EEP/IAS 118 - Introductory Applied Econometrics, Section 11a

Pierre Biscaye and Jed Silver

November 2021

Agenda

1. Data types
2. First Differences
3. Fixed Effects

Terminology: We will refer to **units** as the individual people, cities, firms, etc. in a given dataset.

Data Types: Cross Section

A cross section is a snapshot of (randomly selected) units at one point in time. This is like the data we have used most often is the past.

Notation: we use i to index units (in this case, individuals):

$$wage_i = \beta_0 + \beta_1 edu_i + \beta_2 exper_i + \beta_3 female_i + u_i$$

indiv	wage	edu	exper	female
1	3.10	11	2	1
2	3.24	12	22	1
.	.	.	.	.
100	5.30	12	7	0

Data Types: Pooled Cross Section

We also call this "repeated cross-section". This is multiple snapshots of multiple bunches of (randomly selected) nuits at many points in time. *Pooling* the data means to treat the separate samples over time as one big sample.

Notation: We still only use i to index observations (in this case homes)

$$hprice_i = \beta_0 + \beta_1 bdrms_i + \beta_2 bthrms_i + \beta_3 sqrft_i + \delta y2010_i + u_i$$

Note: With repeated cross-sections, we can now control for the fact that observations are from different years. In this case we do this by using the $y2010_i$ dummy

Data Types: Pooled cross section

Example:

house	year	hprice	bdrms	bthrms	sqrft
1	2000	85,500	3	2.0	1600
2	2000	67,300	3	2.5	1400
.	.	.	.	.	.
100	2000	134,000	4	2.5	2000
101	2010	243,000	4	3.0	2600
102	2010	65,000	2	1.0	1250
.	.	.	.	.	.
200	2010	144,000	3	2	2000

Data Types: Panel

Panel data tracks the *same* units over time. It is like a repeated cross section, but where every time period we observe the same units rather than a new sample each time.

Notation: With panel data we start indexing observations by t as well as i to distinguish between our observations of unit i (in this case cities) at various points in time t (in this case years):

$$crimes_{it} = \beta_0 + \beta_1 pop_{it} + \beta_2 unemp_{it} + \beta_3 police_{it} + a_i + d_t + u_{it}$$

i	t	murder rate	pop density	police
1	2000	9.3	2.24	440
1	2001	11.6	2.38	471
2	2000	7.6	1.61	75
2	2001	10.3	1.73	75
.	.	.	.	.
100	2000	11.1	3.12	520
100	2001	17.2	3.34	493

Two-Period Panel Data

Let's consider an example:

- data on crime and unemployment rates for 46 cities for 1982 and 1987.
- two time periods, $t = 1$, and $t = 2$.

Let's use just the 1987 cross section and run a simple regression of crime on unemployment:

$$\widehat{crmrte} = 128.38 - 4.16 unemp$$

- Interpret the coefficient on unemployment
- Does this make sense?
- What might be the problem?

Two-Period Panel Data

Why did we get such a strange result?: **omitted variable bias**

- Can we solve the problem just by adding more controls?

$$\widehat{crmrte} = \underset{(2.74)}{140.06} - \underset{(1.80)}{6.7}unem + \underset{(1.23)}{0.059}area - \underset{(1.79)}{21.963}west - \underset{(0.53)}{0.002}pcinc$$

Two-Period Panel Data

Why did we get such a strange result?: **omitted variable bias**

- Can we solve the problem just by adding more controls?

$$\widehat{crmrte} = \underset{(2.74)}{140.06} - \underset{(1.80)}{6.7}unem + \underset{(1.23)}{0.059}area - \underset{(1.79)}{21.963}west - \underset{(0.53)}{0.002}pcinc$$

- **No**
- Why? Probably because there are other important omitted variables that we can't control for

Two-Period Panel Data

How can we use panel data to deal with (some) of this problem?

Two-Period Panel Data

How can we use panel data to deal with (some) of this problem?

1) First differences

For unit i, the relationships between y and x in two time periods are as follows

$$y_{i2} = (\beta_0 + \delta_0) + \beta_1 x_{i2} + \alpha_i + u_{i2}$$
$$y_{i1} = \beta_0 + \beta_1 x_{i1} + \alpha_i + u_{i1}$$

- α_i is the subset of variables in $u_{i,t}$ that includes all time-constant characteristics of unit i that affect the outcome y.
- Assume that the effect of x on y is constant over time (β_1)
- Allow a different baseline level (intercept) of y in the two time periods (from δ_0)

Subtracting the second equation from the first:

$$\Delta y_i = \delta_0 + \beta_1 \Delta x_i + \Delta u_i$$

Two-Period Panel Data

$$\Delta y_i = \delta_0 + \beta_1 \Delta x_i + \Delta u_i$$

- a_i has been "differenced away": we don't even have to know what those time-constant omitted variables are - we have accounted for them!
- Analyze using same methods as before (e.g., let $\tilde{y} = \Delta y$), assuming you have key assumptions (most importantly, that Δu_i is uncorrelated with Δx_i)

In our crime rate example, we have panel data so can take first differences:

$$\Delta \widehat{crmrte} = 15.40 + 2.22 \Delta unem$$

That result aligns much better with our intuition! We've eliminated an important source of potential OVB, by controlling for all city characteristics that do not change over time.

Two-Period Panel Data

Another method to deal with OVB using panel data?

2) Unit Fixed Effects

- Individual dummies that control for the unit of interest
- Capture all unobserved, time-constant factors that affect crime rates in city i

We get the following result after adding city FE:

$$\widehat{crmrte} = 91.6 + 2.9unem - 0.06pcinc + \delta_1 city2 + \cdots + \delta_{46} city46 + d87$$

Again, this result makes much more sense than what we got with cross-sectional data.

Two-Period Panel Data

What do we notice about this fixed effects (FE) regression?

$$\widehat{crmrte} = 91.6 + 2.9unem - 0.06pcinc + \delta_1 city2 + \cdots + \delta_{46} city46 + d87$$

Compare to the cross-sectional regression with controls:

$$\widehat{crmrte} = 140.06 - 6.7unem + 0.059area - 21.963west - 0.002pcinc$$

1. We can't include all the same controls if we have FE. $area$ and $west$ are constant within cities, so are absorbed by city FE.
2. We need to leave out one city dummy ($city1$). As with all categorical data, city FE require a reference category.
3. We can include a control for the time period ($d87$). We could also do that if we treated the data as a repeated cross-section, but in that case we could not include unit FE.

Fixed Effects

What exactly are the fixed effects doing for our regression?

- Captures all unobserved, time constant factors within each i that affect y_{it}.
- In effect this is like adding controls for lots of unit-specific characteristics, but this way we don't have to specify what those characteristics are.

1. What type of omitted variables do we still need to worry about?
2. What type of variables does this prevent us from including in the regression?

Fixed Effects

What exactly are the unit fixed effects doing for our regression?

- Captures all unobserved, time constant factors within each i that affect y_{it}
- In effect this is like adding controls lots of individual specific characteristics

1. What type of omitted variables do we still need to worry about? **Time-varying omitted variables**
2. What type of variables does FE prevent us from including in the regression? **Time-invariant variables**

Notation:

- Denote the fixed effect with a_i or α_i for simplicity instead of including all dummies.
- The fact that this term is not indexed by a time subscript t reminds us that it does not change over time

General Period Panel Data

Example of panel data: unit of observation is a city-year. We have data for 3 cities for 3 years $\Rightarrow$ 9 total observations in our dataset.

i	t	crime rate	pop den	C 1	C 2	C 3	Yr00	Yr01	Yr02
1	2000	9.3	2.24	1	0	0	1	0	0
1	2001	11.6	2.38	1	0	0	0	1	0
1	2002	11.8	2.42	1	0	0	0	0	1
2	2000	7.6	1.61	0	1	0	1	0	0
2	2001	10.3	1.73	0	1	0	0	1	0
2	2002	11.9	1.81	0	1	0	0	0	1
3	2000	11.1	6.00	0	0	1	1	0	0
3	2001	17.2	6.33	0	0	1	0	1	0
3	2002	20.3	6.42	0	0	1	0	0	1

Interpreting Panel Regressions

We estimate this model:

$$crmrte_{it} = \beta_0 + \beta_1 popden_{it} + \alpha_2 City2 + \alpha_3 City3 +$$

$$\delta_2 Yr01 + \delta_3 Yr02 + u_{it}$$

- With multiple time periods and observations of the same units, we can include both unit and time fixed effects.
- Time fixed effects capture all variables that change over time in the same way across units (e.g., national laws that apply to all cities).
- How do we interpret β_1, α_3 or δ_3 here?

Interpreting Panel Regressions

$$crmrte_{it} = \beta_0 + \beta_1 popden_{it} + \alpha_2 City2 + \alpha_3 City3 +$$
$$\delta_2 Yr01 + \delta_3 Yr02 + u_{it}$$

1. β_1 is the marginal effect of population density on predicted crime rate controlling for the year and the city
2. α_3 we can interpret as the "effect" of City3 relative to the omitted group (City1). *i.e., what is the average difference in crime rate between City3 and City1*
3. δ_3 we can interpret as the "effect" of Year02 relative to the omitted group (Year00). *i.e., what is the average difference in crime rate between Year2 and Year0*

Interpreting α_3 and δ_3 is analogous to how we interpreted dummy variables previously.

Panel Notation

For fixed effect regressions, we save time by writing δ_t and α_i instead of writing out each dummy variable. You can imagine that if we had 40 cities and years instead of 3, writing out each dummy variable would get tedious.

- **Note the subscripts on these variables:** for a given city, the city dummy variable isn't going to vary by year, and for a given year, the year dummy variable isn't going to vary by city.

So we often write this regression as:

$$crime_{it} = \beta_0 + \beta_1 popden_{it} + \alpha_i + \delta_t + u_{it}$$

You will be asked to write panel models and we will grade you on your subscripts

Panel Regression in R

We have the model:

$$\widehat{mrdrte}_{it} = \hat{\beta}_0 + \hat{\beta}_1 unem_{it} + \underbrace{\alpha_2 State2 + ...\alpha_{50} State50}_{\text{Dummy for all but one state}}$$

$$+ \underbrace{\delta_1 Yr2001 + \delta_2 Yr2002}_{\text{Dummy for all but one year}} + u_{it}$$

How do we run this in R?

- There are a few ways! The most convient is with the `felm` command from the `lfe` package.
- This works very similar to `lm()`. The way you specify which varaibles are fixed effects are to put them after a "|" character in the formula
  - i.e. `felm(mrdrte~unem|year+state, data=mrdr)`.
  - **Note:** you will want to make sure your fixed effect variables are factors first (e.g. `mrdr$year <- as.factor(mrdr$year)`)
  - Treat the output the same way as "lm()", e.g., using "summary()", etc.

Assumptions for Fixed Effect Models

Consider the following model:

$$y_{it} = \beta_1 x_{it1} + \beta_2 x_{it2} + \cdots + \beta_k x_{itk} + \alpha_i + \delta_t + u_{it}$$

1. Assumption 1: Model is linear in parameters
2. Assumption 2: Random sample
3. Assumption 3: Each explanatory variable changes over time (for at least some i), and no perfect linear relationships exist among the explanatory variables
4. Assumption 4: $E(u_{it}|x_{it}, \alpha_i, \delta_t) = 0$, or equivalently $E(\Delta u_i|\Delta x_i) = 0$. This assumption says that we don't want changes in the u's to be correlated with changes in the x's. This assumption says that we don't want the u's in period $t-1$ to be correlated with the x's in period t or $t-1$
5. Assumption 5: $Var(u_{it}|x_{it}, \alpha_i, \delta_t) = \sigma_u^2$

Assumptions for Fixed Effect Models

Implications for recovering true population parameters/causal estimates:

1. From Assumption $A1 \rightarrow A4$ we get that β is unbiased.
2. From Assumption A5: we get an expression we can estimate for $var(\hat{\beta})$.

We have modified our model assumptions so that we know under what circumstances our estimate of β is unbiased

Assumptions for Fixed Effect Models

Consider the two regressions below using the same data:

$$y_{it} = \beta_0 + \beta_1 x_{1,it} + ... + \beta_k x_{k,it} + u_{it} \tag{1}$$

$$y_{it} = \beta_0 + \beta_1 x_{1,it} + ... + \beta_k x_{k,it} + a_i + u_{it} \tag{2}$$

1. What are the MLR.4 assumptions for each model?
2. What kind of omitted variable bias is mitigated by using model (2) instead of model (1)? (Why is model 2 *better* than model 1?)

Assumptions for Fixed Effect Models

Consider the two regressions below using the same data:

$$y_{it} = \beta_0 + \beta_1 x_{1,it} + ... + \beta_k x_{k,it} + u_{it} \tag{3}$$

$$y_{it} = \beta_0 + \beta_1 x_{1,it} + ... + \beta_k x_{k,it} + a_i + u_{it} \tag{4}$$

1. What are the MLR.4 assumptions for each model?
   For (1): $\mathbb{E}[u_{it} | x_{it1}, ..., x_{itk}] = 0$.
   For (2): $\mathbb{E}[u_{it} | x_{it1}, ..., x_{itk}, a_i] = 0$
2. What kind of omitted variable bias is mitigated by using model (2) instead of model (1)?

   Any omitted variable that is constant over time for a unit i will bias (1), but will not bias (2) because the fixed effect will capture any effect they have.

Example Questions

Let's think back to our original example of crime and unemployment, where we found that *changes* in unemployment were positively correlated with *changes* in crime when we added city fixed effects.

$$crime_{it} = \beta_0 + \beta_1 unem_{it} + a_i + \delta_t + u_{it}$$

Suppose we find $\beta_1 = 2.9$ with a standard error of 0.8

1. Interpret β_1. (Think about what we are holding constant)

Example Questions

Let's think back to our original example of crime and unemployment, where we found that *changes* in unemployment were positively correlated with *changes* in crime when we added city fixed effects.

$$crime_{it} = \beta_0 + \beta_1 unem_{it} + a_i + \delta_t + u_{it}$$

Suppose we find $\beta_1 = 2.9$ with a standard error of 0.8

1. Interpret β_1. (Think about what we are holding constant)
    - A one p.p. increase in the unemployment rate in a given city in a given year leads to 2.9 more crimes, holding all attributes about the city that don't change over time and all attributes of a year that affect all cities equally constant

Example Questions

2. Does it seem imprortant to add time dummies here? What do they control for?

3. What would cause a violation of MLR 4 here?

Example Questions

2. Does it seem imprortant to add time dummies here? What do they control for?

3. What would cause a violation of MLR 4 here?

Example Questions

2. Does it seem imprortant to add time dummies here? What do they control for?
   - Time dummies control for nationwide patterns in crime that are common across all cities in a given year. For example, if crime is decreasing everywhere, we might spuriously attribute these trends to the effects of changes in unemployment over time.
3. What would cause a violation of MLR 4 here?
   - We might think there are still a lot of unobservable things happening in cities that *vary across time and space* and are correlated with both unemployment and crime. For example, federal funding allcoated to a city might reduce unemployment through jobs programs and reduce crime by giving more resources to law enforcement.