LUISS BUSINESS SCHOOL

EFMD EQUIS ACCREDITED

Specialised Master Programmes

Econometrics

Econometrics

Giuseppe Brandi

LUISS

June 2018

PANEL DATA - THEORY

A Map

- **Cross Section**
  Multiple entities observed once.

- **Time Series**
  One entity observed multiple times.

- **Panel (Longitudinal) Data**
  Multiple entities observed multiple times.

- **Independently Pooled Cross Section**
  Multiple entities observed once, but in different times.

Independently Pooled Cross Section

Why pooling cross sections?

- To increase sample size $\rightarrow$ more precise estimators.
- To investigate the effect of time (simply add dummy variables)

Each *i* is independent from the other.

What if we have Panel Data - i.e. multiple observations observed multiple times?

Example

- Data on 50 U.S. states, each state is observed in 3 years, for a total of 150 observations.
- Data on 1000 individuals, in four different months, for 4000 observations.
- ...

Formats

Panel Data are available in two formats:

- Long format: one record for each subject for each time point.
- Wide format: one record for each subject.

Table: Long

id	time	tick	ed
A	1	1200	9
A	2	2012	10.2
B	1	1310	6.5
B	2	2166	6.7
C	1	1118	10
C	1	1001	11

Table: Wide

id	tick_1	tick_2	ed_1	ed_2
A	1200	2012	9	10.2
B	1310	2166	6.5	6.7
C	1118	1001	10	11

In *R, melt*.

Why Panel Data?

- It increases your sample size.
- It allows to esimate time effects.

But there is much more!

- We can control for unobserved heterogeneity - factors that do not change over time, or change slowly (while they changes for different entities).

Why Panel Data? cont.ed

If an omitted variable does not change over time, then any change in Y cannot be caused by it!

- α_i is called unobserved effect or unobserved/individual heterogeneity.
- There are other unobservable factors that vary over time and across entities. We call them shocks or idiosyncratic errors, $u_{i,t}$.

Notation

$y_{i,t}$ is the response variable for entity i at time t.

$x_{i,t,k}$ is the explanatory variable k for entity i at time t.

An explanatory variable can be

- time-varying: hours worked, income
- time-invariant ($x_{it} = x_i$): birth country, ethnicity, adults' education level...

This distinction will prove useful: keep it in mind.

Balanced panel

In this course we deal with *balanced* panel data.

A balanced panel is a panel where we observe the same time periods for each entity/unit.

- Easier to achieve for large entities (states, cities).
- Following small units (individuals, small firms) over time can be challenging.
  - Attrition ("dangerous" loss of partecipants). The reason some entities have left the sample may be correlated with unobserved factors.
  - Ex: in a test of a dieting program, those who drops out of the trial may be those for whom it was not working $\rightarrow$ biased results.

Two periods example

- $Y_{i,1} = \beta_0 + \beta_1 X_{i,1} + \alpha_i + u_{i,1} = \beta_0 + \beta_1 X_{i,1} + v_{i,1}$
- $Y_{i,2} = \beta_0 + \beta_1 X_{i,2} + \alpha_i + u_{i,2} = \beta_0 + \beta_1 X_{i,2} + v_{i,2}$

$v_{i,t} = \alpha_i + u_{i,t}$ is called composite error.

Let's take the difference

- $\Delta Y = \beta_1 \Delta X + \Delta u$

where $\Delta Y = Y_{i,2} - Y_{i,1}$ and similarly for ΔX and Δu.

The unobserved heterogeneity α_i is differenced away.

- The interpretation of the coefficients remains the same.
- Inference procedures (t-tests, p-values) remain the same.
  - We need to use the correct standard errors.

Example T=2 - Data

Use the rental dataset:

https://ideas.repec.org/p/boc/bocins/rental.html

Google "Rental.dta IDEAS".

We have data on:

- percentage of population students in the city, **pctstu**
- average log rent, **lrent**
- per capita city log income, **lavginc**
- **year**, 1980 or 1990 (or **y90**=1 if 1990)
- **city**, 64 USA cities.

Example T=2

- $pcstu_{i,80} = \beta_0 + \beta_1 lrent_{i,80} + \beta_2 lavginc_{i,80} + \alpha_i + u_{i,80}$
- $pcstu_{i,90} = \beta_0 + \beta_1 lrent_{i,90} + \beta_2 lavginc_{i,90} + \alpha_i + u_{i,90}$
- $\Delta pcstu = \beta_1 \Delta lrent + \beta_2 \Delta lavginc + \Delta u$
- $\Delta pcstu = -2,625 + 9.496 \Delta lrent - 3.611 \Delta lavginc + \Delta u$

R automatically inserts an intercept to account for trends.

Multiple times

For $T > 2$, we have different possible models.

The choice depends first on the assumption on $Corr(\alpha_i, x_{it})$ in

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

However...

The strength of panel data lies in dealing with unobserved heterogeneity.

If we believe that α_i is uncorrelated with explanatory variables, we may not need panel data at all.

APPETIZERS...

Before starting with the description for models in multiple times, two general features of panel data are worth discussing.

- Exogeneity
- Serial correlation

Exogeneity

In *all* the models now presented, strict/strong exogeneity must hold:

$$Corr(u_{it} x_{i,s}) = 0 \text{ for every } t, s = 1, 2, ...T$$

The idiosyncratic error term has zero mean conditional on past, current and future values of the regressors.

- Violation if $x_{i,t}$ is affected/determined by a shock in the past ($u_{i,t-1}$).
- Cannot include lagged dependent variables.

Much stronger than contemporaneous exogeneity:

$$Corr(u_{it} x_{i,t}) = 0 \text{ for every } t = 1, 2, ...T$$

Serial correlation (autocorrelation)

When error terms from different (usually adjacent) time periods are correlated, we say that the error term is serially correlated.

Serial correlation

- does not affect the unbiasedness or consistency of estimators.
- but does affect their standard errors.

SERIAL CORRELATION CONT.ED

Remember: one of the OLS assumptions is iid observations.

- but, with panel data, errors are likely to be correlated over time for the same entity! Not iid!
- NT correlated observations contain less information than NT independent observations.
- You overestimate the amount of information in your hands.
  - Standard errors are going to be downward biased.
  - The precision of your estimates is overstated.

Panel Data - Models

Multiple times: Models

- Pooled OLS
- Random effects
  - Between estimator
- Fixed effect
  - Within estimator
  - First difference
  - Dummies

WORKING EXAMPLE

We use the "Charity" dataset at

http://instruction.bus.wisc.edu/jfrees/jfreesbooks/Longitudinal%20and%20Panel%20Data/Book/DataFiles.htm

You find it googling "Edward Frees google site" under the section "Longitudinal and Panel Data: Analysis and Applications for the Social Sciences"

Example, dataset

Subset of people always reporting contributions bw 1979-1988

- **CHARITY**, log of sum of cash and other property contributions,
- **INCOME**, log of gross income
- **AGE**, dummy equal one if the individual is over 64.
- **MS**, dummy equal one if individual is married
- **SUBJECT**, subject identifier (47 subjects)
- **TIME**, time identifier (10 years)

Pooled OLS

Similar to cross-section analysis. It does not exploit the fact of having a panel data. You have NT observations and "pretend" you deal with a cross section.

$$y_{it} = x'_{it}\beta + v_{it}$$

$v_{it} = \alpha_i + u_{i,t}$, so α_i is simply included in the error.

We are assuming $Corr(\alpha_i, x_{it}) = 0$...

Pooled OLS cont.ed

You can estimate it with usual OLS, but errors are likely to be serially correlated.

We will see how to correct our standard errors. For the moment:

Pooled OLS

Pros: easy + it finds coefficients also for time-invariant regressors.

Cons: inconsistent if $Corr(\alpha_i, x_{it}) \neq 0$ + wrong SE.

In practice

lm(CHARITY INCOME+AGE+MS, data=ch)

$$log(char)_{i,t} = \beta_0 + \beta_1 log(income)_{i,t} + \beta_2 age_{i,t} + \beta_3 married_{i,t} + u_{i,t}$$

$$log(char)_{i,t} = -2.893 + 0.853 log(income)_{i,t} + 1.398 age_{i,t} + 0.430 married_{i,t} + u_{i,t}$$

Random Effects

$$y_{it} = x'_{it}\beta + \alpha_i + u_{it} = x'_{it}\beta + v_{it}$$

Similar to Pooled OLS.

- α_i is put in the error term assuming $Corr(\alpha_i, x_{it}) = 0$...
- we account for the serial correlation in the composite error.
- We use GLS.

Random effects

We assume equicorrelated composite errors: the correlation does not vary with time difference $t - s$:

$$Corr[v_{it} v_{is}] = Corr[(\alpha_i + u_{it})(\alpha_i + u_{is})] = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_u^2}$$

Random effect

Pros: Consistent estimator of all parameters, including those of time-invariant covariates. Efficient is $Corr(\alpha_i, x_{it}) = 0$.

Cons: Strong, usually invalid assumption $\rightarrow$ biased coefficients.

Bottomline: go with fixed effects.

In practice

plm(CHARITY INCOME+AGE+MS, data=ch, model="random")

$$log(char)_{i,t} = \beta_0 + \beta_1 log(income)_{i,t} + \beta_2 age_{i,t} + \beta_3 married_{i,t} + v_{i,t}$$

$$log(char)_{i,t} = -1.236 + 0.729 log(income)_{i,t} + 0.297 age_{i,t} + 0.123 married_{i,t} + v_{i,t}$$

Between estimator

Presented just for completeness. Take averages for each subject over time:

$$\bar{y}_i = \sum_{t=1}^{T} y_{it}$$

Similarly for $\bar{x}_i$, $\bar{u}_i$ and $\bar{\alpha}_i$. Then estimate with OLS:

$$\bar{y}_i = \bar{x}_i' \beta + \bar{\alpha}_i + \bar{u}_i$$

The between estimator exploits variation between entities.

As for random effects, between estimator is biased when $Corr(\alpha_i, x_{it}) \neq 0$, but Random effects is better since it exploits both between and within entity variation and computes the correct SE.

In practice

plm(CHARITY INCOME+AGE+MS, data=ch,
model="between")

$$log(char)_{i,t} = \beta_0 + \beta_1 log(income)_{i,t} + \beta_2 age_{i,t} + \beta_3 married_{i,t} + v_{i,t}$$

$$log(char)_{i,t} = -4.590 + 0.994 log(income)_{i,t} + 2.159 age_{i,t} + 0.611 married_{i,t} + v_{i,t}$$

Fixed effect

Usually, $Corr(\alpha_i, x_{it}) \neq 0$.

Think about individual ability, cultural components...

Two possible approaches:

- Within estimator (sometimes 'fixed effect estimator')
- Dummy variables

WITHIN ESTIMATOR

Take averages for each subject over time:

$$\bar{y}_i = \sum_{t=1}^{T} y_{it}$$

Similarly for $\bar{x}_i$, $\bar{u}_i$ and $\bar{\alpha}_i$. Notice that $\bar{\alpha}_i = \alpha_i$ since this is constant over time.

We estimate the entity-demeaned model:

$$(y_{it} - \bar{y}_i) = (x_{it} - \bar{x}_i)'\beta + (u_{it} - \bar{u}_i)$$

using a pooled GLS (the error term is still serially correlated) - R will do the job for us.

WITHIN ESTIMATOR

$$(y_{it} - \bar{y}_i) = (x_{it} - \bar{x}_i)'\beta + (u_{it} - \bar{u}_i)$$

- The unobserved heterogeneity α_i disappears!
- But so do all time-invariant regressors (gender, ethnicity)...
- The good news is: we get consistent estimates for the coefficient of the time-varying regressors.
- Moreover, we could interact time invariant variables with time-varying ones (such as time dummies).

WITHIN ESTIMATOR

You can take the average over both time and entity and control for all factors that are either

- constant over time but changing across entities (culture, ability)
- constant across entities but changing over time (fashion, some laws)
- but NOT those varying across both time and entities.

We can use entity-demeaning, time-demeaning or both.

Using time-demeaning prevents us from estimating coefficient of covariates that are the same for all entities but change over time (macro-variables, prices).

Within estimator

Pros: Consistent estimators.

Cons: No coefficient found for time-invariant regressors (unless we interact them).

In practice

$$log(char)_{i,t} = \beta_0 + \beta_1 log(income)_{i,t} + \beta_2 age_{i,t} + \beta_3 married_{i,t} + v_{i,t}$$

Entity-demeaning
plm(CHARITY$\sim$ INCOME+AGE+MS, data=ch, model="within",
effect=c("individual"))
$$log(char)_{i,t} = 0.720 log(income)_{i,t} + 0.155 age_{i,t} + 0.067 married_{i,t} + v_{i,t}$$

Time-demeaning
plm(CHARITY$\sim$ INCOME+AGE+MS, data=ch, model="within",
effect=c("time"))
$$log(char)_{i,t} = 0.828 log(income)_{i,t} + 1.375 age_{i,t} + 0.067 married_{i,t} + v_{i,t}$$

Time&entity-demeaning
plm(CHARITY$\sim$ INCOME+AGE+MS, data=ch, model="within",
effect=c("time","individual"))
$$log(char)_{i,t} = 0.513 log(income)_{i,t} - 0.018 age_{i,t} + 0.143 married_{i,t} + v_{i,t}$$

First difference

An alternative method is taking the first difference:

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

Again, the unobserved heterogeneity α_i disappears, and so do all time-invariant regressors.

Differencing we lose one time period.

In practice

lm(CHARITY INCOME+AGE+MS+as.factor(SUBJECT)-1, data=ch)

lm(CHARITY INCOME+AGE+MS+as.factor(TIME)-1, data=ch)

lm(CHARITY INCOME+AGE+MS+as.factor(SUBJECT)+as.factor(TIME)-1,data=ch)

Entity and Time Dummies, cont.ed

- # parameters: $N + T - 1 + dim(x)$
- Consistently estimated if both $T \to \infty$ and $N \to \infty$
- In short panels (small T) $\hat{\delta}_t$ are consistent
- $\hat{\alpha}_i$ creates problems because you add new parameter at each new entity you observe.

Entity and Time Dummies

If you have few entities or the interest lies in the α_i, use Dummy Variables.

With large N, such model is computing power intensive.

Entity and Time Dummies

Pros: Consistent estimators. Correct standard errors.

Cons: No coefficient found for time-invariant regressors (unless interacted).
Parameters $\hat{\alpha}_i$ rarely consistent (but not a problem is they are not our main interest!)

In practice

Individual dummies
lm(CHARITY INCOME+AGE+MS+factor(SUBJECT)-1,
data=ch)

Time-demeaning
lm(CHARITY INCOME+AGE+MS+factor(TIME)-1, data=ch)

Time&entity-demeaning
lm(CHARITY INCOME+AGE+MS+factor(TIME)+factor(SUBJECT)-
1,
data=ch)

Results are those found in the within estimator case.

SUMMING UP ON FIXED EFFECTS

With only two time periods the within estimator, first difference and dummy variables coefficients are identical.

If $T > 2$ the preferred estimator depends on the coefficients you are interested in and the assumptions about u_{it}.

- If you are interested in fixed effects, go with Dummies.
- If you are not interested and
  - u_{it} are iid, ho with Within estimator.
  - u_{it} follows a Random Walk, go with First Difference.

Some more refinements

Estimate both models. If the results are not sensitive, then fine! But if they differ, reason about possible causes.

In practice, the within estimator is used.

- With unbalanced panel data, for each missing observation First difference loses two observations!
- Measurement errors is a more important problem with the Within estimator.

Panel Data - Practice

Dataset

Use the Jtrain dataset:

https://ideas.repec.org/p/boc/bocins/jtrain.html

Google "Jtrain.dta IDEAS".

VARIABLES

- **hrsemp**, hours of training per employee, at the firm level;
- **grant**, dummy equal to 1 if the firm received a grant during that year;
- **year**, 1987, 1988 or 1989 or **d88,89**, dummies;
- **fcode**, firm code, 157 firms;
- **sales**, annual sales;
- **employ**, number of employees.

Questions

- We want to know whether receiving a training grant during one year (money to train employees) actually results in more hours of training.
- Second, we want to know whether more hours of training in one year have an effect on sales in the following one. (For this one, you may want to look into the plm package, otherwise write the code yourself).

Stepwise Question 1

1. Using simple conditional means, check whether firms that received the grant had more hours of training. Test this.
2. Can you think about possible omitted variables?
3. Estimate the model using Pooled OLS. Is the effect of the grant positive or negative? Is it significant?
4. Estimate the model using Fixed Effect. Is the effect of grant positive or negative? Is it significant?
5. Do larger firms provide their employees with more or less training, on average?