EEP/IAS 118 - Introductory Applied Econometrics, Lecture 12

Gregory Lane

August 2017

Data Types

There are three mains types of data we are concerned with in this class:

1. Cross section
2. Pooled cross section
3. Panel Data

Data Types: Cross Section

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

Notation: we use i to index 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) individuals at many points in time.

Notation: We still only use i to index observations

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

- **Note:** we can still control for the fact that observations are from different years 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

Data Types: Panel

Panel data tracks the *same* observations over time. With panel data we start indexing observations by t as well as i

i	t	crime 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	11.1	520
100	2001	17.2	17.2	493

Two-Period Panel Data

Let's investigate a two period panel data set:

- 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:

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

- 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?

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

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

Two-Period Panel Data

How do we deal with (some) of this problem?

Fixed Effects

- Add back the second year of data and a dummy for the year
- Individual dummies that control for the unit of interest (city)
- Capture all unobserved, time-constant factors that affect crime rates

Incorporating these things we get the following result:

$$\widehat{crmrte} = 91.6 + 2.9unem + 1.8officers - 0.06income + \delta_2 city2 + \cdots + \delta_{46} city46 + d87$$

- Now the coefficient on unemployment makes sense

Fixed Effects

What exactly are the fixed effects doing for our regression?

- In our example, the FE are controlling for which city we are in
  - Captures everything unique about that city (e.g. size, climate, culture, corruption)
- Have $(i-1)$ new parameters in our regression
  - Interpret these parameters as we do other dummy variables $\Rightarrow$ δ_i is the average difference in crime rate for that city relative to the omitted group
- Leave out variables that are constant across time
  - Dropped $area$ and $west$ from the regression because they are perfectly co-linear with the city fixed effects
  - The city fixed effects already control for these constant differences

Fixed Effects

The general fixed effect model is written as:

$$y_{it} = \beta x_{it} + \gamma_t d_t + a_i + u_{it}$$

$$crimes_{it} = \beta_0 + \beta_1 unemp_{it} + \beta_2 income_{it} + a_i + d_t + u_{it}$$

- The a_i capture all unobserved, time constant factors within each i that affect y_{it}
- In effect this is like adding controls for lots of individual specific characteristics
- Note that another way to interpret the a_i is as a separate intercept for each city
- **Question:** What type of omitted variables do we still need to worry about?

Fixed Effects

$$y_{it} = \beta x_{it} + \gamma_t d_t + a_i + u_{it}$$

- What type of omitted variables do we still need to worry about?
    - **Time varying omitted variables:** these variables will *not* be controlled for in the city fixed effects
    - Can be things like changes in police practices within a city (i.e. in response to increases or decreases in crime rate)
    - Note that variables that change over time, but in the *same* way for all cities will be controlled for by d_t. E.g. national GDP growth, federal policy changes, etc.
- Fixed effects take care of *some* types of omitted variables but not all

General Period Panel Data

Expand our analysis beyond a two-year panel - unit of observation is a city-year. Example data for 3 cities for 3 years ⇒ 9 total observations in our dataset.

i	t	crime rate	pop den	C 1	C 2	C 3	Yr00	Yr01	Yr0
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 can expand our two-period model to incorporate the extra year(s):

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

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

- As before, the α capture all time constant characteristics for a given city
- The δ capture effects that are common to all cities within that year

- 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 interpret dummy variables

Panel Notation

We save time by writing δ_t and α_i instead of writing out each dummy variable. If we had 40 years instead of 3, writing out each dummy variable would get tedious.

- **Note the subscripts:** for a given city, the city dummy variable doesn't vary by year, and for a given year, the year dummy variable doesn't vary across cities.

$$crime_{it} = \beta_0 + \beta_1 popden_{it} + a_i + d_t + u_{it}$$

- Anything that is constant for an individual over time is indexed by i
- Variables that are the same for all individuals in a given time are indexed by t
- Vars that move both across time and across individuals are indexed by it

Panel Regression in Stata

We have the model:

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

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

How do we run this in Stata?

- Easiest way is using the " i.var " syntax
- In our example this would look like:

  reg crmrte unem i.stateid i.year

Panel Regression in Stata

Alternatively you could run code to generate dummy variables explicitly:

tab stateid, gen(STATE)
tab year, gen(YEAR)
reg crmrte unem STATE YEAR**

The " * " indicates that the regression should include all variables that begin with STATE or YEAR

Panel Regression in Stata

reg crmrte unem STATE YEAR**

```
      Source |       SS       df       MS              Number of obs =     153
-------------+------------------------------           F( 53,    99) =   17.75
       Model |  11622.5233    53  219.292892           Prob > F      =  0.0000
    Residual |  1222.81484    99   12.351665           R-squared     =  0.9048
-------------+------------------------------           Adj R-squared =  0.8538
       Total |  12845.3381   152  84.5088034           Root MSE      =  3.5145

------------------------------------------------------------------------------
     mrdrte |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       unem |   .2019432   .2947557     0.69   0.495    -.3829162    .7868025
     STATE2 |   2.182073   2.886745     0.76   0.452    -3.545855    7.910001
     STATE3 |   .7759888   2.897709     0.27   0.789    -4.973695    6.525672
...
    STATE50 |  -5.036179   2.927538    -1.72   0.089    -10.84505    .7726923
      YEAR2 |   1.577016   .7433858     2.12   0.036     .1019775    3.052055
      YEAR3 |   1.681938   .6959821     2.42   0.017     .3009584    3.062917
      _cons |   6.077295   3.300348     1.84   0.069    -.4713127     12.6259
------------------------------------------------------------------------------
```

Panel Regression in Stata

Finally, you can use the *xtreg* command:

xtset stateid
xtreg crmrte unem i.year, fe

- You first specify your i variable with *xtset*.
- Then run regression with *xtreg* with fixed effect option "fe"
- Note you still have to specify year dummies

All these approaches will give you the same $\hat{\beta}$ on unemployment

Assumptions for Fixed Effect Models

Consider the following model:

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

1. Assumption 1: Model is linear in parameters
2. Assumption 2: Random sample
3. Assumption 3: Each x_k needs to vary either over time t, and across units i
4. Assumption 4: $E(u_{it}|x_{it}, a_i, \delta_t) = 0$
   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: $Var(u_{it}|x_{it}, a_i, \delta_t) = \sigma_u^2$

Assumptions for Fixed Effect Models

Implications:

1. From Assumption $A1 \to A4$ 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:

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

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

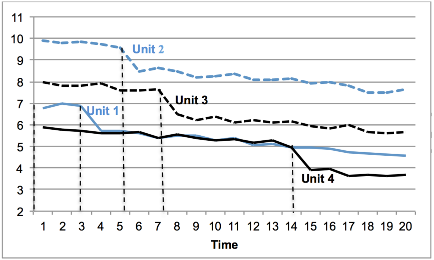
1. What are the MLR.4 assumptions for each model?
   For (1): $\mathbb{E}[u_{it}|x_{it1}, ..., x_{itk}] = 0.$
   For (2): $\mathbb{E}[u_{it}|x_{it1}, ..., x_{itk}, a_i] = 0$
2. What kind of omitted variable bias is mitigated by using model (2) instead of model (1)?

   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.

Generalized Diff-in-Diff

Before we dealt with a simple two period, two group scenario for our Diff-in-Diff estimation. What if we have something more complicated?

- Sometime treatment is introduced to different people at different points in time:



- We can use this staggered roll-out to estimate the effect of the program
- Note that here we don't have any "pure" control - everyone eventually gets treatment!

Generalized Diff-in-Diff

The idea is we want to combine the logic of our diff-in-diff regression with a panel fixed effect model

- Use the units that have not yet been treated as the comparison group for units that have been treated
- Think back to the basic two-period two-group diff-in-diff regression:

$$y = \beta_0 + \beta_1 treat + \beta_2 post + \beta_3 post \times treat + u$$

  This is very close to a two-period panel fixed effect model (with only two groups)

  - *treat* is the unit fixed effect
  - *post* is a time fixed effect
  - $post \times treat$ is the time varying variable of interest

Generalized Diff-in-Diff

We expand this simple diff-in-diff frame work to the many unit and many time period case using a panel fixed effect model:

$$y_{it} = \beta_0 + \beta T_{it} + a_i + \delta_t + u_{it}$$

Key Assumption:

- The annual change in the comparison group is a good counterfactual for the annual change in the treatment group
- As before we want to test for the validity of this assumption
- Three issues we are particularly worried about:
  1. Differential trends
  2. Ashenfelter dip - ("pre-treatment dip")
  3. Confounding policies

Generalized Diff-in-Diff, Assumption Tests

Tests for Validity of Assumption:

1. **Differential Trends:** Show that the entry into the treatment is not correlated with a differential trend in the pre-treatment period.
   - Define the change in outcome variable: $dy = y(t) - y(t-1)$
   - Define the year of introduction of the policy: $policyyear$
   - Regress the change in outcome on the year in which the law was passed in the years before the policy was implemented:

     ```
     reg  dy policyyear  if  year < firstyearpolicy
     ```

   - Want to obtain is a precise zero on the variable $policyyear$. If so, conclude that entry to treatment is not correlated with trends in the outcome variable.

Generalized Diff-in-Diff, Assumption Tests

2. **Absence of Ashenfelter dip:** We are concerned that policy was implemented in response to a sharp change in the outcome variable
    - Add two dummy variables for the year prior to and 2 years before the change in policy
    - Add them in the panel regression

      ```
      xtset state year
      xtreg y policyyear policypre_1 policypre_2  i.year, fe
      ```

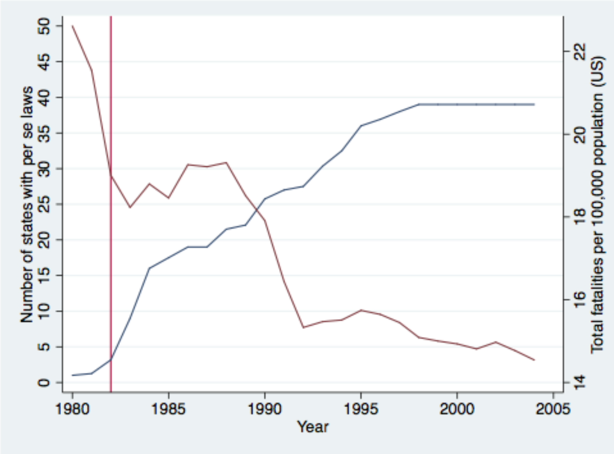
    - Again you want to make sure that the estimated coefficients on $policypre_1$, $policypre_2$ are precise zeros.

Generalized Diff-in-Diff, Assumption Tests

3. **Confounding Policies** Add other policies (or other covariates) that may be responsible for the change in outcome
   - Policies are often introduced as bundles
   - E.g. Increased change in policing coincides with a change in judicial sentencing guidelines
   - Requires knowledge of context in which policy of interest was implemented

Generalized Diff-in-Diff, Example

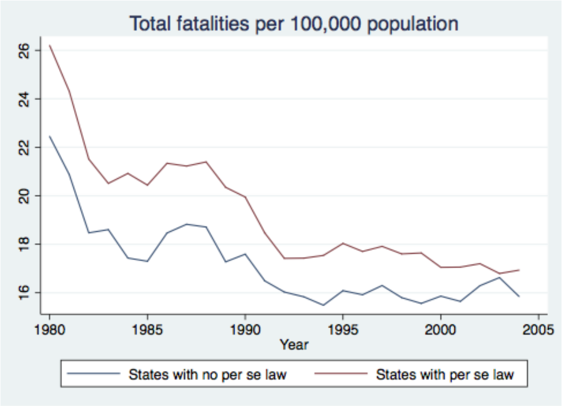
Did the introduction of "per-se" seatbelt laws reduce traffic fatalities (Freeman, D.G., 2007)? Per-se laws mean that the state can revoke your license for a DUI



Generalized Diff-in-Diff, Example

Two things to note:

1. Selection: States with higher rates of fatalities choose to introduce law
2. Time trend: Strong trend even in states without the law



Generalized Diff-in-Diff, Example

Use FE model to test hypothesis:

```
xtreg totfatrte perse i.year, fe i(state)
```

totfatrte	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval]	
perse	-1.848261	.2423821	-7.63	0.000	-2.323831	-1.37269
year						
1981	-1.814749	.4565585	-3.97	0.000	-2.710549	-.9189488
1982	-4.468642	.4566879	-9.78	0.000	-5.364697	-3.572588
....						
2002	-7.001794	.4952416	-14.14	0.000	-7.973493	-6.030095
2003	-7.267836	.4952416	-14.68	0.000	-8.239535	-6.296137
2004	-7.302419	.4952416	-14.75	0.000	-8.274118	-6.33072
_cons	25.53309	.3228739	79.08	0.000	24.89959	26.16659

We have large, negative, and significant effect. But need to test assumptions

Generalized Diff-in-Diff, Example

Differential trends:

```
. reg  dtotfatrte  perseyear d82  if year<1983 & perseyear>1982
                                               Number of obs   =        92
------------------------------------------------------------------------------
  dtotfatrte |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   perseyear |   .0135805   .0415193     0.33   0.744    -.0689175    .0960785
         d82 |  -1.187174   .6440271    -1.84   0.069    -2.466842    .0924946
       _cons |  -28.69584   82.70863    -0.35   0.729    -193.0361    135.6445
------------------------------------------------------------------------------
```

Coefficient on $dperseyear$ is small and insignificant

Generalized Diff-in-Diff, Example

Ashenfelter Dip:

```
gen perse_1 = (year == perseyear-1)
gen perse_1 = (year == perseyear-2)
```

```
. xtreg  totfatrte  perse perse_1 perse_2 i.year, fe i(state)

------------------------------------------------------------------------------
   totfatrte |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       perse |  -1.984322    .260309    -7.62   0.000    -2.495068   -1.473577
     perse_1 |    -.67682   .3903417    -1.73   0.083      -1.4427    .0890595
     perse_2 |  -.3457241   .4076816    -0.85   0.397    -1.145626    .4541778
             |
        year |
       1981  |  -1.785535   .4567127    -3.91   0.000    -2.681639   -.8894303
       1982  |  -4.355375   .4646388    -9.37   0.000    -5.267031   -3.443719
..................
```

Coefficients are not significant, in addition the point estimates are negative (here we would be concerned about positive coefficients)

Generalized Diff-in-Diff, Example

Confounding Policies:

```
. xtreg  totfatrte  perse seatbelt minage slnone zerotol gdl i.year, fe i( state)
```

totfatrte	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
perse	-2.079465	.2494411	-8.34	0.000	-2.568889	-1.590041
seatbelt	.1725957	.1263679	1.37	0.172	-.0753485	.42054
minage	.3597417	.1146316	3.14	0.002	.1348252	.5846583
bac10	-.2905969	.1939357	-1.50	0.134	-.6711148	.0899209
slnone	-.2599742	.9542762	-0.27	0.785	-2.132343	1.612394
zerotol	1.18105	.2877223	4.10	0.000	.6165153	1.745585
gdl	-.4026001	.3219036	-1.25	0.211	-1.034202	.2290014

....

Controlling for other policies doesn't change coefficient on perse