Sustainability of Public Policy

Dott.ssa Rossella Iraci Capuccinello

Università degli Studi di Ferrara

a.a. 2016-2017

Panel Data

"A longitudinal, or panel, data set is one that follows a given sample of individuals over time, and thus provides multiple observations on each individual in the sample." (Hsiao,2003)

More generally speaking it follows over time a given sample of entities (individuals, countries, firms, schools, etc...) and are also known as longitudinal or cross-sectional time-series

Data Example

	country	year	y	y_bin	x1	x2	x3	opinion
1	A	1990	1.343e+09	1	.2779036	-1.107956	.2825536	Str agree
2	A	1991	-1.900e+09	0	.3206847	-.94872	.4925385	Disag
3	A	1992	-11234363	0	.3634657	-.789484	.7025234	Disag
4	A	1993	2.646e+09	1	.246144	-.885533	-.0943909	Disag
5	A	1994	3.008e+09	1	.424623	-.7297683	.9461306	Disag
6	A	1995	3.230e+09	1	.4772141	-.723246	1.02968	Str agree
7	A	1996	2.757e+09	1	.499805	-.7815716	1.092288	Disag
8	A	1997	2.772e+09	1	.0516284	-.7048455	1.415901	Str agree
9	A	1998	3.397e+09	1	.3664108	-.6983712	1.548723	Disag
10	A	1999	39770336	1	.3958425	-.643154	1.794198	Str disag
11	B	1990	-5.935e+09	0	-.08185	1.42512	.0234281	Agree
12	B	1991	-7.116e+08	0	.10616	1.649602	.2603625	Str agree
13	B	1992	-1.933e+09	0	.3537852	1.593719	-.2343988	Agree
14	B	1993	3.073e+09	1	.726777	1.691758	.2562243	Str disag

Notation

- Individual or cross section unit : country, region, state, firm, consumer, individual, student, couple of individuals or countries
- Double index : i (for cross-section unit) and t (for time)
  y_{it} for $i = 1, ..., N$ and t for $t = 1, ..., T$

Micro and Macro Panel Data

- A Micro Panel Dataset is one for which the individual dimension is much larger than the time one, $T < N$
- In a Macro Panel Dataset the time dimension is similar to the individual dimension, $T \simeq N$

Balanced and Unbalanced Panel

- A Panel is balanced if for each individual/entity we have the same time periods
- A Panel is unbalanced when the time dimension is specific to each individual/entity
- When dealing with an unbalanced Panel we need to uncover why the panel is unbalanced. The reason may be related to existence of sample selection or attrition.

Advantages of Panel Data

- They provide a large number of data points reducing collinearity among covariates
- Allow to analyse some issues that cannot be investigated with cross-sectional data
- Allow to isolate the effects of specific actions, treatments, or more general policies.
- Allow to address the omitted variable problem
- In some time series analysis the availability of panel data can simplify the estimation

Addressing omitted variable bias

Example 1:

$$y_{it} = \alpha + \beta' x_{it} + \rho' z_{it} + \epsilon_{it}$$

Let's assume that z_{it} is unobservable and correlated with x_{it}, $cov(x_{it}, z_{it}) \neq 0$;
Assuming $z_{it} = z_i$

$$y_{it} = \alpha + \beta' x_{it} + \rho' z_i + \epsilon_{it}$$

We can take the first difference of individual observations over time

$$y_{it} - y_{i,t-1} = \beta'(x_{it} - x_{i,t-1}) + \epsilon_{it} - \epsilon_{i,t-1}$$

Least squares regression now provides unbiased and consistent estimates of β

Addressing omitted variable bias

Example 2:

$$\text{Assuming } z_{it} = z_t$$

$$y_{it} = \alpha + \beta' x_{it} + \rho' z_t + \epsilon_{it}$$

We can take the deviation from the mean across individuals at a given time

$$y_{it} - \bar{x}_t = \beta'(x_{it} - \bar{x}_t) + \epsilon_{it} - \bar{\epsilon_t}$$

Least squares regression now provides unbiased and consistent estimates of β

Heterogeneity Bias

- One of the main issues occurring when dealing with panel data is Heterogeneity Bias
- Ignoring the individual or time-specific effects that exist among cross-sectional or time-series units but are not captured by the included explanatory variables can lead to parameter heterogeneity in the model specification.

Heterogeneity Bias

Example: Cobb Douglas production function with two factors (labor and capital). We have N countries and T periods.
Let us denote:

- $y_{it} = \alpha_i + \beta_i k_{it} + \gamma_i n_{it} + \epsilon_{it}$
- y_{it} the log of the GDP for country i at time t.
- n_{it} the log of the labor employment for country i at time t.
- y_{it} the log of the capital stock for country i at time t.

Homogeneous specification

Here the elasticities α_i and β_i are specific to each country but, several alternative specifications can be considered. First, we can assume that the production function is the same for all countries: in this case we have an homogeneous specification:

$$y_{it} = \alpha + \beta k_{it} + \gamma n_{it} + \epsilon_{it}$$

$$\alpha_i = \alpha,\ \beta_i = \beta,\ \gamma_i = \gamma$$

Homogeneous Specification

An homogeneous specification of the production function for macro aggregated data is meaningless. We can assume that the mean of TFP (given by $E(\alpha_i + \epsilon_i, t) = \alpha_i$) is different across countries (heterogeneity of the Total Factor Productivity)

We can use a specification with individual effects, α_i and common slope parameters (elasticities β and γ).

$$y_{it} = \alpha_i + \beta k_{it} + \gamma n_{it} + \epsilon_{it}$$

$$\beta_i = \beta, \gamma_i = \gamma$$

Heterogeneous Specification

Finally, we can assume that the labor and/or capital elasticities are different across countries. In this case, we will have an heterogeneous specification of the panel data model (heterogeneous panel)

$$y_{it} = \alpha_i + \beta_i k_{it} + \gamma_i n_{it} + \epsilon_{it}$$

In this case, there are two solutions: 1) Using N times series models to produce some group-mean estimates of the elasticities. 2) Using a model with random (slope) parameters =¿ *random coefficient model*. In this case, we assume that parameters β_i and γ_i and randomly distributed, but follow the same distribution:

$$\beta_i \text{ i.i.i } (\bar{\beta}, \sigma_\beta^2) \; \gamma_i \text{ i.i.i } (\bar{\gamma}, \sigma_\gamma^2)$$

Heterogeneity Bias

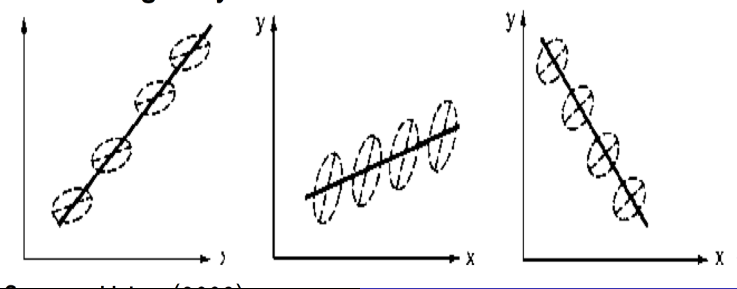
Ignoring such heterogeneity could lead to biased estimates.
Consider a simple linear model with individual effects and only one independent variable x_i (common slope).

$$y_{it} = \alpha_i + \beta x_{it} + \epsilon_{it}$$

Assume that all $N * T$ observations are used to estimate the homogeneous model:

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

The heterogeneity bias.



Heterogeneity Bias

- All of these figures depict situations in which biases (on $\hat{\beta}$) arise in pooled least-squares estimates because of heterogeneous intercepts
- Pooled regression ignoring heterogeneous intercepts should never be used
- Difficult to identify the direction of the bias of the pooled slope estimates

Definitions:

- An heterogeneous panel data model is a model in which all parameters (constant and slope coefficients) vary across individuals.
- An homogeneous panel data model (or pooled model) is a model in which all parameters (constant and slope coefficients) are common

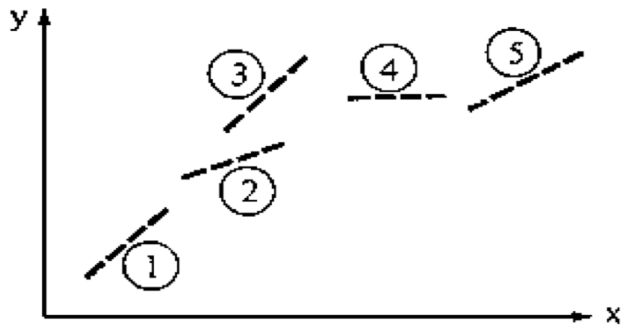
Heterogeneity Bias

Now consider another model

$$y_{it} = \alpha_i + \beta_i x_{it} + \epsilon_{it}$$

Assume that all $N * T$ observations are used to estimate the homogeneous model:

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



Specification tests

We can test the estimated regression coefficients in 3 possible ways:

- the homogeneity of regression slope coefficients
- the homogeneity of regression intercept coefficients.
- the time stability of parameters (slopes and constants). We will not consider this issue (not specific to panel data models) here.

Specification tests

We assume that parameters are constant over time, but can vary across individuals.

$$y_{it} = \alpha_i + \beta_i' x_{it} + \epsilon_{it}$$

Three types of restrictions can be imposed on this model.
Regression slope coefficients are identical, and intercepts are not (model with individual / unobserved effects).

$$y_{it} = \alpha_i + \beta' x_{it} + \epsilon_{it}$$

Regression intercepts are the same, and slope coefficients are not (unusual).

$$y_{it} = \alpha + \beta_i' x_{it} + \epsilon_{it}$$

Both slope and intercept coefficients are the same (homogeneous / pooled panel).

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

Fixed Effects vs. Random Effects

In the traditional approach to panel data models, α_i is called a Random Effects when it is treated as a random variable and a Fixed Effects when it is treated as a parameter to be estimated for each cross section observation i.

- Use fixed-effects (FE) whenever you are only interested in analysing the impact of variables that vary over time.
- You should use fixed effects to control for unobserved heterogeneity when heterogeneity is constant over time and correlated with independent variables.

"Unobserved heterogeneity" refers to omitted variables that are fixed for an individual or an entity (at least over a long period of time). For example: demographics, innate ability, culture, etc...

Distinct Intercept DGP

In a model with individual unobserved effect, we allow each individual *i* to have a different intercept α_i. This intercept accounts for all aspects of unobserved heterogeneity that are fixed over time.
This is called "Distinct Intercept" DGP. It is suitable for panel of States, countries, regions, schools, firms...

Error Component Model

- With longitudinal data on individual workers, students or consumers, we draw a different set of individuals from the population each time we collect a sample.
- Each individual has his/her own set of fixed omitted variables.
- We cannot fix each individual intercept.

Error Component Model

$$y_{it} = \alpha + \beta x_{it} + v_i + \mu_{it}$$

- In this DGP we have the same intercept and slope but separate the error term in two components $v_i + \mu_{it}$
- v_i is fixed for each individual in all time periods
- It includes all fixed omitted variables (i.e.: gender, ethnicity, innate ability, etc..)

Error Component Model

- In the Distinct Intercepts DGP, the unobserved heterogeneity is absorbed into the individual-specific intercept α_i
- In the second DGP, the unobserved heterogeneity is absorbed into the individual fixed component of the error term, v_i

Depending on $E(X_{it}, v_i)$ the OLS can produce unbiased and consistent estimates. This is true when $E(X_{it}, v_i) = 0$ In this case, unobserved heterogeneity is uncorrelated with the explanatory variables.
If $E(X_{it}, v_i) \neq 0$ unobserved heterogeneity is correlated with the explanatory variables and OLS is Biased and Inconsistent

Fixed Effect Estimator

- Using Panel Data we can create a consistent and unbiased estimator, the Fixed Effect Estimator.
- Used with either the distinct intercepts DGP or the error components DGP with $E(X_{it}, v_i) \neq 0$
- Fixed Effects assumes that the individual specific effect is correlated to the independent variable.
- Basic Idea: estimate a separate intercept for each individual
- The simplest way to do so is to use dummy variables (LSDV estimator)

Least Squares Dummy Variable Estimator

The least squares dummy variable estimator in practice

- Create a set of n dummy variables, D_j, such that $D_j = 1$ if $i = j$
- Regress Y_{it} against all the dummies, X_t , and X_{it} variables (you must omit X_i variables and the constant).
- n is sometimes too large for a computer to handle

Least Squares Dummy Variable Estimator in Stata

xtset country year
xi: regress y x1 i.country
predict yhat
or areg y x1, absorb(country)

- The effect of x1 is mediated by the differences across countries.
- By adding the dummy for each country we are estimating the pure effect of x1
- Each dummy is absorbing the effects particular to each country.

Fixed Effect Estimator

- The Fixed Effect Estimator is a less computationally intensive alternative to LSDV
- Even though both methods are usually referred to as Fixed Effect, technically this is the Fixed Effect Estimator while the previously illustrated technique is the LSDV
- The intuition behind the FE Estimator is that if we difference observations ofr the same individual, v_i cancels out

$$y_{it} = \alpha + \beta x_{it} + v_i + \mu_{it}$$
$$-y_{it'} = \alpha + \beta x_{it'} + v_i + \mu_{it'}$$
$$\rightarrow y_{it} - y_{it'} = 0 + \beta(x_{it} - x_{it'}) + 0 + \mu_{it} - \mu_{it'}$$

- OLS would be a consistent estimator of β

Fixed Effect Estimator

- If $T = 2$ we have only 2 observations for each individual/entity and differencing the two of them is efficient
- it $T > 2$ differencing any 2 observations ignores valuable information in the other observations for the same individual/entity
- We can use all the observations for each individual if we subtract the individual-specific mean from each observation.

$$y_{it} = \alpha + \beta x_{it} + v_i + \mu_{it}$$
$$-\bar{y_i} = \alpha + \beta \bar{x_{it}} + v_i + \bar{\mu_{it}}$$
$$\rightarrow y_{it} - \bar{y_i} = 0 + \beta(x_{it} - \bar{x_{it}}) + 0 + \mu_{it} - \bar{\mu_{it}}$$

- The Fixed Effects and DVLS estimators provide exactly identical estimates.

Fixed Effect Estimator

- Fixed effect discard all variations between individuals
- It only uses variation over time within individuals
- Within Estimator
- Fixed Effects uses n degrees of freedom.
- is not efficient if $E(X_{it}, v_i) = 0$
- In this case we cannot use OLS because there is serial correlation within individuals and OLS would be inefficient

Fixed Effect Estimator using Stata

xtreg y x1,fe
The 3 methods provide the same results.

The fixed-effects model controls for all time-invariant differences between the individuals, so the estimated coefficients of the fixed-effects models cannot be biased because of omitted time-invariant characteristics...[like culture, religion, gender, race, etc]

However, we cannot use fixed effect estimator to investigate the effect of time-invariant characteristics on the dependent variable. Moreover, fixed effects will not work well with data for which within-cluster variation is minimal.

Time Fixed Effects

To see if time fixed effects are needed when running a FE model use the command testparm.

It is a joint test to see if the dummies for all years are equal to 0, if they are then no time fixed effects are needed

After running the fixed effect model, type:

testparm i.year

The $Prob > F$ is > 0.05, so we failed to reject the null that the coefficients for all years are jointly equal to zero, therefore no time fixed-effects are needed in this case.

Random Effects

- When $E(X_{it}, v_i) = 0$ we use the Random Effect Estimator to deal with the serial correlation in panel data
- An advantage of random effects is that you can include time invariant variables (i.e. gender). In the fixed effects model these variables are absorbed by the intercept.
- In random-effects you need to specify those individual characteristics that may or may not influence the predictor variables $\rightarrow$ omitted variable bias.
- RE allows to generalise the inferences beyond the sample used in the model.
- The RE estimator provides more precise estimates

Random Effects Estimation

The key idea of random effects:

- Estimate σ_v^2 and σ_μ^2
- Use these estimates to construct efficient weights of panel data observations
- In Stata:
  xtreg y x1, re

Random or Fixed Effects?

To decide between fixed or random effects you can run a Hausman test where the null hypothesis is that the preferred model is random effects vs. the fixed effects (see Green, 2008). It basically tests whether the unique errors (V_i) are correlated with the regressors, the null hypothesis is they are not.

In Stata: Run a fixed effects model and save the estimates, then run a random model and save the estimates, then perform the test.

```
xtreg y x1, fe
estimates store fixed
xtreg y x1, re
estimates store random
hausman fixed random
```

If "Prob> *chi*2 =" this is < 0.05 (i.e. significant) use fixed effects.

Testing for random effects: Breusch-Pagan Lagrange multiplier (LM)

- The LM test helps decide between a random effects regression and a simple OLS regression.
- The null hypothesis in the LM test is that variance across entities is zero. This is equivalent to no significant difference across units
- The command in Stata is xttset0 type it right after running the random effects model.
- If "Prob> *chi*2" is > 0.05 we fail to reject the null and conclude that random effects is not appropriate

Testing for heteroskedasticity

- A test for heteroskedasticity is available for the fixed- effects model using the command xttest3.
- It is a user-written program
- ssc install xtest3
  xttest3
- We find "Prob> $chi2 = 0.000$" therefore there is heteroskedasticity
- Use the option ÔrobustÕ to obtain heteroskedasticity-robust standard errors