Applied Econometrics

Static panel data

Miguel Portela[1,2]

[1]NIPE – UMinho
[2]IZA, Bonn

Universidade do Minho

November, 2021

Panel data: outline

1. Panel data analysis: introduction
2. Regression model
3. Fixed-effects model
4. Test for the presence of fixed-effects
5. Between-groups estimator
6. Random-effects model
7. Hypothesis testing: Wald and Hausman tests
8. High-dimensional fixed effects

Introduction

- Panel data combines time series and cross section data
- We have information on the same unit of analysis over time
- It allows us to follow a given unit under observation over time $\rightarrow$ the different observations of the same unit are not independent
- We need appropriate models to deal with it: we will only discuss single equation models
- We must distinguish panel data from pooled data
- Pooled data: population samples for different periods; a common assumption is independence between sub-samples, so, if true, we can assume no serial correlation between the residuals of different observations (between sub-samples)

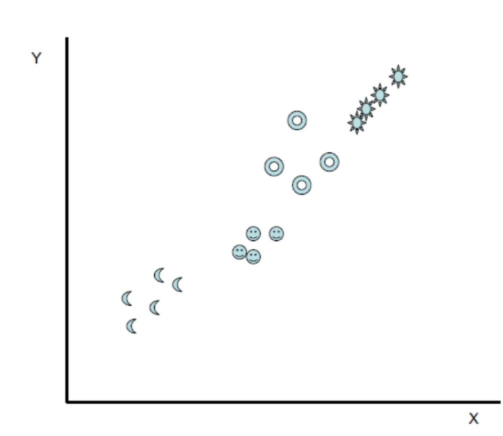
Introduction (cont.)

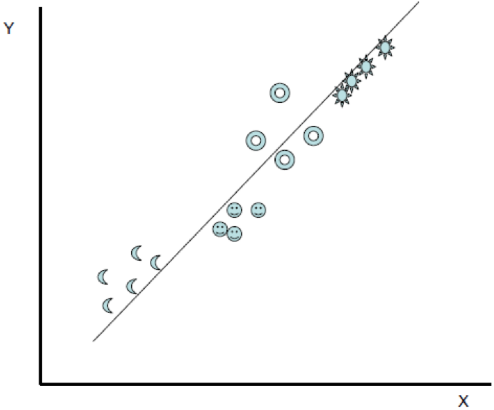
- Increasingly large data sets make it possible to control for different sources of unobserved heterogeneity across the observed units
- These data are sometimes described as multilevel or hierarchical
- Repeated observations allow for the introduction of additional error components that account for unobserved heterogeneity
- If error components are uncorrelated with observed explanatory variables we can use mixed models (error components are treated as random effects)
- If error components are correlated with observed explanatory variables then they need to be treated as fixed effects
- Introducing fixed effects in a linear regression amounts to modelling the error component using dummy variables
- With only one error component this is the usual fixed effects panel data regression

Introduction (cont.)

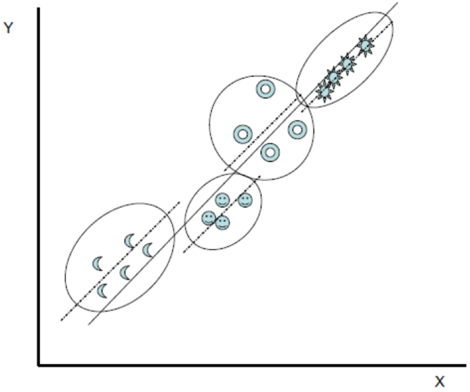
- Employer-employee level data: wages, schooling, experience, tenure, location, industry, ...
- Data on countries: GDP, average education, physical capital, language, inland, ...
- Hospital-doctor-patient data
- School-class-teacher-student level data
- In a typical panel, for example when we analyze the labour market, we have a great number of cross section units/individuals and a short number of periods
- Advantages when using panel data:
    - Flexibility in modelling unit behavior
    - Improved efficiency in estimators and gains in terms of identification
    - Control for unobserved time-invariant variables potentially correlated with the error term

Graphical illustration

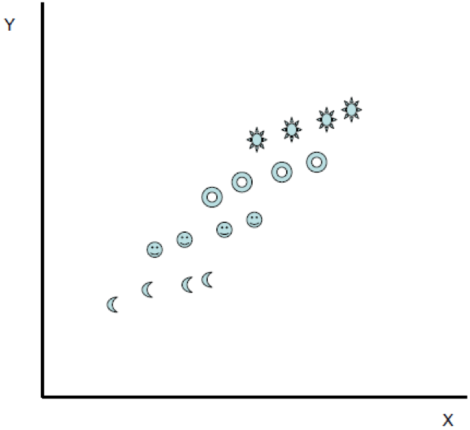



Y
X

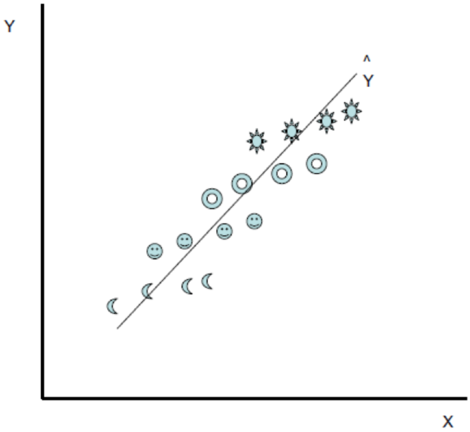
Graphical illustration (cont.)



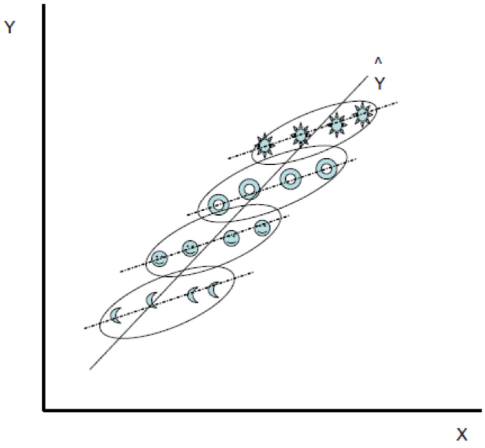
Graphical illustration (cont.)



Graphical illustration (cont.)



Graphical illustration (cont.)



Regression model

$$y_{it} = x_{it}^{'}\beta + z_{i}^{'}\alpha + \varepsilon_{it} \qquad (1)$$

where i stands for the unit and t stands for the time

- $x_{it}^{'}$: dimension $(t \times k)$; vector x_{it} contains k regressors, excluding the constant
- β: column vector $(k \times 1)$
- $z_{i}^{'}\alpha$: it includes the constant and other variables that assume a constant value within individuals/units; it includes unobserved heterogeneity, which is a specific component of unit i
- Example: $(i = 1, 2), (t = 1, 2)$

$$y_{it} = \begin{bmatrix} 730 \\ 790 \\ 830 \\ 870 \end{bmatrix} \qquad x_{it}^{'} = \begin{bmatrix} 23 & 4 \\ 24 & 5 \\ 21 & 1 \\ 22 & 2 \end{bmatrix} \qquad z_{i}^{'} = \begin{bmatrix} 1 & 12 \\ 1 & 12 \\ 1 & 10 \\ 1 & 10 \end{bmatrix}$$

The model (cont.)

- $z_i^{'}$: can contain both unobserved and observed variables
- Examples of unobserved variables: worker's skill, consumer's preferences, ...
- If all the variables are observed we can use OLS to estimate the model Estimation solutions:
- (1) OLS: estimation applied to group data, where z_i only contains a constant
- (2) fixed effects model: when z_i contains unobserved elements which are correlated with x_{it}
  - The OLS estimator is biased and inconsistent when estimating β as a result of omitted variables

Alternative formulation of the model

$$y_{it} = x_{it}^{'}\beta + \alpha_i + \varepsilon_{it} \tag{2}$$

- $\alpha_i = z_i^{'}\alpha$ is a specific component of unit i; it is constant for this unit, which means that it does not vary within unit/individual
- This element captures the unobserved heterogeneity associated with each unit under analysis
- α_i: this is a conditional mean, which incorporates all elements, observed and unobserved, that do not vary within individuals
- α_i can be estimated

Random effects model

- (3) random effects model: its appropriate when the unobserved heterogeneity, α_i, is not correlated with the remaining variables included in the model

$$
\begin{aligned}
y_{it} &= x_{it}^{'}\beta + E\left[z_i^{'}\alpha\right] + \left\{z_i^{'}\alpha - E\left[z_i^{'}\alpha\right]\right\} + \varepsilon_{it} = \\
&= x_{it}^{'}\beta + \alpha + \underbrace{u_i + \varepsilon_{it}}_{\text{composite error term}}
\end{aligned} \tag{3}
$$

where u_i is a random element specific to unit i

- The model specified in equation (3) can be estimated by OLS, but this solution is not efficient

Fixed effects model

$$Y_i = X_i\beta + i\alpha_i + \varepsilon_i \tag{4}$$

where α_i is unknown, and can be estimated

- The differences between the units are captured as differences in the constant term of the model
- T: is the number of observation for unit $i \rightarrow$ for now we assume that all units i have the same number of observations (balanced panel)
- i: $(T \times 1) \rightarrow$ it represents a column of *ones* with dimension $(T \times 1)$
- Grouping the data for all units we obtain

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ \vdots \\ X_n \end{bmatrix} \beta + \begin{bmatrix} i & 0 & 0 & \cdots & 0 \\ 0 & i & 0 & \cdots & 0 \\ 0 & 0 & i & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & i \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Fixed effects model (cont.)

$$Y = \left[\begin{array}{ccccc} X & d_1 & d_2 & \cdots & d_n \end{array} \right] \left[\begin{array}{c} \beta \\ \alpha \end{array} \right] + \varepsilon$$

- d_i: is a vector with dummy variables which indicate unit i

$$\underset{(nT \times n)}{D} = \left[\begin{array}{cccc} d_1 & d_2 & \cdots & d_n \end{array} \right]$$

- The matrix D has n columns (total number of units); nT corresponds to the total number of observations

$$Y = X\beta + D\alpha + \varepsilon \tag{5}$$

- The equation (5) can be estimated by OLS, where $D\alpha$ represents a set of dummy variables to be included in the model
- This procedure is called *Least Squares Dummy Variable* (LSDV)
- In this case, we estimate $k + n$ parameters

Fixed effects model (cont.)

The model to be estimated can be defined as

$$\hat{\beta} = \left[X' M_D X\right]^{-1} \left[X' M_D Y\right] \tag{6}$$

with

$$M_D = I - D\left(D' D\right)^{-1} D'$$

- We can see it as an OLS regression of $Y_* = M_D Y$ on $X_* = M_D X$
- This procedure is identical to running a regression of $\left[y_{it}' - \bar{y}_{i.}\right]$ on $\left[x_{it}' - \bar{x}_{i.}\right]$; i.e., we transform each variable as the deviation to its mean within each unit

$$\hat{\alpha}_i = \bar{y}_{i.} - \hat{\beta}\bar{x}_{i.} \tag{7}$$

$$\bar{y}_{i.} = \frac{\sum_t y_{it}}{T} \qquad \bar{x}_{i.} = \frac{\sum_t x_{it}}{T}$$

Fixed effects model (cont.)

The model based on the transformed variables can be defined as

$$y_{it} - \bar{y}_{i.} = \left(x_{it} - \bar{x}_{i.}\right)' \beta + \left(\varepsilon_{it} - \bar{\varepsilon}_{i}\right) \tag{8}$$

The consistency of the estimator depends on

$$E\left\{\left(x_{it} - \bar{x}_{i.}\right) \varepsilon_{it}\right\} = 0 \tag{9}$$

Variance-covariance matrix

$$\widehat{V\left(\hat{\beta}\right)} = s^2 \left[X' M_D X\right]^{-1} \tag{10}$$

where

$$s^2 = \frac{\left(M_D Y - M_D X\hat{\beta}\right)' \left(M_D Y - M_D X\hat{\beta}\right)}{\left(nT - n - k\right)} \tag{11}$$

Test for the presence of fixed effects

$$F_{(n-1,nT-n-k)} = \frac{\frac{R^2_{LSDV}-R^2_{OLS}}{n-1}}{\frac{1-R^2_{LSDV}}{nT-n-k}} \tag{12}$$

$H_0 : \alpha_2 = \alpha_3 = ... = 0$
H_1 : H_0 is not true

- We are testing if there are no differences between the units
- When implementing the estimation the model has one general constant and $(n-1)$ *dummies*, so we have an omitted group and we estimate the difference to the general constant, $(\alpha_i - \alpha)$

Between-groups estimator (pooled regression)

$$y_{it} = x_{it}^{'}\beta + \alpha + \varepsilon_{it} \tag{13}$$

- We now have OLS applied to grouped data
- The variables are transformed as deviations to individual means

$$y_{it} - \bar{y}_{i.} = \left(x_{it} - \bar{x}_{i.}\right)' \beta + \left(\varepsilon_{it} - \bar{\varepsilon}_{i.}\right) \tag{14}$$

- The model can be defined as unit means

$$\bar{y}_{i.} = \bar{x}_{i.}^{'}\beta + \alpha + \bar{\varepsilon}_{i.} \tag{15}$$

- In this case we use only n observations
- The three models can be estimated consistently by OLS, depending on the existence, or not, of fixed effects
- If there are fixed effects, OLS for grouped data will give biased and inconsistent estimators

Between-groups estimator (cont.)

Sum of squares and cross products

$$S_{XX}^{inter} = \sum_{i=1}^{n} T\left(\bar{x}_{i.} - \bar{\bar{x}}\right)\left(\bar{x}_{i.} - \bar{\bar{x}}\right)'$$
$$S_{XY}^{inter} = \sum_{i=1}^{n} T\left(\bar{x}_{i.} - \bar{\bar{x}}\right)\left(\bar{y}_{i.} - \bar{\bar{y}}\right)'$$

$$\begin{aligned}
\hat{\beta}^{Total} &= \left[S_{XX}^{intra} + S_{XX}^{inter}\right]^{-1}\left[S_{XY}^{intra} + S_{XY}^{inter}\right] \\
\hat{\beta}^{inter} &= \left[S_{XX}^{inter}\right]^{-1} S_{XY}^{inter} \Leftrightarrow S_{XY}^{inter} = S_{XX}^{inter}\hat{\beta}^{inter} \\
\hat{\beta}^{intra} &= \left[S_{XX}^{intra}\right]^{-1} S_{XY}^{intra} \Leftrightarrow S_{XY}^{intra} = S_{XX}^{intra}\hat{\beta}^{intra}
\end{aligned}$$

$$\begin{aligned}
\hat{\beta}^{Total} &= \left[S_{XX}^{intra} + S_{XX}^{inter}\right]^{-1}\left[S_{XX}^{intra}\hat{\beta}^{intra} + S_{XX}^{inter}\hat{\beta}^{inter}\right] \\
&= \left[S_{XX}^{intra} + S_{XX}^{inter}\right]^{-1} S_{XX}^{intra}\hat{\beta}^{intra} + \left[S_{XX}^{intra} + S_{XX}^{inter}\right]^{-1} S_{XX}^{inter}\hat{\beta}^{inter}
\end{aligned}$$

- OLS is a weighted average of the fixed effects and between-groups estimators

Random effects model

- The random error term represents all factors that influence the dependent variable, but are not included in the model as regressors
- α_i: random factors independently and identically distributed across individuals

$$y_{it} = \mu + x_{it}^{'}\beta + \underbrace{\alpha_i + \varepsilon_{it}}_{\text{error term with two components}}$$

$$\varepsilon_{it} \sim IID\left(0, \sigma_{\varepsilon}^{2}\right); \quad \alpha_i \sim IID\left(0, \sigma_{\alpha}^{2}\right)$$

- Specific component to the unit/individual, which does not vary over time
- Residual component of the error term, which is not serially correlated over time, nor correlated with the regressors
- The serial correlation is associated with α_i
- α_i and ε_{it} are independently distributed and independent of the explanatory variables included in the model
- OLS produces consistent estimates of μ and β

Random effects model (cont.)

- However, the standard-errors associated with OLS are not correct
- We can use a more efficient estimator exploring the structure of the Variance-covariance matrix of the error term
- The fixed effects and the random effects estimators are identical when T is large

$$\hat{\beta}_{GLS} = \Delta\hat{\beta}_B + (I_k - \Delta)\,\hat{\beta}_{FE} \tag{16}$$

  where Δ is a weighting matrix proportional to the inverse of the Variance-covariance matrix of $\hat{\beta}_B$ (between-groups estimator)
- The $\hat{\beta}_{GLS}$ is a weighted average of the between-groups and fixed effects estimators, where the weighting depends on the variance between the two estimators
- This estimator is an optimal combination of both estimators, being as a result more efficient

Random effects model (cont.)

- The random effects estimator is unbiased when the explanatory variables are independent of ε_{it} and α_i

$$E\left[\bar{x}_i \varepsilon_{it}\right] = 0 \quad ; \qquad\qquad E\left[\bar{x}_i \alpha_i\right] = 0 \tag{17}$$

- Consistency is assured when N or T, or both, go to infinity

$$y_{it} - \upsilon \bar{y}_i = \mu\left(1 - \upsilon\right) + \left(x_{it} - \upsilon \bar{x}_i\right)' \beta + u_{it} \tag{18}$$

- The random effects estimator is more efficient than the fixed effects estimator (as long as $\psi > 0$; $\upsilon = 1 - \sqrt{\psi}$)
- The efficiency gain results from the use of the between-groups variation $\left(\bar{x}_{i.} - \bar{x}\right)$
- For the fixed effects model we can think that the results are only valid for the units included in the sample used in the estimation
- How far have the units/individuals come from a big population?

Random effects model (cont.)

- For the random effects model we have fewer parameters (compared to the fixed effects estimator) to estimate
- This procedure has a cost related with potential bias of the estimator when the underlying assumptions are not valid

$$y_{it} = x_{it}^{'}\beta + (\alpha + u_i) + \varepsilon_{it} \tag{19}$$

where u_i represents the random heterogeneity specific to unit i, and constant over time,

$$\begin{aligned}
E[\varepsilon_{it}|X] &= E[u_i|X] = 0 \\
E\left[\varepsilon_{it}^2|X\right] &= \sigma_\varepsilon^2 \\
E\left[u_i^2|X\right] &= \sigma_u^2 \\
E[\varepsilon_{it}u_j|X] &= 0, \forall i, t, j \\
E[\varepsilon_{it}\varepsilon_{js}|X] &= 0, \forall t \neq s \text{ and } i \neq j \\
E[u_i u_j|X] &= 0, \forall i \neq j
\end{aligned}$$

Random effects model (cont.)

Composite error term: $\eta_{it} = \varepsilon_{it} + u_i$

$$
\begin{aligned}
E\left[\eta_{it}^2 | X\right] &= \sigma_\varepsilon^2 + \sigma_u^2 \\
E\left[\eta_{it}\eta_{is} | X\right] &= \sigma_u^2,\ t \neq s \\
E\left[\eta_{it}\eta_{js} | X\right] &= 0,\ \forall t, s \text{ if } i \neq j \\
\boldsymbol{\eta}_i &= \left[\eta_{i1}, \eta_{i2}, ..., \eta_{iT}\right]' \\
\Sigma &= E\left[\boldsymbol{\eta}_i \boldsymbol{\eta}_i' | X\right],\ T \text{ observations for unit } i\text{:}
\end{aligned}
$$

$$
\begin{aligned}
\Sigma &= \begin{bmatrix} \sigma_\varepsilon^2 + \sigma_u^2 & \sigma_u^2 & \sigma_u^2 & \cdots & \sigma_u^2 \\ \sigma_u^2 & \sigma_\varepsilon^2 + \sigma_u^2 & \sigma_u^2 & \cdots & \sigma_u^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_u^2 & \sigma_u^2 & \sigma_u^2 & \cdots & \sigma_\varepsilon^2 + \sigma_u^2 \end{bmatrix} \\
&= \sigma_\varepsilon^2 I_T + \sigma_u^2 i_T i_T'
\end{aligned}
$$

where i_T is a column vector of dimension $(T \times 1)$ containing $1's$.

Random effects model (cont.)

- We assume independence between observations i and j
- The Variance-covariance matrix for the error term for the complete data ($n \times T$ observations) is defined as

$$\Omega = \begin{bmatrix} \Sigma & 0 & \cdots & 0 \\ 0 & \Sigma & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma \end{bmatrix} = I_n \otimes \Sigma$$

- We can transform the data and apply OLS to the transformed information (GLS procedure: *Generalized Least Squares*)

$$\hat{\beta} = \left(X'\Omega^{-1}X\right)^{-1} X'\Omega^{-1}Y = \left(\sum_{i=1}^{n} X_i'\Omega^{-1}X_i\right)^{-1} \left(\sum_{i=1}^{n} X_i'\Omega^{-1}Y_i\right) \tag{20}$$

- This formulation depends on the knowledge of the Variance-covariance matrix
- This is a regression of partial deviations of y_{it} on partial deviations of the regressors

Random effects model (cont.)

- OLS inefficiency results from the inefficient weighting matrix
- Compared to the GLS, the OLS gives too much weight to the between-groups variation

$$
\begin{aligned}
\Omega^{-1/2} &= [I_n \otimes \Sigma]^{-1/2} \\
\Sigma^{-1/2} &= \frac{1}{\sigma_\varepsilon}\left[I - \frac{\upsilon}{T} i_T i_T'\right] \\
\upsilon &= 1 - \frac{\sigma_\varepsilon}{\sqrt{\sigma_\varepsilon^2 + T\sigma_u^2}}
\end{aligned}
$$

$$
\Sigma^{-1/2} y_i = \frac{1}{\sigma_\varepsilon}\begin{bmatrix} y_{i1} - \upsilon\bar{y}_{i.} \\ y_{i2} - \upsilon\bar{y}_{i.} \\ \vdots \\ y_{iT} - \upsilon\bar{y}_{i.} \end{bmatrix}
$$

- When $\upsilon = 1$ we have the *LSDV* estimator (Fixed effects model)

Random effects model (cont.)

- The $\hat{\beta}_{GLS}$ is a weighted average of the within- and between-groups estimators

$$
\begin{aligned}
\hat{\beta} &= \hat{F}^{within}\hat{\beta}^{within} + \left(I - \hat{F}^{within}\right)\hat{\beta}^{between} \\
\hat{F}^{within} &= \left[S_{XX}^{within} + \psi S_{XX}^{between}\right]^{-1} S_{XX}^{within} \\
\psi &= \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_u^2} = (1 - \upsilon)^2
\end{aligned}
$$

- If $\psi = 1$ we have $GLS = OLS$. In this case $\sigma_u^2 = 0$.
- If $\psi = 0$ we have $GLS = LSDV$. In this case $\upsilon = 1$.
- (1) $\sigma_\varepsilon^2 = 0 \Rightarrow$ all the variation results from the u_i's. So, we cannot distinguish between the fixed effects model and the random effects model
- (1) $T \to \infty$: if T goes to infinity the unobserved u_i "becomes observable"

Random effects model (cont.)

- When the Variance-covariance components are unknown we should use the *FGLS* (Feasible GLS), as we have to estimate the variances

$$\hat{\sigma}_\varepsilon^2 = s^2_{LSDV} = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T} e_{it}^2}{nT - n - k}$$

- There are different ways to estimate σ_u^2
- Using the within-groups estimator we get a consistent estimator for

$$\sigma_B^2 = \sigma_u^2 + \frac{\sigma_\varepsilon^2}{T} \rightarrow \hat{\sigma}_u^2 = \hat{\sigma}_B^2 - \frac{1}{T}\hat{\sigma}_\varepsilon^2$$

- In some cases, the estimators used for σ_u^2 induce a negative result
- Since we only need consistent estimators we can abandon the correction for the degrees of freedom when computing $\hat{\sigma}_\varepsilon^2$ and $\hat{\sigma}_B^2$ (this can be obtained from the residuals of the within-groups estimator)

Random effects model (cont.)

- The random effects model is not conditional on the values of u_i
- There can be other specification problems which might justify rejection of the null H_0

Another random effects test: Breush and Pagan, 1980

- This is a Lagrange multiplier type of test, which is based on OLS residuals

$$H_0 : \sigma_u^2 = 0 \quad (corr(\eta_{it}, \eta_{is}) = 0)$$
$$H_1 : \sigma_u^2 \neq 0$$

$$\begin{aligned} LM &= \frac{nT}{2(T-1)} \left[\frac{\sum_{i=1}^{n} \left[\sum_{t=1}^{T} e_{it} \right]^2}{\sum_{i=1}^{n} \sum_{t=1}^{T} e_{it}^2} - 1 \right]^2 \\ &= \frac{nT}{2(T-1)} \left[\frac{\sum_{i=1}^{n} \left[T\bar{e}_{i.} \right]^2}{\sum_{i=1}^{n} \sum_{t=1}^{T} e_{it}^2} - 1 \right]^2 \sim \chi^2_{(1)} \end{aligned}$$

Test for random effects: the Hausman test

- Which model should we use? FE or RE?
- The LSDV implies a significant loss in the degrees of freedom
- The random effects has the issue of inconsistency associated with possible correlation between regressors and the specific effect
- A test for the independence between the regressors and the individual effect u_i can be implemented in the following way:
- Under the null hypothesis lack of correlation is correct, so OLS, LSDV and GLS are consistent estimator, although OLS is inefficient
- Under the null the estimates should not differ in a systematic way. The statistic of the test is defined as

$$w = \left[b - \hat{\beta}\right]' \left[Var\left(b\right) - Var\left(\hat{\beta}\right)\right]^{-1} \left[b - \hat{\beta}\right] \sim \chi^2_{(k)} \qquad (21)$$

where k is the number of elements in b, and, under the null hypothesis, b is a consistent estimator and $\hat{\beta}$ is an efficient estimator

Hypothesis testing: Wald test

$$\begin{aligned} H_0 &: R\beta = r \\ H_1 &: H_0 \text{ is not true} \end{aligned}$$

where R is a matrix of dimension $q \times k$, $q \leq k$, q represents the number of restrictions over the vector β of dimension $k \times 1$, and r is a vector of dimension $q \times 1$ with known constants.
Wald test's statistic:

$$W = \left(R\hat{\beta} - r\right)' \left[R\left(X'X\right)^{-1} R'\right]^{-1} \left(R\hat{\beta} - r\right) / \hat{\sigma}^2 \tag{22}$$

- Under H_0, $W \sim \chi^2_q$
- If $\hat{\sigma}^2 = SSR/\left(n-k\right)$, where k is the number of coefficients to be estimated, $W/q \sim F\left(q, n-k\right)$

From a single fixed effects model (...)

- Suppose that you want to estimate the model

$$y_{it} = \mu + \mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{u}_i\boldsymbol{\eta} + \alpha_i + u_{it}$$

  where you have observations of multiple individuals observed over time. The subscript i indexes individual and t stands for time. Strict exogeneity, $E(u_{it}|\mathbf{x}'_{it}, \mathbf{u}_i, \alpha_i) = 0$, is assumed.
  $\mathbf{x}_{it}-$ time-varying individual level observed explanatory variables
  $\mathbf{u}_i-$ time invariant individual level observed explanatory variables
  α_i-time invariant individual level unobserved explanatory variables
- The vector $\boldsymbol{\beta}$ can be estimated by running the regression

$$y_{it} - \bar{y}_i = (\mathbf{x}'_{it} - \bar{\mathbf{x}}'_i)\boldsymbol{\beta} + \widetilde{u}_{it}$$

- The estimates for $\boldsymbol{\beta}$ are the same as if we had included a dummy variable per individual
- time invariant individual level observed variables are absorbed
- This will work regardless of the number of individuals (high-dimensional)

(...) to a two high dimensional fixed effects

- An example: wage regression for employee-employer data:

$$y_{it} = \mu + \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{w}_{j(i,t)}\gamma + \mathbf{u}_i\boldsymbol{\eta} + \mathbf{q}_{j(i,t)}\boldsymbol{\rho} + \alpha_i + \phi_{j(i,t)} + \mu_t + u_{it}$$

y_{it}—wage of worker i at time t
$\mathbf{x}_{it}$— time-varying worker observed explanatory variables
$\mathbf{w}_{j(i,t)}$—time-varying firm observed explanatory variables
$\mathbf{u}_i$— time invariant worker observed explanatory variables
$\mathbf{q}_{j(i,t)}$—time invariant firm observed explanatory variables
α_i—time invariant worker unobserved explanatory variables
$\phi_{j(i,t)}$—time invariant firm unobserved explanatory variables
μ_t— unobserved time effect
u_{it}—usual error term

- In practice we have

$$y_{it} = \mu + \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{w}_{j(i,t)}\gamma + \theta_i + \psi_{j(i,t)} + \mu_t + u_{it}$$

where $\theta_i \equiv \alpha_i + \mathbf{u}_i\boldsymbol{\eta}$ and $\psi_j \equiv \mathbf{q}_j\boldsymbol{\rho} + \phi_j$

Spell Fixed Effects

- Combine all fixed effects into a single fixed effect
- Example: calculate unique combination of worker-firm (the spell)
- Treat the model as if the spell is the (single) fixed effect
- This is fine if our interest is on $\boldsymbol{\beta}$ and γ
- But note that we do not obtain estimates for θ_i and ψ_j
- With this approach we are controlling for θ_i, ψ_j and their interactions
- The same as running the regression

$$y_{it} = \mu + \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{w}_{j(i,t)}\gamma + \lambda_{s(i,t)} + \mu_t + u_{it}$$

where λ_s is a fixed effect that captures the spell

Iterative Procedures

- Rewrite the two fixed-effects model in matrix form:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{D}_1\boldsymbol{\alpha} + \mathbf{D}_2\gamma + \epsilon$$

- $\mathbf{D}_1$ is $n \times G_1$ and $\mathbf{D}_2$ is $n \times G_2$ and both G_1 and G_2 are large numbers
- Direct estimation of this model is complicated
- But a "zigzag" approach is simple to implement:

$$\begin{bmatrix} \beta^{(j+1)} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\left(\mathbf{Y} - \mathbf{D}_1\boldsymbol{\alpha}^{(j)} - \mathbf{D}_2\gamma^{(j)}\right) \\ \boldsymbol{\alpha}^{(j)} = \left(\mathbf{D}_1'\mathbf{D}_1\right)^{-1}\mathbf{D}_1'\left(\mathbf{Y} - \mathbf{X}\beta^{(j)} - \mathbf{D}_2\gamma^{(j)}\right) \\ \gamma^{(j)} = \left(\mathbf{D}_2'\mathbf{D}_2\right)^{-1}\mathbf{D}_2'\left(\mathbf{Y} - \mathbf{X}\beta^{(j)} - \mathbf{D}_1\boldsymbol{\alpha}^{(j)}\right) \end{bmatrix}$$

- Standard errors (clustered or not) can also be easily calculated
- To degrees of freedom of the regression are

$$dof = n - (k + G_1 + G_2 - M)$$

where M is the number of mobility groups

The problem of identification: One fixed effect

id
1
1
2
2
3
3

$cons$	$other$	$id = 1$	$id = 2$	$id = 3$
1	.	1	0	0
1	.	1	0	0
1	.	0	1	0
1	.	0	1	0
1	.	0	0	1
1	.	0	0	1

The problem of identification: two fixed effects (case1)

$id1$	$id2$
1	1
1	2
2	1
2	2
3	1
3	2

$id1=1$	$id1=2$	$id1=3$	$id2=1$	$id2=2$
1	0	0	1	0
1	0	0	0	1
0	1	0	1	0
0	1	0	0	1
0	0	1	1	0
0	0	1	0	1

The problem of identification: two fixed effects (case2)

$id1$	$id2$
1	1
1	2
2	1
2	2
3	3
3	4

$id1 = 1$	$id1 = 2$	$id1 = 3$	$id2 = 1$	$id2 = 2$	$id2 = 3$	$id2 = 4$
1	0	0	1	0	0	0
1	0	0	0	1	0	0
0	1	0	1	0	0	0
0	1	0	0	1	0	0
0	0	1	0	0	1	0
0	0	1	0	0	0	1

General considerations

- Remember that with short-panels the estimates of the fes may be inconsistent
- But averages, kernel-densities, etc should be fine!
- With 2 hdfe estimates of the fes may only be compared within each mobility group. That is why researchers work with the "largest connected set"
- Correlation between the estimates of the fixed effects may be biased (eg: limited mobility)
- The iterative technique can be extended to several sets of fixed effects or even to interacted fixed effects
- With more than 2 hdfes the exact degrees of freedom may not be known. But do we really care?

Stata commands for estimation of high dimensional models

- Models with 2hdfe
  - `areg` - Implements the exact least squares solution proposed by Abowd, Creecy and Kramarz (2002). Does not compute standard errors - by Amine Ouazad
  - `felsdvreg` - Uses a "memory-saving" procedure - by Thomas Cornelissen
  - `reg2hdfe` - uses iterative procedure - by Paulo Guimarães
  - `gpreg` - another implementation of the iterative procedure - by Johannes F. Schmieder
- Models with interacted hdfe
  - `regintfe` - estimates models with one high-dimensional interacted fe - by Paulo Guimarães
- The gold standard!
  - `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

Advice for estimation

- Prepare a "clean dataset"
- Use `reghdfe`!
- Singletons should be dropped (default on `reghdfe`)
- If you use clustered standard errors make sure the number of clusters is high enough (+50)
- If you plan on doing secondary analysis of fes restrict your data to a "connected set"
- If you have a large data set then:
  - read the `reghdfe` help file
  - take advantage of multiple cores on your computer
  - use a lower convergence criterion
  - be patient!

References

- Abowd, J., Kramarz, F. and Margolis, D. (1999), "High wage workers and high wage firms", Econometrica, 67(2), 251-233.
- Andrews, M., Schank, T., and Upward, R. (2006), "Practical fixed-effects estimation methods for the three-way error-components model", Stata Journal, 6(4), 461-481.
- Arellano, M. (2003), Panel Data Econometrics, Oxford University Press: New York.
- Correia, S. (2015), "Singletons, cluster robust standard errors and fixed effects: a bad mix", unpublished paper.
- Greene, W. H. (2017), Econometric Analysis, 8th ed., Pearson: New York.
- 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.
- Verbeek, M. (2017), A Guide to Modern Econometrics, 5th ed., John Wiley & Sons, Ltd.: Chichester, England.
- Wooldridge, J. (2010), Econometric Analysis of Cross Section and Panel Data, 2nd ed., The MIT Press: Cambridge, Massachusetts.