

BSM946: Topic 5 Panel Data

David Morris

Economics and Strategy Group

d.morris5@aston.ac.uk

Today's Plan

- ① What is Panel Data?
- ② When is panel data useful?
- ③ Pooled OLS and problems with Pooled OLS
- ④ First Difference Models
- ⑤ Fixed Effects Models (FE)
- ⑥ Issues with panel estimators?
- ⑦ Published Example

Will now expand that

- 1 This week we will look at models, building up to fixed effects
- 2 Will also move on to look at some of the issues that exist with panel data

- 1 What is panel data?
- 2 Why is it useful?

- i.e. we have numerous observations but observe them all at the same time, or at only one point in time
- X_i where $i = 1, 2, \dots, n$

- i.e. we have a sample of 10,000 individuals that are in the labour force, and we ask the same individuals the same questions every year
- $X_{i,t}$ where $i = 1, 2, \dots, N$, and $t = 1, 2, \dots, T$

- Panel data or longitudinal data are repeated measurements at different points in time on the same individual unit, such as a person, firm or country
- Regressions can then capture both variation over units, similar to regression on cross-section data, and variation over time
- In this lecture we will focus on methods for short panels, meaning data on many individual units and few time periods
- Examples include panel surveys of many individuals or firms
- We will see that one important advantage of panel data is the ability to control for certain types of omitted variable bias

Why is it useful?

Baltagi (2001) puts, “*Panel data give more informative data, more variability, less collinearity among the variables, more degrees of freedom and more efficiency*” (p.6).

- 1 Provide ways of dealing with heterogeneity
- 2 Can start to investigate how time impacts certain relationships
- 3 Can imply more causality
 - If A precedes B we can be more confident in suggesting A caused B

Why is it useful?

Heterogeneity is the key reasons we are going to explore regarding why panel data is used

What do we mean by heterogeneity?

- Can't we just add more independent variables to control for this?

What about unobserved heterogeneity?

- This is the real reason for using panel data

Unobserved Heterogeneity

- 1 Looking at firm's and trying to measure their productivity...
 - What if there are characteristics that we cannot measure that make a given firm perform better or worse?
- 2 Looking at how people's happiness varies with their commuting behaviour....
 - Some people are happier than others....
 - Can we really control for this?

If we believe we have unobserved heterogeneity the is related to our independent variables OLS estimation will be biased (unobserved variable bias)

Panel data models we will look at today can solve this unobserved heterogeneity problem

- ① First idea is to show how pooled OLS is not appropriate in many situations
- ② Second step is to introduce the First Difference (FD) and Fixed Effects (FE) models
- ③ There is also a model called Random Effects (RE) that we will not look at
 - Can find further reading on this in Dougherty chapter 14 if you are interested

Example: The returns to education

A simple model for returns to wages might be: 1982 and 1987 is:

$$wage_{i,t} = \beta_0 + \delta_0 year_t + \beta_1 education_{i,t} + a_i + u_{i,t}$$

Where i denotes different individuals, a_i is an unobserved individual effect

- It represents all the factors affecting individuals wages that do not change over time

This could include:

- Work ethic
- Ability
- Informal networks
- Other factors that may be roughly constant over a five-year period

How do we estimate the returns to education model?

$$wage_{i,t} = \beta_0 + \delta_0 year_t + \beta_1 education_{i,t} + v_{i,t}$$

where $v_{i,t} = a_i + u_{i,t}$ is often called the **composite error**

As we do not control for a_i it is included in our error term v_{it}

We could pool the different years of data and estimate the model using OLS

- Treat all observations as the same and run an OLS on both time periods at once
 - We have included a year variable to pick up any time trends...

Drawbacks of Pooled OLS

$$wage_{i,t} = \beta_0 + \delta_0 year_t + \beta_1 education_{i,t} + v_{i,t}$$

This method has the important drawback that for OLS to be consistent we need to assume that the errors and the independent variables are not correlated - $E[\epsilon | X]=0$

- i.e. $education_{i,t}$ and v_{it} are uncorrelated
- i.e. or $education_{it}$ and $(u_{it} + a_i)$ are uncorrelated
- Not likely as a_i is likely to impact on the education level of an individual ($education_{it}$)

Even if $education_{i,t}$ is uncorrelated with $u_{i,t}$ pooled OLS is biased and inconsistent if a_i and $education_{i,t}$ are correlated

- This is a form of omitted variable bias (or an endogenous regressor)
 - a_i is not controlled for and is in our error term

Assumption 1: Short Panels

As we have just mentioned we are looking at cases where there are a large number of observations over multiple time periods

What we are saying here is that we have:

- Small T, Large N!

With more time periods some of the problems from time series start to become important

We will start looking at this next week.....

If you want to do more time series work there is an option in final year

- Applied Econometrics and Forecasting

Not the only assumption....

Assumption 2: Serial correlation

One of the assumptions that needs to hold for OLS (Gauss-Markov assumptions) to be BLUE is:

$$\text{Cov}(u_i, u_s) \neq 0 \text{ where } i \neq s$$

What does this mean?

- There should be no correlation between two errors in the population!

If this does not hold then OLS will not be biased, but will be inefficient - each observation is not providing equal amounts of information

There is likely to be correlation between the error terms in two different time periods

Assumption 2

$$\text{Corr}(v_{i1}, v_{i2}) = 0$$

$$\text{Corr}(v_{i1}, v_{i3}) = 0$$

$$\text{Corr}(v_{i2}, v_{i3}) = 0$$

$$\text{Corr}(v_{i1-T}, v_{iT}) = 0$$

Not just a problem caused by time... can be caused by spatial factors

This will be more of an issue with more time periods and is something of real importance in time series analysis

Recap

Hopefully now we know:

- ① What panel data is
- ② What it is useful
- ③ Assumptions that are involved

Let's start looking at the actual models!

Crime rate model

Moving on to look at another example, if we are interested in how crime rates in an area are impacted by gender we may think of modelling it as such:

$$\text{Crime rate}_{i,t} = \beta_0 + \delta_0 \text{year}_t + \beta_1 \text{Percentage of Males}_{i,t} + v_{i,t}$$

- where i denotes different cities, and t denotes different years

But we may have something important missing such poverty levels

- which may be related to our independent variable!
- $E[\varepsilon | X] \neq 0$

Standard pooled OLS results might be biased

Crime rate model - pooled OLS estimates

regress crmrte pctymle i.year

```
. reg crmrte pctymle i.year
```

Source	SS	df	MS	Number of obs = 630		
Model	.012081785	7	.001725969	F(7, 622) = 5.52		
Residual	.19446028	622	.000312637	Prob > F = 0.0000		
Total	.206542066	629	.000328366	R-squared = 0.0585		
				Adj R-squared = 0.0479		
				Root MSE = .01768		

crmrte	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
pctymle	.1707876	.0291916	5.85	0.000	.1134616	.2281137
year						
82	.0001395	.0026361	0.05	0.958	-.0050373	.0053163
83	-.0015609	.0026371	-0.59	0.554	-.0067396	.0036178
84	-.0025824	.0026386	-0.98	0.328	-.0077641	.0025993
85	-.0020223	.002641	-0.77	0.444	-.0072087	.0031641
86	.0008261	.002645	0.31	0.755	-.0043682	.0060203
87	.0023619	.00265	0.89	0.373	-.0028421	.0075659
_cons	.0167974	.0033028	5.09	0.000	.0103114	.0232835

First Differencing



An alternative to pooled OLS - First Differencing

In most applications the reason for collecting panel data is to allow for the unobserved effect, a_i

It turns out that this is simple to allow: because a_i is constant over time, we can therefore difference the data across the two years

For a cross-sectional observation i write the two years as:

$$CR_{i,1} = \beta_0 + \beta_1 \text{Perc Male}_{i,1} + a_i + u_{i,1}, (t = 1)$$

$$CR_{i,2} = \beta_0 + \delta_0 + \beta_1 \text{Perc Male}_{i,2} + a_i + u_{i,2}, (t = 2)$$

An alternative to pooled OLS - First Differencing

If we subtract the second equation from the first we obtain:

$$(CR_{i,2} - CR_{i,1}) = \delta_0 + (a_i - a_i) + \beta_1(\text{Perc Male}_{i,2} - \text{Perc Male}_{i,1}) + (u_{i,2} - u_{i,1})$$

This can instead be written:

$$\Delta CR_i = \delta_0 + \beta_1 \Delta \text{Perc Male}_i + \Delta u_i$$

where Δ is the change in the variable from period 1 to period 2

The unobserved effect a_i does not appear in this equation, it has been “differenced away”

An alternative to pooled OLS - First Differencing

$$\Delta CR_i = \delta_0 + \beta_1 \Delta \text{Perc Male}_i + \Delta u_i$$

This parameters in this equation can be consistently estimated by OLS provided that Δu_i is uncorrelated with $\Delta \text{Perc Male}_i$

- No longer biased

We call the OLS estimator of β_1 the **first-differenced** estimator

Estimating the city crime rate by first differencing

In this model assuming that Δu_i is uncorrelated with Δx_i may be reasonable, but could fail

- For example, if law enforcement effort increases more in cities with more males this will lead to bias in the OLS estimator

Another crucial assumption is that there is some variation in Δx_i across i

- This fails if the explanatory variable does not change over time. This is not a problem in the crime model if the percentage of men in a city varies over time, but can often be an issue (E.g. a model where i denotes an individual and $x_{i,t}$ is a dummy variable for being gender)

Crime rate model - first differenced estimates

xtset county year

gen ccrmte = D.crmte

gen ppctymle = D.pctymle

reg ccrmte ppctymle i.year

. reg ccrmte ppctymle i.year

Source	SS	df	MS	Number of obs = 540		
Model	.001193114	6	.000198852	F(6, 533) = 2.51		
Residual	.042254447	533	.000079277	Prob > F = 0.0211		
Total	.043447561	539	.000080608	R-squared = 0.0275		
				Adj R-squared = 0.0165		
				Root MSE = .0089		

ccrmte	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
ppctymle	.0734199	.4572267	0.16	0.872	-.8247675	.9716073
year						
83	-.0018392	.0013273	-1.39	0.166	-.0044466	.0007682
84	-.001157	.0013274	-0.87	0.384	-.0037646	.0014506
85	.0004143	.0013276	0.31	0.755	-.0021937	.0030223
86	.0026631	.0013445	1.98	0.048	.0000219	.0053044
87	.0013561	.0013406	1.01	0.312	-.0012774	.0039897
_cons	1.83e-06	.0011397	0.00	0.999	-.002237	.0022406

The returns to education revisited

We have seen how first differencing can be used to overcome omitted variable bias when the omitted variable is constant over time

Let's reconsider the problem of estimating the returns to education, now using panel data on individuals for two years:

$$wage_{i,t} = \beta_0 + \delta_0 year_t + \beta_1 educ_{i,t} + a_i + u_{i,t}, t = 1, 2$$

here a_i contains unobserved ability which is probably correlated with $educ_{i,t}$

What happens if we first difference this?

The returns to education revisited

Well the first differenced equation is:

$$\Delta wage_i = \delta_0 + \beta_1 \Delta educ_i + \Delta u_i$$

The problem here is that for most employed individuals education does not change over time so $\Delta educ_i = 0$

This gives....

$$\Delta wage_i = \delta_0 + \Delta u_i$$

Not the best estimate!

Fixed Effects



Fixed effects (FE) estimation

First differencing is only one way to eliminate the unobserved effect, a_i

An alternative method is called the *fixed effects* transformation

Consider the crime rate model we were previously working with:

$$CR_{it} = \beta_0 + \delta_0 year_t + \beta_1 Perc\ Male_{it} + a_i + u_{it}$$

Problem is the a_i term as it ends up in our composite error term:

$$CR_{it} = \beta_0 + \delta_0 year_t + \beta_1 Perc\ Male_{it} + \eta_{it}$$

Where $\eta_{it} = a_i + u_{it}$

And we believe that $Cov(\eta_{it}, X_{it}) \neq 0$

Fixed effects (FE) estimation

How does FE work?

Idea is that we find the average of the variables over time (time average) and take this away from the original equation

$$FE = \text{Normal equation} - \text{Time averaged equation}$$

What is the time average?

$$\overline{CR}_{it} = \frac{1}{T} \sum_{t=T}^T CR_{it}$$

Fixed effects (FE) estimation

Original equation:

$$CR_{it} = \beta_0 + \delta_0 year_t + \beta_1 Perc\ Male_{it} + a_i + u_{it}$$

Time averaged equation:

$$\overline{CR}_i = \beta_0 + \delta_0 year_t + \beta_1 \overline{Perc\ Male_i} + a_i + \overline{u_i}$$

Fixed effects equation:

$$CR_{it} - \overline{CR}_i = (\beta_0 - \beta_0) + \delta_0 + \beta_1 (Perc\ Male_{it} - \overline{Perc\ Male_i}) + (a_i - a_i) + (u_{it} - \overline{u_i})$$

Fixed effects (FE) estimation

As we end up with $a_i - a_i$ the unobserved heterogeneity is difference away

Tend to write the equation as:

$$\widetilde{CR}_{i,t} = \delta_0 + \beta_1 \widetilde{\text{Perc Male}}_{i,t} + \widetilde{u}_{i,t}$$

where $\widetilde{CR}_{i,t} = CR_{i,t} - \overline{CR}_i$ is the time-demeaned data on y, and similarly for $\widetilde{\text{Perc Male}}_{i,t}$ and $\widetilde{u}_{i,t}$

- OLS estimation of this will now be consistent as long as $\text{Cov}(\text{Perc Male}_{it}, u_{it}) = 0$

Crime rate model - fixed effects estimates

xtset county year

xtreg crmrte pctymle i.year, fe

corr(u_i, Xb) = -0.3867 F(7, 533) = 4.24
 Prob > F = 0.0001

crmrte	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
pctymle	-.1438467	.1823299	-0.79	0.431	-.50202	.2143267
year						
82	-.0003054	.0010591	-0.29	0.773	-.002386	.0017752
83	-.0024482	.0011488	-2.13	0.034	-.0047049	-.0001915
84	-.0039017	.0012805	-3.05	0.002	-.0064172	-.0013861
85	-.0038066	.0014576	-2.61	0.009	-.0066699	-.0009434
86	-.0015509	.0017183	-0.90	0.367	-.0049264	.0018246
87	-.0005897	.0019952	-0.30	0.768	-.0045093	.0033298
_cons	.0461865	.0170464	2.71	0.007	.0127001	.079673
sigma_u	.0181605					
sigma_e	.00689124					
rho	.87413185	(fraction of variance due to u_i)				

F test that all u_i=0: F(89, 533) = 40.02 Prob > F = 0.0000

Disadvantages of FE

Removing the time averaged value allows us to remove anything that does not vary with time

Such as:

- a_i

Also removes anything else that does not vary with time

Such as:

- If the house is close to the coast and thus a tourist area

First Differences vs Fixed Effects

With two time periods:

$$FE \equiv FD$$

With ≥ 3 periods:

$$FE \neq FD$$

Which is better?

- **Not simple!**
- FE deals with lack of strict exogeneity better
 - i.e. $\text{Cov}(u_{it}, X_{is}) \neq 0$

FE Summary

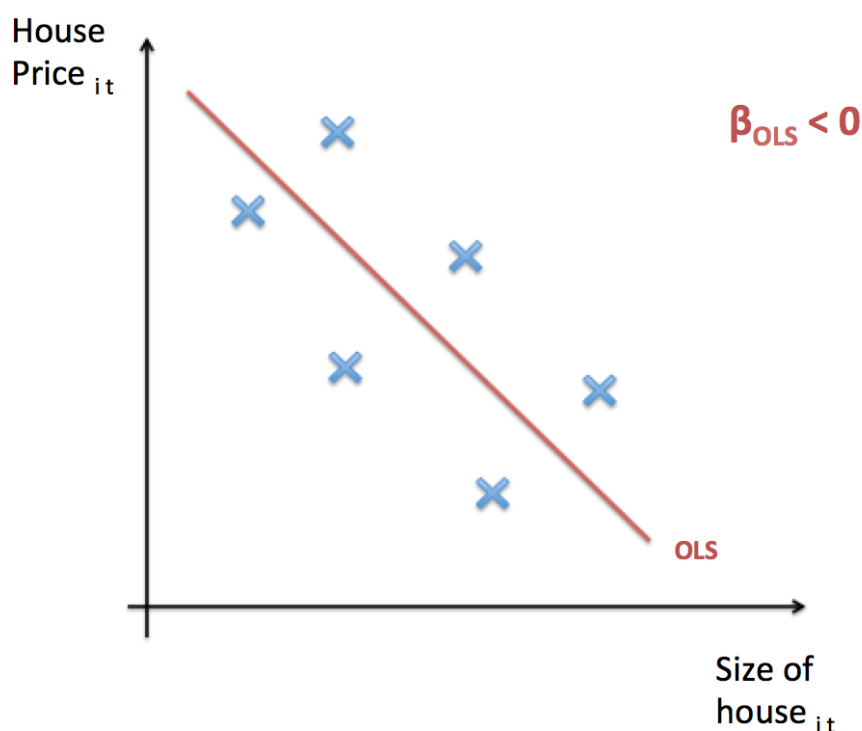
The fixed effects estimator allows us to remove unobserved heterogeneity (a_i), just like first differencing

Because of this transformation any explanatory variable that is constant over time for all i is swept away by the fixed effects method

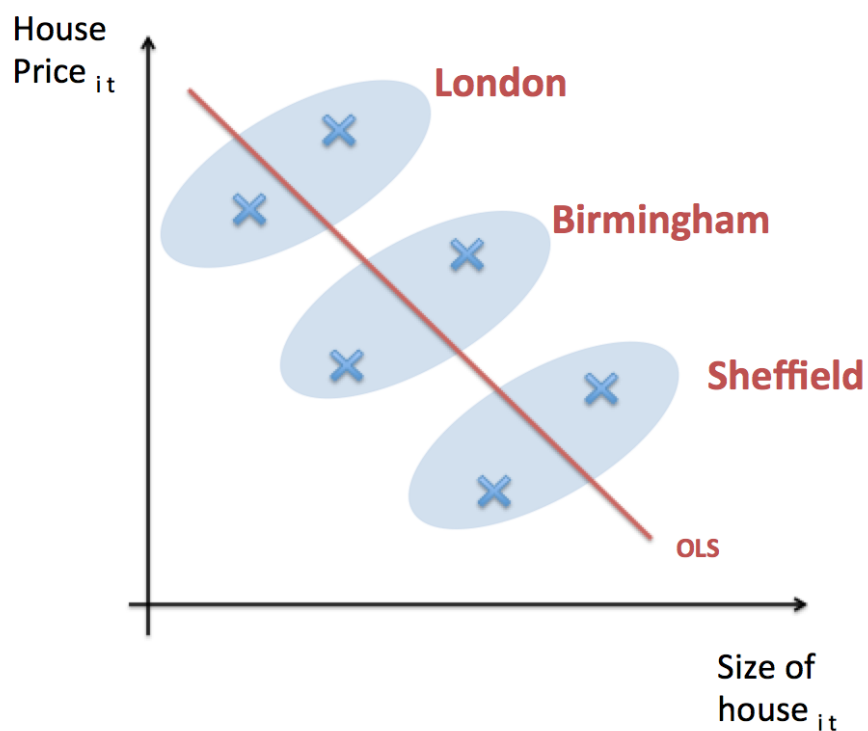
- Therefore we cannot include variables such as gender or a city's distance to a river etc

In applications where the effect of a **time-invariant variable** on the dependent variable is the main topic of interest the fixed effects estimator is **not very useful!**

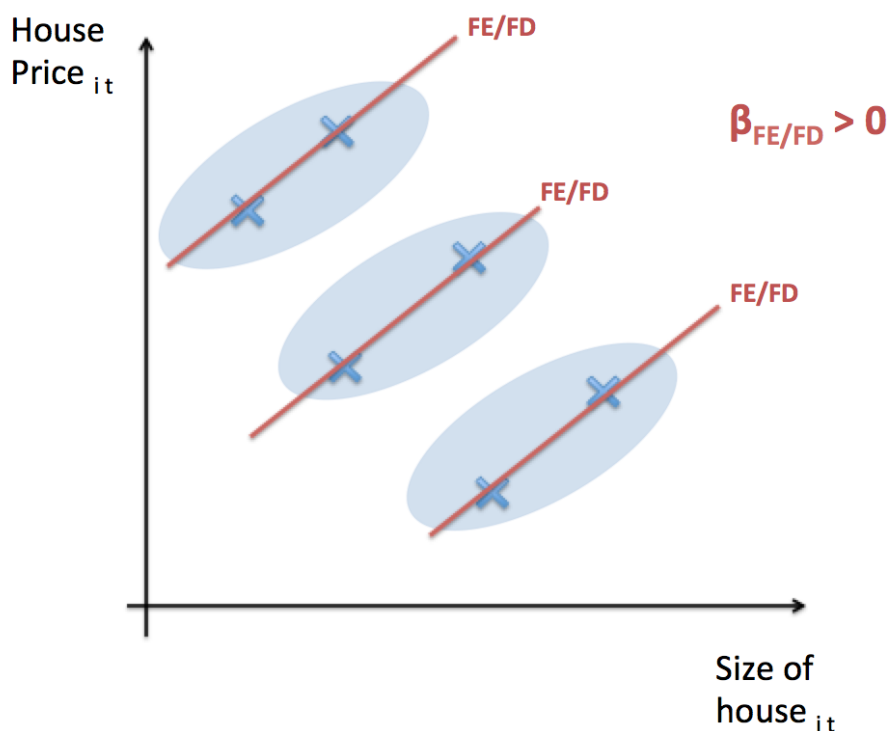
How important is it to remove the unobserved heterogeneity?



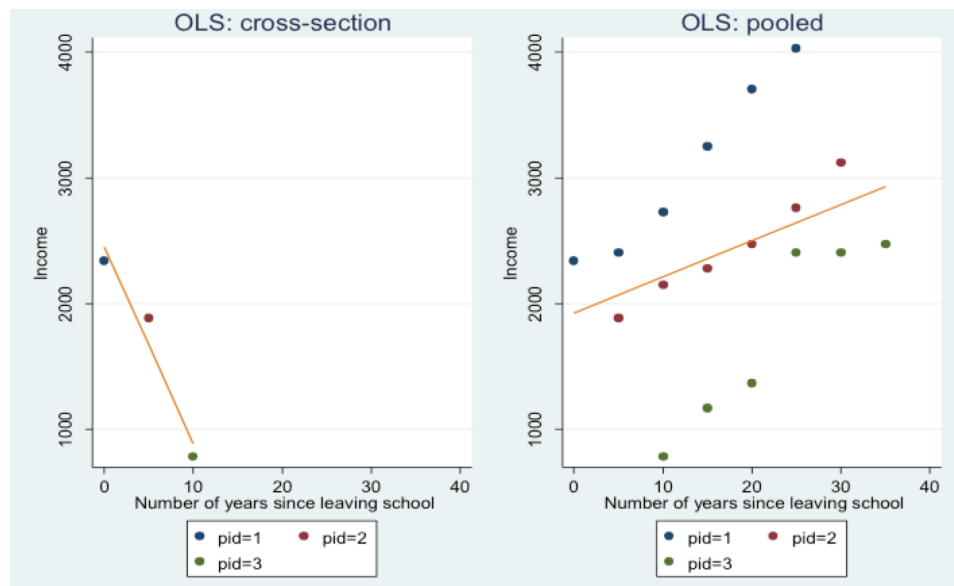
How important is it to remove the unobserved heterogeneity?



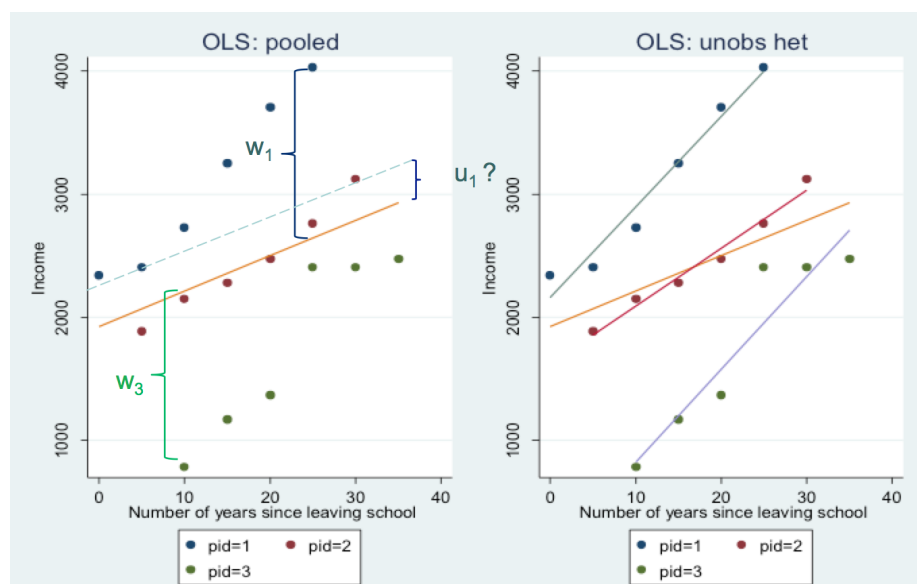
How important is it to remove the unobserved heterogeneity?



How important is it to remove the unobserved heterogeneity? STATA Example



How important is it to remove the unobserved heterogeneity? STATA Example





Problems with Panel Data

There are some negatives to panel data though.....

- ① Non-random attrition
- ② Available information
- ③ Panel Conditioning

Issue 1: Non-random attrition

Panels require the same individual or firm to be interviewed in each time period

Sometimes individuals drop out and stop responding

- This is what we call attrition

Can be an issue if those that drop out are not random...

- i.e. if individuals with certain characteristics are more likely to drop out

Less of an issue in cross-sectional data as we don't require people to complete more than one survey/interview

Issue 2: Availability

- In many countries there is a lack of genuine panel data where specific individuals or firms are followed over time
- Lots of repeated cross-sectional surveys may be available where a random sample is taken from the population at consecutive points in time
not the same individuals each time however

While panels may allow us to conduct robust analysis we need the data!

Issue 3: Panel Conditioning

Genuine panel dataset can suffer from conditioning effects

- People adapt to being asked questions again and again
- Leads them to change their behaviour

Recent work around panel data has found increasing evidence that people do adapt their behaviour around the subject of the questionnaire

Issue 3: Panel Conditioning

- Zwane (2011) split their survey sample in two, with half randomly allocated to the health section of the questionnaire and half allocated to the household finances section
 - The authors find that the conditioning effects change peoples behaviour to a large enough extent to change the mean of the outcome variable as well as the estimated coefficients from the regression analysis
 - For example, being in the health half of the survey increased the take-up of medical insurance and the use of water treatment products

Issue 3: Panel Conditioning

- Similar findings have been found by Crossley (2014) with saving behaviour
- Das et al (2011) show a panel conditioning effect by comparing refreshment samples to those that have been in the panel for longer
 - also find conditioning effects

Does seem to be a real issue for panel data!

Panel Data

So panel data is useful as it allows us to remove troublesome heterogeneity, α , which cannot be controlled for with cross sectional work

- But has its down sides....

Cross Sectional Data

Cross sectional data suffer much less from typical panel data problems like attrition and nonresponse, and are very often substantially larger, both in number of individuals or households and in the time period that they span

- Cannot control for α

Is there anyway to get the best of both models?

Pseudo Panels

In a seminal paper, Deaton (1985) suggests the use of cohorts to estimate a fixed effects model from repeated cross-sections

In his approach, individuals sharing some common characteristics (most notably year of birth) are grouped into cohorts, after which the averages within these cohorts are treated as observations in a pseudo panel

More work followed suit such as: Moffitt (1993) and Collado (1997)

- *They do of course have their own downsides...*
- **Not going to look at them in any detail in this course!**

Learning Outcomes

- ① **Be able to review the limitations of simple liner regression analysis and the situations where it is not the most appropriate method to use;**
 - Why might we want to use panel models?
- ② **Have developed an awareness of which models are appropriate to use when an OLS regression may not be robust, and be able to apply these models correctly;**
 - How does First Differences (FD) work?
 - How does Fixed Effects (FE) work?
- ③ **Command an experienced understanding of the empirical approach to economics, and be able to undertake a robust piece of econometric analysis;**
 - Chance in the labs to apply this....
 - What problems might we face with panel data?
 - Can you explain the assumptions you need for robust panel analysis?

Questions

Your questions please....

Glossary

Panel Data - Panel data can be thought of as cross sectional data, that is observed over numerous time periods. The same individuals answer the same questions year upon year.

Endogeneity - Occurs when an explanatory variable is correlated with the error term. Normally due to an unobserved variable impacting both the dependent and an independent variable or due to the being a reverse causality link between the dependent and independent variables

Unobserved Heterogeneity - Unobserved variation in our sample, i.e. factors that we cannot control for but that might have an impact on our dependent variable.

First differenced - $y_t - y_{t-1}$, the difference between this periods value and last periods value

Time demeaned - The FE model is a time demeaned model, i.e. the mean across all time periods is removed from the value.