Lecture 9 - Panel Data

Samuel DeWitt

Panel Data - Organization

1) What is panel data?
2) Why is panel data valuable?
3) How do we analyze panel data?

Panel Data - What is it?

Panel data consist of repeated observations of the same units over at least two time periods.

The NLSY is an example of a panel data set - youth were first interviewed in 1997 and then annually thereafter.

Panel Data - What is it?

Mathematical notation changes slightly in the context of panel data when we have multiple observations of the same measure at different time periods.

We generally notate individuals with the subscript i and time with the subscript t:

Outcomes: Y_{it}

Independent Variables: X_{it}

Error Terms: ϵ_{it}

Panel Data - What is it?

We can also examine treatment effects within-units over time because treatment can turn *on* and *off* at the different observation periods.

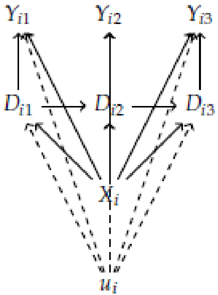
Treatment is generally notated as such:

$$D_{it}$$

One complication is that a unit treated at time t generally tends to also be treated at time t+*1* (or, at least, these statuses are correlated).

Panel Data - What is it?

More generally, the depicted DAG describes the typical structure assumed for panel data:



Panel Data - What is it?

In cross-sectional data each row corresponds to one individual unit, but in panel data one unit can have multiple rows, one for each time period captured in the data.

Here's what traditional data look like:

Unit ID	Age	Crimes Reported
100	19	5
101	17	2
102	18	3

Here's what panel data look like:

Unit ID	Age	Crimes Reported
100	19	5
100	20	4
100	21	3
101	17	2
101	18	4
101	19	3

Panel Data - Why is it valuable?

Panel data solves a prominent issue with strictly cross-sectional data - we *can* know what treated units look like **before** and **after** their change in treatment status.

An additional benefit is that we can effectively ignore one particular source of bias due to unobserved heterogeneity across units (u_i)

Panel Data - Why is it valuable?

With panel data, we can ignore sources of time-stable unobserved heterogeneity across units.

In real people language, time-stable unobserved heterogeneity means that there are characteristics about units that we do not measure or observe that do not change over time *within* a unit, but can influence the outcome variable we are interested in.

Panel Data - Why is it valuable?

Why do you think that unobserved heterogeneity (u_i) is not subscripted by t?

Panel Data - Why is it valuable?

Why do you think that unobserved heterogeneity (u_i) is not subscripted by t?

If you said that panel data only can account for time-stable sources of unobserved heterogeneity, you are correct!

Panel Data - Why is it valuable?

Basically, if the time-stable characteristic influences a person's behavior, it influences it both **before** and **after** some treatment takes affect.

This is why panel data allows us to effectively ignore one important source of bias - the influence of time-invariant unit characteristics that we both **can** and **cannot** observe (since in both cases they have the same values before and after treatment occurs).

Panel Data - How do we analyse it?

We have a few options for analysis methods when we have panel data, and we will primarily talk about just one of these methods in this class.

1) Pooled OLS
2) Fixed Effects
3) Random Effects

Panel Data - How do we analyse it?

For the purposes of this class, we will focus on pooled OLS and fixed effects analysis, but will probably only have time to really dive into the former this semester.