling, experience, location, industry, ...

Example

Data on countries: GDP, average education, physical capital, ...

Example

Hospital-doctor-patient data

Example

School-class-teacher-student level data

Panel Data - *Personnel Records* (Quadros de Pessoal): Matched Employer-employee Data Set

Personnel Records - Firms

	FIRMID	LOCATION	INDUSTRY	NUMBER OF EMPLOYEES	SALES	...
2001	10000	1301	52111	10	€	
2002	10000	1301	52111	10	€	
2003	10000	1301	52111	12	€	
2001	20000	1602	17400	25	€	
2002	20000	1602	17400	20	€	
2003	20000	1602	17400	18	€	
2002	160000	1105	65121	2	€	
2003	160000	1105	65121	4	€	
	...	...	...	...	...	...

Panel Data - *Personnel Records* (Quadros de Pessoal): Matched Employer-employee Data Set

Personnel Records - Workers

	WORKERID	FIRMID	GENDER	AGE	EDUC	WAGE	· · ·
2001	999999990	10000	1	21	9	€	
2002	999999990	10000	1	22	9	€	
2003	999999990	10000	1	23	9	€	
2001	999999991	10000	2	40	4	€	
2002	999999991	10000	2	41	6	€	
2001	999999992	20000	1	18	12	€	
2002	999999992	20000	1	19	12	€	
2003	999999992	20000	1	20	12	€	
2002	999999993	160000	1	50	16	€	
2003	999999993	160000	1	51	16	€	
2002	999999994	160000	2	48	9	€	
	· · ·	· · ·	· · ·	· · ·	· · ·	· · ·	· · ·

Advantages of Panel Data (Hsiao, 2003)

- They usually give the researcher a large number of data points, increasing the degrees of freedom and reducing the collinearity among explanatory variables - improved efficiency in estimators and gains in terms of identification.

- Longitudinal data allow a researcher to analyze a number of important economic questions that cannot be addressed using cross-sectional or time-series data sets alone: e. g., firm survival, worker mobility, patient survival, etc.

- Control for unobserved time-invariant variables potentially correlated with the error term.

Specification

$$y_{it} = \alpha_i + \mathbf{x}_{it}\boldsymbol{\beta} + u_{it}, \qquad t = 1, ..., T; i = 1, ..., N \tag{1}$$

- where $\mathbf{x}_{it}$ is a $(1 \times K)$ vector of explanatory variables and $\boldsymbol{\beta}$ $(K \times 1)$ the vector of unknown parameters.
- α_i: unobserved heterogeneity, individual effect, fixed effect.
- u_{it}: is the random error term assumed to be independent and identically distributed IID$(0, \sigma_u^2)$.

Specification

- **Key issue**: is α_i correlated with the observed explanatory variables $\mathbf{x}_{it}, t = 1, ..., T$?
- If α_i is referred as an "individual random effect", α_i is being assumed to be uncorrelated with the $\mathbf{x}_{it}$, i. e., $Cov(\mathbf{x}_{it}, \alpha_i) = \mathbf{0}$, $t = 1, ..., T$.
- If α_i is referred as a "fixed effect", α_i is allowed to be correlated with the $\mathbf{x}_{it}$, i. e., $Cov(\mathbf{x}_{it}, \alpha_i) \neq \mathbf{0}$, $t = 1, ..., T$.

Pooled OLS (POLS)

- The pooled model treats all observations as independent observations.
- The **pooled OLS (POLS) estimator** is consistent if the composite error term is uncorrelated with the regressors, i. e.,

$$cov(\mathbf{x}_{it}, u_{it}) = \mathbf{0}$$
$$cov(\mathbf{x}_{it}, \alpha_i) = \mathbf{0}, \qquad t = 1, ..., T \tag{2}$$

- Even if assumption (2) holds, the composite errors will be serially correlated due to the presence of α_i in each time period ($Cov(v_{it}, v_{is}) = \sigma^2_\alpha$).
- Therefore, inference using pooled OLS requires the robust variance matrix estimator and robust test statistics.

Random Effects Model (RE): Assumptions

- The RE model accomodates this correlation $\Longrightarrow$ apply GLS methods.
- Assumption RE.1:

$$y_{it} = \alpha_i + \mathbf{x}_{it}\boldsymbol{\beta} + u_{it}$$

- Assumption RE.2: We have a random sample in the cross-sectional dimension.
- Assumption RE.3: There are no perfect linear relationships among the explanatory variables.

Random Effects Model (RE): Assumptions

- Assumption RE.4:

$$E(u_{it}|\mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{iT}, \alpha_i) = 0, \qquad t = 1, ..., T$$

$$E(\alpha_i|\mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{iT}) = 0, \qquad t = 1, ..., T$$

- Assumption RE.5:

$$var(u_{it}|\mathbf{x}_i, \alpha_i) = var(u_{it}) = \sigma_u^2, \qquad \text{for all } t = 1, ..., T$$

$$var(\alpha_i|\mathbf{x}_i) = \sigma_\alpha^2, \qquad t = 1, ..., T$$

where $\mathbf{x}_i \equiv (\mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{iT})$.

- Assumption RE.6:

$$cov(u_{it}, u_{is}|\mathbf{x}_i, \alpha_i) = 0, \qquad \text{for all } t \neq s$$

Random Effects Estimator

- Under these assumptions the RE estimator is consistent as N gets large for fixed T.

- The FGLS estimator of the RE model is the **random effects (RE) estimator**:

$$\widehat{\beta}_{RE} = \left(\sum_{i=1}^{N} \mathbf{X}_i' \widehat{\Omega}^{-1} \mathbf{X}_i \right)^{-1} \left(\sum_{i=1}^{N} \mathbf{X}_i' \widehat{\Omega}^{-1} \mathbf{y}_i \right) \tag{3}$$

- The RE estimator is fully efficient under the RE assumptions, though the efficiency gain compared to POLS need not to be great.

Random Effects Estimator

- Where $\mathbf{\Omega}$ takes the special form

$$\mathbf{\Omega} \equiv E(\mathbf{v}_i \mathbf{v}_i^{'}) =$$

$$= \begin{pmatrix} \sigma_\alpha^2 + \sigma_u^2 & \sigma_\alpha^2 & \cdots & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\alpha^2 + \sigma_u^2 & \cdots & \vdots \\ \vdots & & \ddots & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \cdots & & \sigma_\alpha^2 + \sigma_u^2 \end{pmatrix}$$

Random Effects Estimator

- $\widehat{\sigma}_{\alpha}^2$? $\widehat{\sigma}_{u}^2$?
- These estimators can be based on the pooled OLS or fixed effects residuals.
- Let $\widehat{\widehat{v}}_{it}$ denote the pooled OLS residuals. A consistent estimator of σ_v^2 under RE.1-RE.3 is:

$$\widehat{\sigma}_{v}^2 = \frac{1}{(NT-K)} \sum_{i=1}^{N} \sum_{t=1}^{T} \widehat{\widehat{v}}_{it}^2 \tag{4}$$

which is the usual variance estimator from the OLS regression on the pooled data.

Random Effects Estimator

- A consistent estimator of σ_α^2 under RE.1-RE.6 is:

$$\widehat{\sigma}_\alpha^2 = \frac{1}{[NT(T-1)/2 - K]} \sum_{i=1}^{N} \sum_{t=1}^{T-1} \sum_{s=t+1}^{T} \widehat{\widehat{v}}_{it} \widehat{\widehat{v}}_{is} \tag{5}$$

$$\widehat{\sigma}_u^2 = \widehat{\sigma}_v^2 - \widehat{\sigma}_\alpha^2 \tag{6}$$

Testing for the Presence of an Unobserved Effect

- The absence of an unobserved effect is statistically equivalent to H_0: $\sigma_\alpha^2 = 0$.
- Notice that:

$$Corr(v_{it}, v_{is}) = \sigma_\alpha^2 / (\sigma_\alpha^2 + \sigma_u^2) \geq 0, \qquad s \neq t.$$

- Thus, if $\sigma_\alpha^2 = 0$ then $Corr(v_{it}, v_{is}) = 0$.

- Lagrange multiplier test for the random effects model based on the pooled OLS residuals ($\widehat{v}_{it}$) proposed by Breusch and Pagan (1980). For

$$H_0 : \sigma_\alpha^2 = 0$$

$$H_1 : \sigma_\alpha^2 \neq 0$$

Random Effects Estimator

- If $\sigma_\alpha^2 = 0$ then $\mathbf{\Omega}$ takes the special form

$$\mathbf{\Omega} \equiv E(\mathbf{v}_i \mathbf{v}_i^{'}) =$$

$$= \begin{pmatrix} \sigma_u^2 & 0 & \cdots & 0 \\ 0 & \sigma_u^2 & \cdots & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & & \sigma_u^2 \end{pmatrix} = \sigma_u^2 \mathbf{I}_T$$

Testing for the Presence of an Unobserved Effect

- The test statistic is

$$LM = \frac{NT}{2(T-1)} \left[\frac{\sum_{i=1}^{N} \left[\sum_{t=1}^{T} \widehat{v}_{it} \right]^2}{\sum_{i=1}^{N} \sum_{t=1}^{T} \widehat{v}_{it}^2} - 1 \right]^2$$

- Under the null hypothesis, LM is distributed as chi-squared with one degree of freedom. Large values of LM favor the RE model.

Endogeneity Due to Omitted Variables

- Let $v_{it} = \alpha_i + u_{it}$
- If $Cov(\mathbf{x}_{it}, \alpha_i) \neq \mathbf{0}$, then $Cov(\mathbf{x}_{it}, v_{it}) \neq \mathbf{0} \Rightarrow$ endogeneity due to omitted variable.
- OLS and RE estimators are inconsistent.
- Use FE to control for individual unobserved heterogeneity.

Fixed Effects (FE) Model: Assumptions

- Assumption FE.1:

$$y_{it} = \alpha_i + \mathbf{x}_{it}\boldsymbol{\beta} + u_{it}$$

- Assumption FE.2: We have a random sample in the cross-sectional dimension.

- Assumption FE.3: There are no perfect linear relationships among the explanatory variables.

FE Model: Assumptions

- Assumption FE.4 (strict exogeneity assumption):

$$E(u_{it}|\mathbf{x}_i, \alpha_i) = 0, \qquad t = 1, ..., T \tag{7}$$

where $\mathbf{x}_i \equiv (\mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{iT})$.

- Assumption FE.5:

$$var(u_{it}|\mathbf{x}_i, \alpha_i) = var(u_{it}) = \sigma_u^2, \qquad \text{for all } t = 1, ..., T$$

- Assumption FE.6:

$$cov(u_{it}, u_{is}|\mathbf{x}_i, \alpha_i) = 0, \qquad \text{for all } t \neq s$$

FE Estimator

- α_i allowed to be arbitrarily correlated with $\mathbf{x}_i \Longrightarrow$ control for unobserved heterogeneity.

- The idea for estimating $\boldsymbol{\beta}$ under FE.1 is to transform the equations to eliminate the unobserved effect $\alpha_i \Rightarrow$ **FE transformation/within transformation**.

- The FE transformation is obtained by first averaging equation (1) over $t = 1, ..., T$ to get the cross section equation

$$\overline{y}_i = \overline{\mathbf{x}}_i \boldsymbol{\beta} + \alpha_i + \overline{u}_i \tag{8}$$

where $\overline{y}_i = T^{-1} \sum_{t=1}^{T} y_{it}$, $\overline{\mathbf{x}}_i = T^{-1} \sum_{t=1}^{T} \mathbf{x}_{it}$ and $\overline{u}_i = T^{-1} \sum_{t=1}^{T} u_{it}$.

FE Estimator

- Subtracting equation (8) from equation (1) for each t gives FE transformed equation,

$$y_{it} - \overline{y}_i = (\mathbf{x}_{it} - \overline{\mathbf{x}}_i)\boldsymbol{\beta} + u_{it} - \overline{u}_i \tag{9}$$

FE Estimator

- OLS estimator of equation (9) will be consistent, if $cov\left[(\mathbf{x}_{it} - \overline{\mathbf{x}}_i), (u_{it} - \overline{u}_i)\right] = \mathbf{0},\ t = 1, ..., T.$

- The **fixed effects (FE) estimator,** denoted by $\widehat{\beta}_{FE}$ is the OLS estimator from the regression

$$(y_{it} - \overline{y}_i) \text{ on } (\mathbf{x}_{it} - \overline{\mathbf{x}}_i) \tag{10}$$

Identifying Assumption

- Identifying assumption: unobservable factors that might simultaneously affect the dependent variable and the independent variables of the regression are time-invariant.
- Identification of $\boldsymbol{\beta}$ is based on within group variation over time.
- Coefficients of variables that are time-invariant are not identified as $(\mathbf{x}_{it} - \overline{\mathbf{x}}_i) = 0, \forall_i$.

Least Square Dummy Variable (LSDV) Estimator

- If the α_i are parameters to estimate, how would we estimate each α_i along with $\boldsymbol{\beta}$?

- One possibility is to define N dummy variables, one for each cross section observation: $dn_i = 1$ if $n = i$, $dn_i = 0$ if $n \neq i$. Then, run the OLS regression

$$y_{it} \text{ on } d1_i, d2_i, ..., dN_i, \mathbf{x}_{it} \tag{11}$$

- Then, $\widehat{\alpha}_1$ is the coefficient on $d1_i$, $\widehat{\alpha}_2$ is the coefficient on $d2_i$, and so on. The estimator of $\boldsymbol{\beta}$ obtained from regression (11) is, in fact, the FE estimator.

- This is why $\widehat{\boldsymbol{\beta}}_{FE}$ is sometimes referred to as the **least square dummy variable estimator** (**LSDV**).

Testing for Fixed Effects

- Run the OLS regression

$$y_{it} \text{ on constant}, d1_i, d2_i, ..., dN-1_i, \mathbf{x}_{it} \tag{12}$$

- Let SSR_{ur} be the unrestricted sum of squared residuals from regression (12).

- Then, run the OLS regression

$$y_{it} \text{ on constant}, \mathbf{x}_{it} \tag{13}$$

- Let SSR_r be the restricted sum of squared residuals from regression (13).

Testing for Fixed Effects

- $H0 : \alpha_1 = \alpha_2 = \ldots = \alpha_{N-1} = 0$
- Then, under $H0$

$$F = \frac{(SSR_r - SSR_{ur})}{SSR_{ur}} . \frac{[N(T-1) - K]}{N-1} \sim F_{N-1, N(T-1)-K}$$

- If $H0$ is not rejected, using FE does not cause bias.
- Though, in this case, controlling for FE when all $\alpha_i = 0$ will lead to larger standard errors.

Testing for Fixed Effects

- If $H0$ is rejected $\Rightarrow$ use FE to control for them.
- Notice, however, that the pooled OLS model and the FE model will not necessarily produce different results.
- Bias occurs when $Cov(\mathbf{x}_{it}, \alpha_i) \neq \mathbf{0}$.
- To cause bias, FE must exist and be correlated with the independent variables.

Estimating the Fixed Effects

- Sometimes it is useful to obtain the $\widehat{\alpha}_i$ even when regression (11) is unfeasible (N is too large). Using the OLS first-order conditions, each $\widehat{\alpha}_i$ can be shown to be

$$\widehat{\alpha}_i = \overline{y}_i - \overline{\mathbf{x}}_i \widehat{\boldsymbol{\beta}}_{FE}, \qquad i = 1, ..., N$$

- Econometric software that employs fixed effects usually supresses the "estimates" of the α_i, although an overall intercept is often reported.

RE vs FE - Hausman (1978) Test

- Comparing the FE and RE estimates can be a test for whether there is correlation between α_i and $\mathbf{x}_{it}$, assuming that the idiosyncratic errors and explanatory variables are uncorrelated across all time periods.

- Given the random effects specification:

$$y_{it} = \alpha_i + \mathbf{x}_{it}\boldsymbol{\beta} + u_{it}$$

- and assuming that the orthogonality condition holds for the u_{it}, the null hypothesis that we want to test is:

$$H_0 : Cov(\mathbf{x}_{it}, \alpha_i) = \mathbf{0}$$

while the alternative hypothesis is:

$$H_1 : Cov(\mathbf{x}_{it}, \alpha_i) \neq \mathbf{0}$$

RE vs FE - Hausman (1978) Test

- Under H_0:

  - the fixed effects estimator is consistent but inefficient;
  - the random effects estimator is consistent and efficient;

- Under H_1:

  - the fixed effects estimator remains consistent;
  - the random effects estimator becomes inconsistent.

- Therefore, under the null hypothesis the two estimators should not differ and this observation provides the basis for the test.

RE vs FE - Hausman (1978) Test

- Let $\widehat{\delta}_{RE}$ denote the vector of RE estimates for the coefficients on the regressors that change across both i and t, and let $\widehat{\delta}_{FE}$ denote the corresponding FE estimates; let these each be Kx1 vectors.

- The Hausman test statistic can be written as:

$$H = (\widehat{\delta}_{FE} - \widehat{\delta}_{RE})' [\widehat{\mathbf{V}}(\widehat{\delta}_{FE}) - \widehat{\mathbf{V}}(\widehat{\delta}_{RE})]^{-1} (\widehat{\delta}_{FE} - \widehat{\delta}_{RE}) \overset{H0}{\sim} \chi^2_K$$

  where the $\widehat{\mathbf{V}}$ denote the estimates of the true covariance matrices.

- If $H0$ is rejected, we have to conclude that the RE estimator is inconsistent because the orthogonality condition fails.

Decomposing OLS Estimates

- Ignoring the individual effects, we have the standard model

$$y_{it} = \alpha + \mathbf{x}_{it}\boldsymbol{\beta} + u_{it} \tag{14}$$

- We can decompose the OLS estimator by considering two variations on this model.
- Deviations from means (within estimator):

$$y_{it} - \overline{y}_i = (\mathbf{x}_{it} - \overline{\mathbf{x}}_i)\boldsymbol{\beta} + (u_{it} - \overline{u}_i) \tag{15}$$

- Group means only (between estimator):

$$\overline{y}_i = \alpha_i + \overline{\mathbf{x}}_i\boldsymbol{\beta} + \overline{u}_i$$

$$\overline{y}_i = \alpha + \overline{\mathbf{x}}_i\boldsymbol{\beta} + (\alpha_i - \alpha + \overline{u}_i) \tag{16}$$

Decomposing OLS Estimates

- The **between estimator (BE)** is the OLS estimator from regression of $\overline{y}_i$ on an intercept and $\overline{\mathbf{x}}_i$.
- The between estimator is consistent if the regressors $\overline{\mathbf{x}}_i$ are independent of the composite error $(\alpha_i - \alpha + \overline{u}_i)$ in (16).
- This will be the case for the random effects model, but the between estimator effectively discards the time series information in the data set (uses only variation between the cross section observations).
- It is more efficient to use the RE estimator.
- In contrast, for the fixed effects model the between estimator is inconsistent as α_i is then assumed to be correlated with $\mathbf{x}_{it}$ and hence $\overline{\mathbf{x}}_i$.

Decomposing OLS Estimates

- In fact, model (14) implies

$$y_{it} - \overline{\overline{y}} = (\mathbf{x}_{it} - \overline{\mathbf{x}}_i)\boldsymbol{\beta} + (\overline{\mathbf{x}}_i - \overline{\overline{\mathbf{x}}})\boldsymbol{\beta} + (u_{it} - \overline{\overline{u}})$$

where $\overline{\overline{\mathbf{x}}} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{x}_{it}$ and $\overline{\overline{y}} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}y_{it}$.

Decomposing OLS Estimates - Sum of Squares and Cross Products

Let

$$S_{xx}^{t} = \sum_{i=1}^{N}\sum_{t=1}^{T}(\mathbf{x}_{it} - \overline{\overline{\mathbf{x}}})'(\mathbf{x}_{it} - \overline{\overline{\mathbf{x}}})$$

$$S_{xy}^{t} = \sum_{i=1}^{N}\sum_{t=1}^{T}(\mathbf{x}_{it} - \overline{\overline{\mathbf{x}}})(y_{it} - \overline{\overline{y}})$$

total group variation

$$S_{xx}^{w} = \sum_{i=1}^{N}\sum_{t=1}^{T}(\mathbf{x}_{it} - \overline{\mathbf{x}}_{i})'(\mathbf{x}_{it} - \overline{\mathbf{x}}_{i})$$

$$S_{xy}^{w} = \sum_{i=1}^{N}\sum_{t=1}^{T}(\mathbf{x}_{it} - \overline{\mathbf{x}}_{i})(y_{it} - \overline{y}_{i})$$

within group variation

Decomposing OLS Estimates - Sum of Squares and Cross Products

$$S_{xx}^{b} = \sum_{i=1}^{N} T(\overline{\mathbf{x}}_i - \overline{\overline{\mathbf{x}}})'(\overline{\mathbf{x}}_i - \overline{\overline{\mathbf{x}}})$$

$$S_{xy}^{b} = \sum_{i=1}^{N} T(\overline{\mathbf{x}}_i - \overline{\overline{\mathbf{x}}})(\overline{y}_i - \overline{\overline{y}})$$

between group variation

It is easy to verify that

$$S_{xx}^{t} = S_{xx}^{w} + S_{xx}^{b}$$

$$S_{xy}^{t} = S_{xy}^{w} + S_{xy}^{b}$$

Decomposing OLS Estimates - Sum of Squares and Cross Products

- There are, therefore, three possible least squares estimators of β corresponding to the decomposition.

- The least squares estimator is

$$\widehat{\beta}^t = [S_{xx}^t]^{-1} S_{xy}^t = [S_{xx}^w + S_{xx}^b]^{-1} [S_{xy}^w + S_{xy}^b]$$

- The **within-groups** estimator is

$$\widehat{\beta}^w = [S_{xx}^w]^{-1} S_{xy}^w$$

- An alternative estimator would be the **between-groups** estimator,

$$\widehat{\beta}^b = [S_{xx}^b]^{-1} S_{xy}^b$$

Decomposing OLS Estimates - Sum of Squares and Cross Products

- The standard OLS estimator of model (14) can be decomposed into

$$\widehat{\boldsymbol{\beta}}^{t} = [S_{xx}^{t}]^{-1} S_{xy}^{t} = F^{w}\widehat{\boldsymbol{\beta}}^{w} + F^{b}\widehat{\boldsymbol{\beta}}^{b}$$

  where

$$F^{w} = ([S_{xx}^{w} + S_{xx}^{b}])^{-1} S_{xx}^{w}$$

$$F^{b} = (1 - F^{w})$$

- which shows that the OLS estimator can be interpreted as weighted average of the within and between estimators with weights that depend on the "within" versus "between" variability of the explanatory factors.

Decomposing OLS Estimates - Sum of Squares and Cross Products

However:

- in general this is not the most efficient way to exploit jointly the within and between variability;
- it leads to biased and inconsistent estimates of the true causal effect β if the individual specific effects are correlated with the regressors.

Decomposing RE Estimates

- Like the OLS estimator, also the random effects estimator can be interpreted as a weighted average of the within and between estimator:

$$\widehat{\boldsymbol{\beta}}^{re} = F^{w}\widehat{\boldsymbol{\beta}}^{w} + (F^{b})\widehat{\boldsymbol{\beta}}^{b}$$

where

$$F^{w} = ([S_{xx}^{w} + \theta S_{xx}^{b}])^{-1} S_{xx}^{w}$$

$$\theta = \sigma_{u}^{2} / (\sigma_{u}^{2} + T\sigma_{\alpha}^{2})$$

Decomposing RE Estimates

- The RE estimator is obtained by estimating the transformed equation

$$y_{it} - \lambda \overline{y}_i = (\mathbf{x}_{it} - \lambda \overline{\mathbf{x}}_i)\boldsymbol{\beta} + (v_{it} - \lambda \overline{v}_i) \tag{17}$$

where

$$\lambda = 1 - [\sigma_u^2 / (\sigma_u^2 + T\sigma_\alpha^2)]^{1/2}$$

- The GLS estimator is simply the pooled OLS estimator of equation (17).

Decomposing RE Estimates

Note that:

- If $\sigma_{\alpha}^2 = 0 \rightarrow \lambda = 0$: the RE estimator is identical to the OLS estimator because there is no individual heterogeneity.

- If $\sigma_{u}^2 = 0 \rightarrow \lambda = 1$: in which case the only existing ignorance would be the individual-specific one captured by α_i and the RE estimator would be identical to the FE estimator.

- If $T \rightarrow \infty \rightarrow \lambda = 1$: if T goes to infinity the unobserved α_i "becomes observable" and the RE estimator would be identical to the FE estimator.

First-Difference (FD) Estimator

- Assumption FD.1:

$$y_{it} = \alpha_i + \mathbf{x}_{it}\boldsymbol{\beta} + u_{it} \tag{18}$$

- Assumption FD.2: We have a random sample in the cross-sectional dimension.

- Assumption FD.3: There are no perfect linear relationships among the explanatory variables.

First-Difference (FD) Estimator

- Assumption FD.4:

$$E(u_{it}|\mathbf{x}_i, \alpha_i) = 0, t = 1, 2, ..., T$$

where $\mathbf{x}_i \equiv (\mathbf{x}_{i1}, \mathbf{x}_{i2}, ..., \mathbf{x}_{iT})$.

- Lagging the model (18) one period and subtracting gives

$$\Delta y_{it} = \Delta \mathbf{x}_{it}\boldsymbol{\beta} + \Delta u_{it}, \qquad t = 2, 3, ..., T$$

where $\Delta y_{it} = y_{it} - y_{i,t-1}$, $\Delta \mathbf{x}_{it} = \mathbf{x}_{it} - \mathbf{x}_{i,t-1}$, and $\Delta u_{it} = u_{it} - u_{i,t-1}$.

First-difference Estimator

- The **first-difference (FD) estimator**, $\widehat{\beta}_{FD}$, is the OLS estimator from the regression

$$\Delta y_{it} \text{ on } \Delta \mathbf{x}_{it}, \quad t = 2, 3, ..., T \tag{19}$$

- OLS estimation of the first-differenced equations will be consistent because the strict exogeneity holds in the first-differenced equation:

$$E(\Delta u_{it} | \Delta \mathbf{x}_{i2}, \Delta \mathbf{x}_{i3}, ..., \Delta \mathbf{x}_{iT}) = 0, \qquad t = 2, 3, ..., T$$

which means the FD estimator is actually unbiased conditional on $\mathbf{X}$.

FD or FE?

- Under FE.1-FE.6, the FE estimator is asymptotically efficient in the class of estimators using the strict exogeneity assumption FE.4.
- Therefore, the first difference estimator is less efficient than fixed effects under FE.1-FE.6.
- FE.5-FE.6 assumes homoskedasticity and no serial correlation in u_{it}.
- Assuming that the $\{u_{it} : t = 1, 2, ..., T\}$ are serially uncorrelated may be too strong.

FD or FE?

- An alternative assumption is that the first difference of the idiosyncratic errors, $\{e_{it} \equiv \Delta u_{it} : t = 2, ..., T\}$, are serially uncorrelated (and have constant variance):

- Assumption FD.5:

$$var(e_{it}|\mathbf{x}_i, \alpha_i) = \sigma_e^2, \qquad \text{for all } t = 2, ..., T$$

- Assumption FD.6:

$$cov(e_{it}, e_{is}|\mathbf{x}_i, \alpha_i) = 0, \qquad \text{for all } t \neq s$$

- Under FD.6 we can write $u_{it} = u_{i,t-1} + e_{it}$, so that no serial correlation in the e_{it} implies that u_{it} is a random walk.

FE vs FD

- When $T = 2$, fixed effects estimation and first differencing produce identical estimates and inference.
- When $T > 2$, the choice between FD and FE hinges on the assumptions about the idiosyncratic errors, u_{it}.
- The FE estimator is more efficient under FE.6 - the u_{it} are serially uncorrelated - while the FD estimator is more efficient when u_{it} follows a random walk.

FE vs FD - Testing for Serial Correlation

- Under FD.6, the errors $e_{it} \equiv \Delta u_{it}$ should be serially uncorrelated.
- As suggested by Wooldridge (2010), we can easily test this assumption given the pooled OLS residuals from regression (16).
- Since the strict exogeneity assumption holds, we can use the t statistic on $\widehat{\rho}$ in the regression

$$\widehat{e}_{it} = \rho \widehat{e}_{i,t-1} + error_{it}, \quad t = 3, 4, ..., T \tag{20}$$

- If the idiosyncratic errors $\{u_{it} : t = 1, 2, ..., T\}$ are uncorrelated to begin with, $\{e_{it} : t = 2, 3, ..., T\}$ will be autocorrelated.
- In fact, under FE.6 it is easily shown that $Corr(e_{it}, e_{i,t-1}) = -0.5$.
- In any case, a finding of significant serial correlation in the e_{it} warrants computing the robust variance matrix for the FD estimator.

Comparison of Estimators: Wrap-up

PANEL DATA MODELS

Assumed Model

	POLS	RE	FE
POLS	Consistent	Consistent	Inconsistent
RE	Consistent	Consistent	Inconsistent
BE	Consistent	Consistent	Inconsistent
FE	Consistent	Consistent	Consistent
FD	Consistent	Consistent	Consistent

FE Model Pros and Cons: Wrap-up

PROS	CONS
Accounts for permanent observed and unobserved heterogeneity	Needs panel data
Mitigates endogeneity issues due to omitted variables	Identification of the parameters of the regressors is based on intragroup variation over time
Always consistent whether the fixed effect is correlated or uncorrelated with the regressors	Less efficient than POLS and RE if the fixed effect is not correlated with the regressors
	Coefficients of time-invariant covariates are not identified

"Besides the advantage that panel data allow us to construct and test more complicated behavioral models than purely cross-sectional or time-series data, the use of panel data also provides a means of resolving or reducing the magnitude of a key econometric problem that often arises in empirical studies, namely, the often-heard assertion that the real reason one finds (or does not find) certain effects is because of omitted (mismeasured, not observed) variables that are correlated with explanatory variables."

Hsiao (2003)

Estimation of a High-Dimensional FE Model

- Increasingly large data sets make it possible to control for different sources of unobserved heterogeneity
- Examples are:
  - Employer-employee level data

- With high-dimensional models explicit introduction of dummy variables to account for fixed effects is not an option.

Estimation of a High-Dimensional FE Model

- Increasingly large data sets make it possible to control for different sources of unobserved heterogeneity
- Examples are:
  - Employer-employee level data
  - Hospital-doctor-patient data
- With high-dimensional models explicit introduction of dummy variables to account for fixed effects is not an option.

Estimation of a High-Dimensional FE Model

- Increasingly large data sets make it possible to control for different sources of unobserved heterogeneity
- Examples are:
  - Employer-employee level data
  - Hospital-doctor-patient data
  - School-class-teacher-student level data
- With high-dimensional models explicit introduction of dummy variables to account for fixed effects is not an option.

Estimation of a High-Dimensional FE Model

- There are a few commands that allow to estimate a high-dimensional fixed effects model
- The gold standard!
  - the Stata command **reghdfe** - absorbs any number of fixed effects and their interactions, implements IV estimation, much faster and takes advantage of multiple cores, excellent support (github) - by Sergio Correia
  - the R package **lfe** uses the same algorithm as in tha Stata module **reg2hdfe** developed by Paulo Guimarães

References

- Bailey, M. A. (2016), Real Econometrics: The Right Tools to Answer Important Questions, Oxford University Press.

- Cameron, A. C. and P. K. Trivedi (2005), Microeconometrics – Methods and Applications, Cambridge University Press.

- Correia, S. (2016), "REGHDFE: Stata Module to Perform Linear or Instrumental-variable Regression Absorving any Number of High-dimensional Fixed Effects". Statistical Software Components s457874, Boston College Department of Economics, revised 25 Jul 2015.

- Gaure, S. (2013), "LFE: Linear Group Fixed Effects", The R Journal, 5(2) 104-116.

References

- Guimarães, P. and Portugal, P. (2010). "A Simple Feasible Procedure to Fit Models with High-dimensional Fixed Effects", the Stata Journal, 10(4) 628-649.
- Hausman, J. A. (1978). "Specification Tests in Econometrics", Econometrica 46(6): 1251-1271.
- Hsiao, C. (2003). Analysis of Panel Data, Cambridge University Press, Cambridge.
- Verbeek, M. (2017), A Guide to Modern Econometrics, John Wiley & Sons.
- Wooldridge, J. M. (2010), Econometric Analysis of Cross Section and Panel Data, MIT Press, Cambridge, MA.

Appendix

$$\mathbf{y}_i = \begin{bmatrix} \mathbf{y}_{i1} \\ \mathbf{y}_{i2} \\ \vdots \\ \mathbf{y}_{iT} \end{bmatrix}$$

$$\mathbf{X}_i = \begin{bmatrix} \mathbf{X}_{i1}^1 & \mathbf{X}_{i1}^2 & \cdots & \mathbf{X}_{i1}^k \\ \mathbf{X}_{i2}^1 & \mathbf{X}_{i2}^2 & \cdots & \mathbf{X}_{i2}^k \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{X}_{iT}^1 & \mathbf{X}_{iT}^2 & \cdots & \mathbf{X}_{iT}^k \end{bmatrix} \qquad \mathbf{v}_i = \begin{bmatrix} \mathbf{v}_{i1} \\ \mathbf{v}_{i2} \\ \vdots \\ \mathbf{v}_{iT} \end{bmatrix}$$

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_N \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_N \end{bmatrix} \qquad \mathbf{v} = \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \vdots \\ \mathbf{v}_N \end{bmatrix}$$

Appendix - Wald Test

- Wald test

- Any hypothesis involving the coefficients of a regression equation can be expressed as one or more restrictions on the coefficient vector:

$$R\beta = r$$

where R is a $(q \times k)$ matrix and r is a q-element column vector; each row of R imposes one restriction on the coefficient vector.

- The Wald test uses the point and variance estimates from the unrestricted model to evaluate whether there is evidence that the restrictions are false.

Appendix - Wald Test

- Given an hypothesis expressed as $H_0 : R\beta = r$, we can construct the Wald statistic as:

$$W = \left(\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{r}\right)' \{\mathbf{R}\widehat{\mathbf{Var}(\widehat{\boldsymbol{\beta}})}\mathbf{R}'\}^{-1} \left(\mathbf{R}\widehat{\boldsymbol{\beta}} - \mathbf{r}\right) \sim \chi^2_{(h)}$$

where $\widehat{Var(\widehat{\boldsymbol{\beta}})}$ is a consistent estimator of $Var(\widehat{\boldsymbol{\beta}})$ and h the number of restrictions under the null.