Panel Data

From the perspective of an applied micro economist

Jesse Burkhardt (Assistant Professor in DARE)

Slides partially taken from Stephen Koontz

What is panel data?

What is panel data?

Generally, a mixture of cross-sectional and time series data

$y_{it} = \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \ldots + \beta_k x_{kit} + e_{it}$

where $i = 1, \ldots, N$ and $t = 1, \ldots, T$. Sample size is $N \times T$. This is a Balanced Design.

Example of an Unbalanced Design: $i = 1, \ldots, N_t$ and $t = 1, \ldots, T_i$. Each i has a different number of T. Or...?

What do the data matrices look like?

$$y = \begin{bmatrix} y_{11} \\ . \\ y_{1T} \\ y_{21} \\ . \\ y_{2T} \\ . \\ . \\ y_{N1} \\ . \\ y_{NT} \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_{1\,11} & \ldots \\ . & . & \\ 1 & x_{1\,1T} & \\ 1 & x_{1\,21} & \\ . & . & \\ 1 & x_{1\,2T} & \\ . & . & \\ . & . & \\ 1 & x_{1\,N1} & \\ . & . & \\ 1 & x_{1\,NT} & \end{bmatrix}$$

What are some examples of panel data?

Balanced vs. unbalanced panels

- What are they?
- When is an unbalanced panel a problem?

Why might we prefer panel data?

- We can exploit variation within an individual (i) over time
- We can exploit variation within time periods across individuals

- But why might this help us as econometricians?

We are interested in…

- Establishing an argument for causality: x causes a β change in Y

$$Y_{it} = \alpha + \beta_1 x_{it} + \epsilon_{it}$$

- What are the key threats to this argument?
    - This is called identification

Bias and omitted variables

- What is omitted from this equation that could lead to biased estimates of β_1?

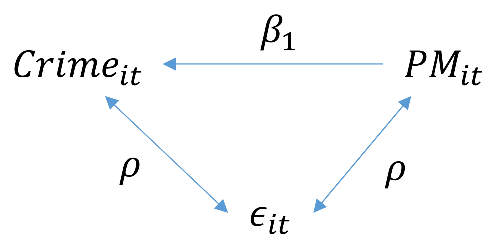
$$Y_{it} = \alpha + \beta_1 x_{it} + \epsilon_{it}$$

Example: Pollution and Crime

$$Crime_{it} = \alpha + \beta_1 PM_{it} + \epsilon_{it}$$

- $Crime_{it}$= crime in county i during time t
- PM_{it}=particulate matter in county i during time t

- What is omitted from this equation that could lead to biased estimates of β_1?



Omitted unobservables

- Let's consider two categories of unobservables
  - Things that are county constant but vary over time
  - Things that are time constant but vary across counties

- How might we control for these unobservables?
- Hint: how do we control for gender?

Two options (for today)

- Fixed effects
- Random effects

Fixed-Effects Models

Suppose we want each ith individual to have its own mean...

$$y_{it} = \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \ldots + \beta_k x_{kit} + \sum \alpha_i D_i + e_{it} \qquad e_{it} \sim N(0, \sigma^2)$$

where D_i = 1 for observation on ith individual, and 0 otherwise.

Suppose we want each tth time period to have its own mean...

$$y_{it} = \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \ldots + \beta_k x_{kit} + \sum \theta_t D_t + e_{it} \qquad e_{it} \sim N(0, \sigma^2)$$

where D_t = 1 for observation on tth period, and 0 otherwise.

Example: Pollution and Crime

$$Crime_{it} = \beta_1 PM_{it} + \sum_i \alpha_i D_i + \sum_t \gamma_t D_t + \epsilon_{it}$$

- $Crime_{it}$= crime in county i during time t
- PM_{it}=particulate matter in county i during time t
- $\alpha_i D_i$= fixed effects for each county
- $\gamma_t D_t$= fixed effects for each time period

- What do these fixed effects control for?
- Are there still omitted variables that could lead to biased estimates of β_1?

Quick Aside: how are fixed implemented?

1. Dummy variables

$$Crime_{it} = \beta_1 PM_{it} + \sum_i \alpha_i D_i + \epsilon_{it}$$

2. Demean by i:

$$(Crime_{it} - \overline{Crime_i}) = \beta_1 (PM_{it} - \overline{PM_i}) + (\epsilon_{it} - \overline{\epsilon_i})$$

3. Equivalent to first differences with 2 time periods

Can throw in time period dummies in either model.
Why are these equivalent?

Fixed Effects Assumptions

For the model $Y_{it} = \beta_1 x_{it1} + \cdots + \beta_k x_{itk} + a_i + e_{it}, t = 1, \ldots, T$

1) β_k are the parameters to estimate and a_i is the unobserved effect
2) We have a random sample from the cross sections (unbalanced?)
3) Each x changes over time. **Why?** And no perfect multicollinearity
4) For each t, the expected value of the idiosyncratic error given the explanatory variables in *all* time periods and the unobserved effect is zero: $E(e_{it}|x_{ik}, a_i) = 0$
   - This is the strict exogeneity assumption

- Under these assumptions, the FE estimator is unbiased

Fixed Effects Assumptions Cont.

For the model $Y_{it} = \beta_1 x_{it1} + \cdots + \beta_k x_{itk} + a_i + e_{it}, t = 1, \ldots, T$

5) $var(e_{it}|x_i, a_i) = var(e_{it}) = \sigma_e^2$, for all t=1,…,T.

This can be addressed with heteroskedasticity robust standard errors

6) For all $t \neq s$, the idiosyncratic errors are uncorrelated: $cov(e_{it}, e_{is}|x_i, a_i) = 0$

Implies…

Benefits of FE

- Makes no assumptions about the correlation between a_i and x_i

Drawbacks of FE

- Suppose we have the model:

$$Crime_{it} = \beta_1 PM_{it} + \sum_i \alpha_i D_i + \sum_t \gamma_t D_t + \epsilon_{it}$$

We cannot include variables that are constant within counties and we cannot include variables that are constant within a year.

* Examples: whether or not a county is urban, geographic region of the US, national policies that do not vary over time.

Random Effects

- What if we want to estimate parameters of variables that are constant within counties, but still control for county specific unobservables?
- Random effects allow us to do this, with an additional assumption.

RE assumptions

Given the model

$$Y_{it} = \beta_1 x_{it1} + \cdots + \beta_k x_{itk} + a_i + e_{it}$$

- Fixed effects allows for correlation between a_i and x's.
- But what if we think a_i and the x's are uncorrelated in all time periods?
  - Example of when this might be the case?
- Thus, the RE assumptions are the same as the fixed effects assumptions with the additional assumption:
  - a_i is independent of all explanatory variables in all time periods: $cov(x_{itj}, a_i) = 0$

How is RE implemented?

$$Y_{it} = \beta_0 + \beta_1 x_{it1} + \cdots + \beta_k x_{itk} + a_i + e_{it}$$

Combined error:

$$Y_{it} = \beta_0 + \beta_1 x_{it1} + \cdots + \beta_k x_{itk} + v_{it}$$

where

$$v_{it} = a_i + e_{it}$$

Because a_i is contained in v_{it}, the composite errors are serially correlated, described by

$$corr(v_{it}, v_{is}) = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}, t \neq s$$

Where $\sigma_a^2 = var(a_i)$ and $\sigma_e^2 = var(e_{it})$

To address this use weighted LS with weights defined as follows

$$\lambda = 1 - \left[\frac{\sigma_a^2}{T\sigma_a^2 + \sigma_e^2}\right]^{1/2}$$

Which is between 0 and 1 (this is important)

The transformed RE equation is

$$Y_{it} - \lambda\overline{Y_i} = \beta_0(1-\lambda) + \beta_1(x_{it1} - \lambda\overline{x_{i1}}) + \cdots + \beta_k(x_{k1} - \lambda\overline{x_{k1}}) + (v_{it} - \lambda\overline{v}_i)$$

- The FE estimator subtracts the time averages
- The RE estimator subtracts a fraction of the time averages
- This also solves the serial correlation in v
- Sample analogs are computed from OLS estimates of v
- Pooled OLS is obtained when $\lambda = 0$
- The RE estimator tends towards the FE estimator as λ goes to 1

Drawback of RE

- Need to assume a_i are uncorrelated with x_i in all time periods which is unlikely.

Benefit of RE

- Plausibly controls for time constant individual specific unobservables while allowing for the recovery of parameters on time constant individual specific covariates.

Example

$$Violent\ Crime_{it} = \beta_1 PM_{it} + D_i + D_t + \epsilon_{it}$$

- $Violent\ Crime_{it}$ count in county i on day t
- PM_{it} is a measure of air pollution in county i on day t
- D_i is a location fixed effect or random effect
- D_t is a time fixed effect

Table 8: Violent Crimes RE and FE

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$PM_{2.5}$	0.123***	0.112***	0.059***	0.060***	0.012***	0.013***	0.012***	0.012***
	(0.008)	(0.008)	(0.008)	(0.008)	(0.002)	(0.002)	(0.002)	(0.002)
year FE		Y	Y	Y	Y	Y	Y	Y
state			FE	RE				
county					FE	RE	FE	RE
month FE							Y	Y
N	77489	77489	77489	77489	77489	77489	77489	77489

Example 2: What explains wages?

$$Wage_{it} = \beta_0 + \beta_1 educ_i + \beta_2 black_i + \beta_3 hispan_i + \beta_4 exper_i$$
$$+\beta_5 exper_{it}^2 + \beta_6 married_{it} + \beta_7 union_{it} + \phi + e_{it}$$

Which variables will drop out with individual FE?

- Time constant parameters are similar for OLS and RE
- Marriage and union premiums fall from OLS to RE. Why?
- Eliminate the household unobservable entirely using FE, the parameters fall even more (why?)
- Captures the idea that people that are more able (higher a_i) are more likely to be married and more likely to have higher wages.
- In OLS, a large part of marriage coefficient is due to the fact that most people who are married would earn more even if they weren't married.

TABLE 14.2

Three Different Estimators of a Wage Equation

Dependent Variable: log(*wage*)			
Independent Variables	**Pooled OLS**	**Random Effects**	**Fixed Effects**
educ	.091 (.005)	.092 (.011)	——
black	−.139 (.024)	−.139 (.048)	——
hispan	.016 (.021)	.022 (.043)	——
exper	.067 (.014)	.106 (.015)	——
exper^2	−.0024 (.0008)	−.0047 (.0007)	−.0052 (.0007)
married	.108 (.016)	.064 (.017)	.047 (.018)
union	.182 (.017)	.106 (.018)	.080 (.019)

Final thoughts

- There is a test called the Hausman test for Fixed versus Random Effects
- Null hypothesis is that the effects are uncorrelated with the data (x's), or random effects are acceptable.
- Most often will reject in favor of FE and that's why you see FE used in most economics studies.