Silvia Mendolia

Labour market, health and well-being: economic analysis using panel data

Plan for this lecture

- Using Stata to construct a Panel dataset
- Types of questions, types of variables: time-invariant, time-varying and trend
- Between- and within-individual variation
- Concept of individual heterogeneity
- Linear Panel regression
- Problems with Pooled OLS and possible solutions with panel data

Using Stata to construct Panel data

- Remember the LONG structure of the data
- Use any cross section to select the variables you need
- Use append to append all waves one after the other
- sort pidp wave
- Browse the data and check how it looks!

Using Stata to construct Panel data

- Useful files in Understanding Society:
  - 1. indresp (individual information)
  - 2. hhresp (household information)
- Waves are indicated with letters (a is wave 1, etc)
- Variables are generally consistent across waves but there are some differences. Check the documentation and the questionnaires!
- You can get information from both and merge

Using Stata to construct Panel data

1. Open a do-file (so you can remember what you did!)
2. Set your working directory

cd “……”

3. Run a loop to select your variable in each wave

```
foreach j in  a b c d e f g h i j{
use  `j'_indresp, clear
renpfix `j'_      // Strip off wave prefix
 // Now select variables needed
keep pidp  hid  dvage  ……/*
save indiv`j', replace
}
```

Using Stata to construct Panel data

1. Gen a wave indicator in each individual file (eg in indiva gen wave=1)
2. Use the individual files to create a panel by appending one wave after the other
3. use indiva
4. append using indivb indivc indivd indive indivf indivg indivh indivi indivj
5. sort pidp wave
6. br pidp wave marstat sex…….

Types of variable

- Some variables vary between individuals but hardly ever over time:
  - Sex
  - Ethnicity
  - Parents' social class when you were 14
  - The type of primary school you attended (once you've become an adult)
- Others vary over time, but not between individuals:
  - The retail price index
  - National unemployment rates
  - Age, in a cohort study

Types of variable

- Other variables vary both over time and between individuals
  - Income
  - Health
  - Psychological wellbeing
  - Number of children you have
  - Marital status
- Trend variables
  - Vary between individuals and over time, but in highly predictable ways:
  - Age
  - Year

Between- and within-individual variation

- If you have a sample with repeated observations on the same individuals, there are two sources of variance within the sample:
  - The fact that individuals are systematically different from one another **(between-individual variation)**
  - The fact that individuals' behaviour varies between observations over time **(within-individual variation)**

Total variation is the sum over all individuals and years, of the square of the difference between each observation of x and the mean

Within variation is the sum of the squares of each individual's observation from his or her mean

Between variation is the sum of squares of differences between individual means and the whole-sample mean

Between- and within-individual variation

Use the dataset “longitudinal_td” from week 3

```
. xtsum sex age urban_dv scghq2_dv
```

Variable		Mean	Std. Dev.	Min	Max	Observations
sex_dv	overall	1.577576	.4939465	1	2	N = 192146
	between		.49535	1	2	n = 36618
	within		0	1.577576	1.577576	T-bar = 5.24731
age_dv	overall	52.40156	17.05193	16	104	N = 192152
	between		18.32819	16	102.5	n = 36619
	within		2.271404	47.51267	57.29045	T-bar = 5.24733
urban_dv	overall	1.23784	.4257617	1	2	N = 192083
	between		.4041759	1	2	n = 36619
	within		.0966778	.348951	2.126729	T-bar = 5.24545
scghq2~v	overall	1.708655	2.943055	0	12	N = 174113
	between		2.419806	0	12	n = 33982
	within		1.998091	-8.791345	12.37532	T-bar = 5.12368

All variation is “between”

Most variation is “between”, because it’s fairly rare to switch between urban and rural area

Between- and within-individual variation

```
. xtsum sex age urban_dv scghq2_dv

Variable                |      Mean   Std. Dev.       Min        Max |    Observations
------------------------+--------------------------------------------+----------------
sex_dv          overall |  1.577576   .4939465         1          2 |     N =  192146
                between |             .49535           1          2 |     n =   36618
                within  |                 0    1.577576   1.577576 |  T-bar = 5.24731
------------------------+--------------------------------------------+----------------
age_dv          overall |  52.40156   17.05193        16        104 |     N =  192152
                between |             18.32819        16      102.5 |     n =   36619
                within  |             2.271404    47.51267   57.29045 |  T-bar = 5.24733
------------------------+--------------------------------------------+----------------
urban_dv        overall |   1.23784   .4257617         1          2 |     N =  192083
                between |             .4041759         1          2 |     n =   36619
                within  |             .0966778     .348951   2.126729 |  T-bar = 5.24545
------------------------+--------------------------------------------+----------------
scghq2~v        overall |  1.708655   2.943055         0         12 |     N =  174113
                between |             2.419806         0         12 |     n =   33982
                within  |             1.998091   -8.791345   12.37532 |  T-bar = 5.12368
```

Observations with non-missing variable

Number of individuals

Average number of time-points

Min & max refer to individual deviation from own averages, with global averages added back in.

xttab

```
. xttab jbstat
```

jbstat	Overall Freq.	Overall Percent	Between Freq.	Between Percent	Within Percent
self emp	14648	7.62	4388	11.98	60.42
Paid emp	87500	45.54	19819	54.13	81.25
unemploy	8105	4.22	4633	12.65	44.64
retired	57483	29.92	11643	31.80	85.27
on mater	1085	0.56	939	2.56	24.02
Family c	11398	5.93	4141	11.31	58.66
full-tim	3902	2.03	2297	6.27	67.73
LT sick	6892	3.59	2382	6.51	58.92
Govt tra	108	0.06	99	0.27	35.32
Unpaid,	126	0.07	98	0.27	25.49
On appre	21	0.01	19	0.05	13.23
doing so	850	0.44	661	1.81	28.72
Total	192118	100.00	51119	139.61	71.63
			(n = 36616)		

Pooled sample, broken down by person/years

Number of people who spent *any* time in this state

Of those who spent any time in this state, the proportion of their time (on average) they spent in it.

Which statistical model?

First consider your research question:

Is your research question more suitable for cross-sectional or longitudinal data?

- Cross-sectional: exploit variation between individuals
- Longitudinal: exploit variation "within" individuals over time and permit causal interpretation of effects
  - and can consider "between" variation if needed

Example: How does mental health change after unemployment?

- What is the difference in mental health between individuals who are employed and unemployed?
- What is the difference in mental health before and after a job loss?
  - What is the difference in mental health between men and women and before and after job loss?

Unobserved heterogeneity

- Remember: **people are different!**
- **In economics, when we talk about heterogeneity, we are really talking about unobservable (or unobserved) heterogeneity**:

- **Observed heterogeneity**: differences in education levels, or parental background, or anything else that we can measure and control for in regressions
- **Unobserved heterogeneity**: anything which is fundamentally unmeasurable, or which is rather poorly measured, or which does not happen to be measured in the particular data set we are using.

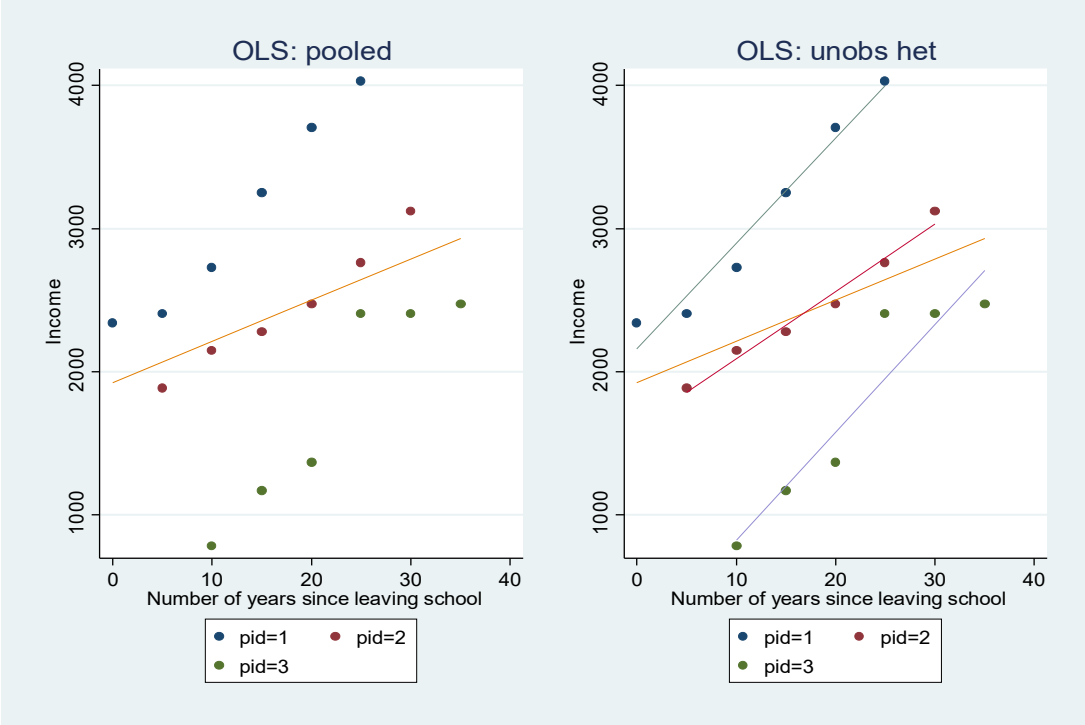
- **The intuition**: With panel data we can do something about unobserved heterogeneity as we can differentiate between person-level unobserved x that are identical over time and those that vary over time!

OLS with panel data

- Cross sectional results with one wave can be misleading
- By pooling multiple waves we add more data for the same individuals at different points in time and we can get more precise estimates
- But OLS assumptions can be violated and estimates can be biased!

OLS with panel data

- Error terms for individual 1, 2 and 3 differ systematically
- The association between x and y appears to be biased
- Panel data allows us to break down the error term (w_i) in two components: the unobservable characteristics of the person (u_i), and genuine "error" (e_i)



OLS with panel data

$$Y_{it} = \alpha + x_{it}\beta + a_i + \varepsilon_{it}$$

- a_i captures all unobserved, time-invariant factors that affect Y_{it}
- a_i is specific to an individual and does not vary over time
- Examples?

OLS with panel data

$$Y_{it}=\alpha + x_{it}\beta + a_i + \varepsilon_{it}$$

- How can we estimate β if we have a panel data?
- We can pool all waves and use pooled OLS (normal **reg** command in Stata)
- Rewrite:

$$Y_{it}=\alpha + x_{it}\beta + v_{it}$$

$$v_{it}= a_i + \varepsilon_{it}$$

Problem with OLS: Serially Correlated Errors

- If a variable is overpredicted for an individual in a certain year, it is likely to be overpredicted in the following years
- **An individual with higher ability today will have higher ability tomorrow as well**
- Cor (v_{1t}, v_{is}) ≠0 for t≠ s

- Each additional observation for a given person provides less than an independent piece of new information.
- With serially correlated errors, standard errors are biased.

Problem with OLS: Serially Correlated Errors

- One possible solution is to use Clustered Robust standard errors
- This allows correlation within clusters
- Use vce (**cluster id**) option in Stata
- Id is the individual indicator variable

Problem with OLS: Omitted variable bias

- In order to consistently estimate β, we must assume that v_{it} is uncorrelated with x_{it}
- If ai is correlated with x_{it}, pooled OLS will be biased and inconsistent
- This is called **heterogeneity bias**
- It results from **omitting a time-constant variables**

Problem with OLS: Omitted variable bias

- Why would a_i be correlated with x_{it}?
- Imagine we are estimating the impact of job loss on mental health
- Unobserved factors that affect mental health may also have affected job loss (personality traits, attitudes, ability)
- Individuals who have certain traits (eg low levels of work ethics) may be more likely to be in poor mental health and loose their job at the same time

- X_{it} (Job loss) may be positively or negatively correlated with a certain personality trait

Between and within variation

- According to the counterfactual approach to causality an individual causal effect is defined as:

$$\Delta_i = Y^T_{i,t_0} - Y^C_{i,t_0}, \quad T: \text{treatment}, C: \text{control}$$

- However, this is not estimable (fundamental problem of causal inference)
- Estimat $\widehat{\Delta_i} = Y^T_{i,t_0} - Y^C_{j,t_0}$:a (assuming no unobserved heteroc
- Estimation $\widehat{\Delta_i} = Y^T_{i,t_1} - Y^C_{i,t_0}$
- **We compare the same person at t_1 and t_0,** assuming no time effects (within estimation)
- Estimation with panel data (adding a control group) and

$$\widehat{\Delta_i} = \left(Y^T_{i,t_1} - Y^C_{i,t_0}\right) - \left(Y^C_{j,t_1} - Y^C_{j,t_0}\right)$$

Between and within variation

- Between estimation can be used with experimental data (and randomization of individuals in treatment)

- However, it’s not feasible with non-experimental data because of:
  - Self-selection into treatment
  - Unobserved heterogeneity

- We can’t randomly assign people to unemployment to study the effect on their mental health!

Data exercise

- Use the “longitudinal_td” dataset

- Run a basic regression (OLS, no panel) in Stata to understand the relationship between income and life satisfaction
- Which control variables? Why?
- How do you insert binary variables in a regression (e.g. sex)?
- What about categorical variables (e.g. educational qualifications)?