Panel Data Methods

Ketevani Kapanadze

Brno, 2020

Pooled Cross Sectional and Panel Data

An ***independently pooled cross section*** (or ***repeated cross sectional***) is obtained by sampling randomly from a large population at different points in time (for example, annual labor force surveys)

A ***panel dataset*** contains observations on multiple entities (individuals, states, companies…), where each entity is observed at two or more points in time.

Hypothetical examples:

- Data on 420 California school districts in 2010 *and again* in 2012, for 840 observations total.
- Data on 50 U.S. states, each state is observed in 3 years, for a total of 150 observations.
- Data on 1000 individuals, in four different months, for 4000 observations total.

Panel Data

A double subscript distinguishes **entities** (states) and **time** periods (years)

i = entity (state), n = number of entities,

so $i = 1,\ldots,n$

t = time period (year), T = number of time periods

so $t =1,\ldots,T$

Data: Suppose we have 1 regressor. The data are:

$$(X_{it}, Y_{it}),\ i = 1,\ldots,n,\ t = 1,\ldots,T$$

Panel Data

Panel data with *k* regressors:

$$(X_{1it}, X_{2it},\ldots,X_{kit}, Y_{it}),\ i = 1,\ldots,n,\ t = 1,\ldots,T$$

n = number of entities (states)

T = number of time periods (years)

Some jargon…

- Another term for panel data is ***longitudinal data***
- ***balanced panel***: no missing observations, that is, all variables are observed for all entities (states) and all time periods (years)

Why are Panel Data Useful?

With panel data we can control for factors that:

- Vary across entities but do not vary over time
  - These could cause omitted variable bias if they are omitted
- Are unobserved or unmeasured – and therefore cannot be included in the regression using multiple regression

Here's the key idea:

If an omitted variable does not change over time, then any *changes* in ***Y*** over time cannot be caused by the omitted variable.

Panel Data: Example of a Dataset

Observational unit: a year in a U.S. state

- 48 U.S. states, so n = # of entities = 48
- 7 years (2002,…, 2008), so T = # of time periods = 7
- Balanced panel, so total # observations = 7 × 48 = 336

Variables:

- Traffic fatality rate (# traffic deaths in that state in that year, per 10,000 state residents)
- Tax on a case of beer
- Other (legal driving age, drunk driving laws, etc.)

Policy Analysis with Pooled Cross Sections

Two or more independently sampled cross sections can be used to evaluate the impact of a certain event or policy change

- **Example: Effect of new garbage incinerator(ინსინერეიტორ) on housing prices (Kiel and McClain (1995))**
  - Examine the effect of the location of a house on its price before and after the garbage incinerator was built:

$$\widehat{rprice} = \underset{(3,093.0)}{101,307.5}\ \underset{(5,827.71)}{\boxed{-\ 30,688.27}}\ nearinc$$

After incinerator was built (1981)

$$\widehat{rprice} = \underset{(2,653.79)}{82,517.23}\ \underset{(4,744.59)}{\boxed{-\ 18,824.37}}\ nearinc$$

Before incinerator was built (1978)

Policy Analysis with Pooled Cross Sections

- **Example: Garbage incinerator and housing prices (cont.)**
  - It would be wrong to conclude from the regression after the incinerator is there that being near the incinerator depresses prices so strongly
  - One has to compare with the situation before the incinerator was built:

$$\widehat{\delta}_1 = -30,688.27 - (-18,824.37) = \boxed{-11,863.9}$$

Incinerator depresses prices but location was one with lower prices anyway

  - In the given case, this is equivalent to

$$\widehat{\delta}_1 = (\overline{rprice}_{1,nr} - \overline{rprice}_{1,fr}) - (\overline{rprice}_{0,nr} - \overline{rprice}_{0,fr})$$

- **This is the so called difference-in-differences estimator (DiD)**

Policy Analysis with Pooled Cross Sections

- **Difference-in-differences in a regression framework**

$$rprice = \beta_0 + \delta_0\, after + \beta_1\, nearinc + \boxed{\delta_1}\, after \cdot nearinc + u$$

Differential effect of being in the location and after the incinerator was built

  - In this way standard errors for the DiD-effect can be obtained
  - If houses sold before and after the incinerator was built were systematically different, further explanatory variables should be included
  - This will also reduce the error variance and thus standard errors

- **Before/After comparisons in „natural experiments“**
  - DiD can be used to evaluate policy changes or other exogenous events

Policy Analysis with Pooled Cross Sections

- **Policy evaluation using difference-in-differences**

$$y = \beta_0 + \delta_0 after + \beta_1 treated + \delta_1 after \cdot treated + other factors$$

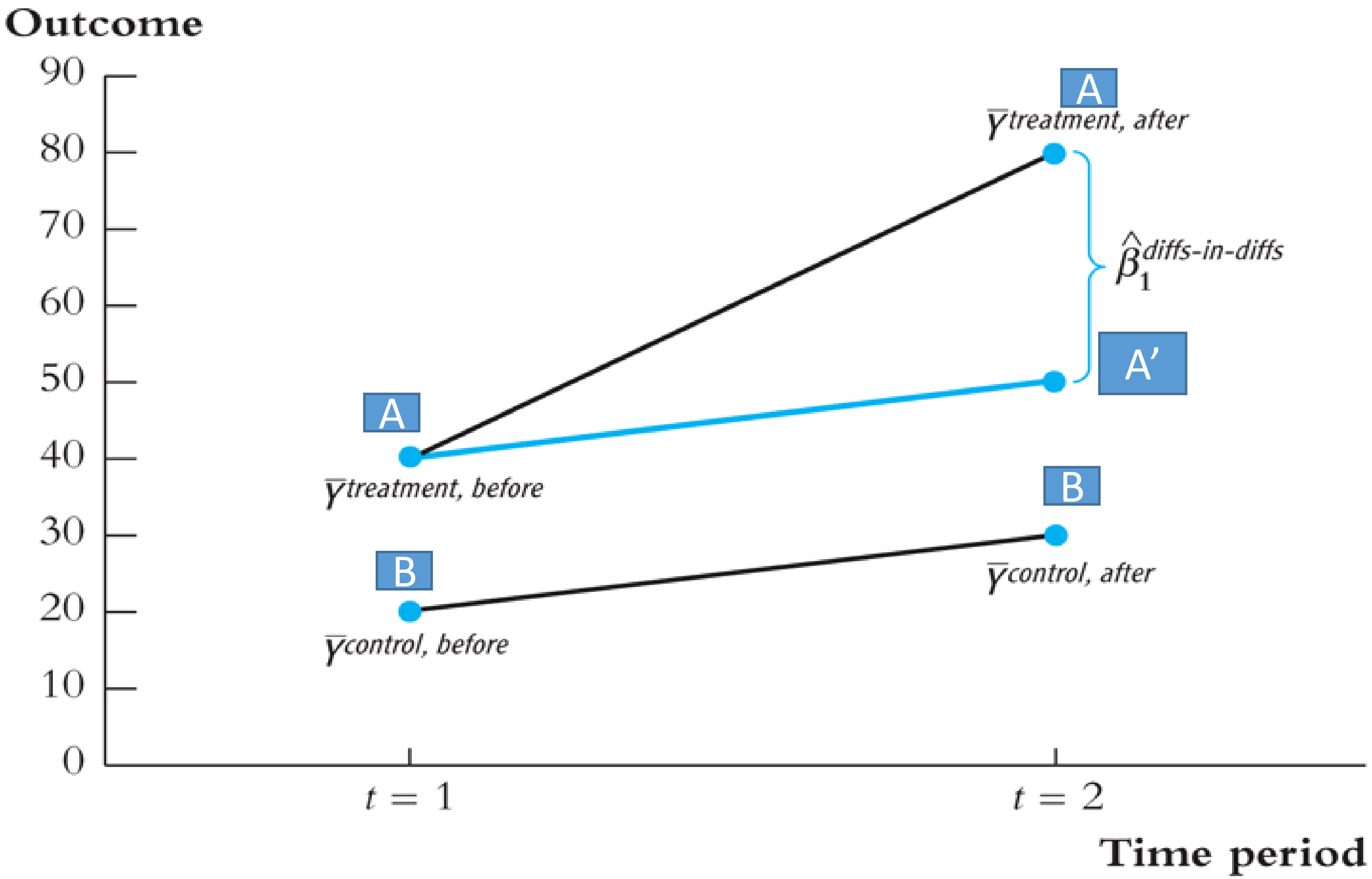
$$\widehat{\delta}_1 = (\overline{y}_{1,T} - \overline{y}_{1,C}) - (\overline{y}_{0,T} - \overline{y}_{0,C}) \leftarrow$$

Compare outcomes of the two groups before and after the policy change

Caution: Difference-in-differences only works if the difference in outcomes between the two groups is not changed by other factors than the policy change (e.g. there must be no differential trends).

Diff-in-Diff Estimator (DID)

$$\hat{\beta}_1^{diffs-in-diffs} = (\bar{Y}^{treat,after} - \bar{Y}^{treat,before}) - (\bar{Y}^{control,after} - \bar{Y}^{control,before})$$



Two-Period Panel Data Analysis

- **Example: Effect of unemployment on city crime rate**

$$crmrte_{it} = \beta_0 + \delta_0 \, d87_{it} + \beta_1 \, unem_{it} + a_i + u_{it}, \quad t = 1982, 1987$$

Time dummy for the second period

Unobserved time-constant factors (= <u>fixed effect</u>)

Other unobserved factors (= idiosyncratic error)

Two-Period Panel Data Analysis

- **Example: Effect of unemployment on city crime rate (cont.)**

$$crmrte_{i1987} = \beta_0 + \delta_0 \cdot 1 + \beta_1\, unem_{i1987} + a_i + u_{i1987}$$

$$crmrte_{i1982} = \beta_0 + \delta_0 \cdot 0 + \beta_1\, unem_{i1982} + a_i + u_{i1982}$$

Subtract: $\Rightarrow \boxed{\Delta crmrte_i} = \delta_0 + \beta_1 \boxed{\Delta unem_i} + \boxed{\Delta u_i}$

- **Estimate differenced equation by OLS:**

$$\Delta \widehat{crmrte} = \underset{(4.70)}{15.40} + \underset{(.88)}{2.22}\, \Delta unem \quad \leftarrow$$

$$n = 46, R^2 = .127$$

Secular increase in crime

Two-Period Panel Data Analysis

- **Discussion of first-differenced panel estimator**
  - Further explanatory variables may be included in the original equation
  - Note that there may be arbitrary correlation between the unobserved time-invariant characteristics and the included explanatory variables
  - OLS in the original equation would therefore be inconsistent
  - <u>The first-differenced panel estimator is thus a way to consistently estimate causal effects in the presence of time-invariant endogeneity</u>
  - For consistency, strict exogeneity has to hold in the original equation
  - First-differenced estimates will be imprecise if explanatory variables vary only little over time (no estimate possible if time-invariant)

Fixed Effects Estimation

Consider the panel data model,

$$FatalityRate_{it} = \beta_0 + \beta_1 BeerTax_{it} + \beta_2 Z_i + u_{it}$$

$\boldsymbol{Z_i}$ is a factor that does not change over time, at least during the years on which we have data

(examples: ; density of cars on the road;).

- Suppose $\boldsymbol{Z_i}$ is not observed, so its omission could result in omitted variable bias.
- The effect of $\boldsymbol{Z_i}$ can be eliminated using T = 2 years by method described above (diff- diff).

Fixed Effects Estimation

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + u_{it},\ i = 1,\ldots,n,\ T = 1,\ldots,T$$

We can rewrite this in two useful ways:

1. "***n*-1 binary regressor**" regression model

2. "**Fixed Effects**" regression model

Fixed Effects Estimation

Population regression for California (that is, *i* = CA):

$$Y_{CA,t} = \beta_0 + \beta_1 X_{CA,t} + \beta_2 Z_{CA} + u_{CA,t}$$

$$= (\beta_0 + \beta_2 Z_{CA}) + \beta_1 X_{CA,t} + u_{CA,t}$$

Or

$$Y_{CA,t} = \alpha_{CA} + \beta_1 X_{CA,t} + u_{CA,t}$$

- $\alpha_{CA} = \beta_0 + \beta_2 Z_{CA}$ **doesn't change over time**
- α_{CA} is the intercept for CA, and β_1 is the slope
- The intercept is unique to CA, but the slope is the same in all the states: parallel lines.

Fixed Effects Estimation

$$
\begin{aligned}
Y_{TX,t} &= \beta_0 + \beta_1 X_{TX,t} + \beta_2 Z_{TX} + u_{TX,t} \\
&= (\beta_0 + \beta_2 Z_{TX}) + \beta_1 X_{TX,t} + u_{TX,t} \quad (\underline{\textit{population regression for Texas}})
\end{aligned}
$$

or

$$Y_{TX,t} = \alpha_{TX} + \beta_1 X_{TX,t} + u_{TX,t}, \text{ where } \alpha_{TX} = \beta_0 + \beta_2 Z_{TX}$$

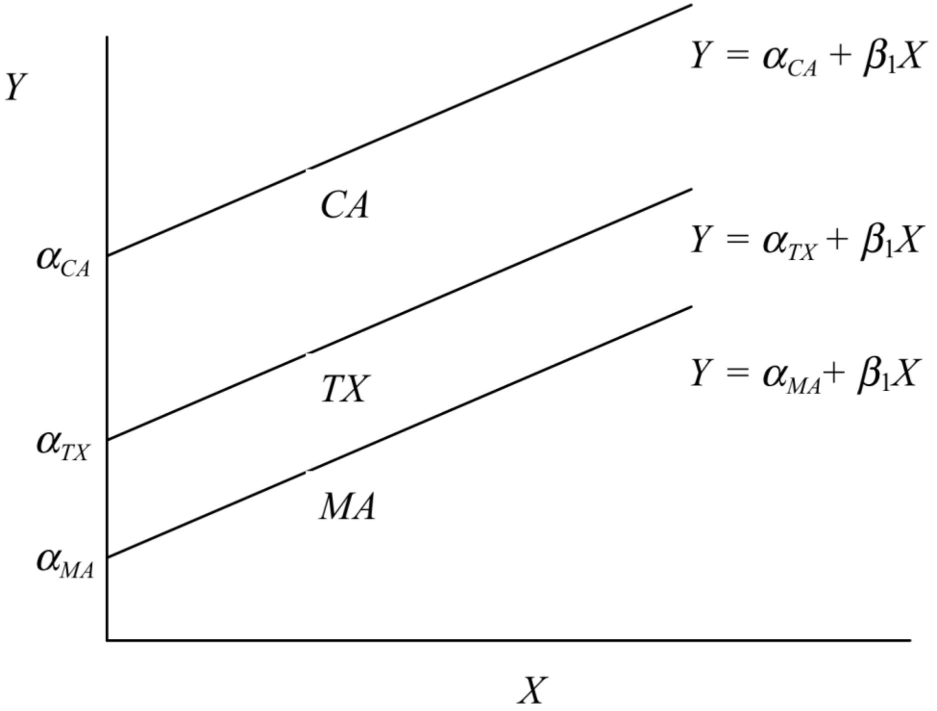
Collecting the lines for all three states:

$$
\begin{aligned}
Y_{CA,t} &= \alpha_{CA} + \beta_1 X_{CA,t} + u_{CA,t} \\
Y_{TX,t} &= \alpha_{TX} + \beta_1 X_{TX,t} + u_{TX,t} \\
Y_{MA,t} &= \alpha_{MA} + \beta_1 X_{MA,t} + u_{MA,t}
\end{aligned}
$$

or

$$Y_{it} = \alpha_i + \beta_1 X_{it} + u_{it},\ i = \text{CA, TX, MA},\ T = 1,\ldots,T$$

Fixed Effects Estimation



Y
$Y = \alpha_{CA} + \beta_1 X$
CA
α_{CA}
$Y = \alpha_{TX} + \beta_1 X$
TX
α_{TX}
$Y = \alpha_{MA} + \beta_1 X$
MA
α_{MA}
X

In binary regressor form:

$$Y_{it} = \beta_0 + \gamma_{CA} DCA_i + \gamma_{TX} DTX_i + \beta_1 X_{it} + u_{it}$$

- DCA_i = 1 if state is *CA*, = 0 otherwise
- DTX_t = 1 if state is *TX*, = 0 otherwise
- leave out DMA_i (*why*?)

Fixed Effects Estimation

1. **“*n*-1 binary regressor” form**

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \gamma_2 D2_i + \ldots + \gamma_n Dn_i + u_{it}$$

where $D2_i = \begin{cases} 1 \text{ for } i\text{=2 (state \#2)} \\ 0 \text{ otherwise} \end{cases}$, etc.

2. **“Fixed effects” form:**

$$Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it}$$

- α_i is called a “state fixed effect” or “state effect” – it is the constant (fixed) effect of being in state i

Fixed Effects Estimation

- **Fixed effects estimation**

Fixed effect, potentially correlated with explanatory variables

$$y_{it} = \beta_1 x_{it1} + \cdots + \beta_k x_{itk} + \enclose{circle}{a_i} + u_{it}, \quad i = 1, \ldots, N, t = 1, \ldots, T$$

$$\bar{y}_i = \beta_1 \bar{x}_{i1} + \cdots + \beta_k \bar{x}_{ik} + \bar{a}_i + \bar{u}_i$$

Form time-averages for each individual

$$\Rightarrow \quad \boxed{[y_{it} - \bar{y}_i]} = \beta_1 \boxed{[x_{it1} - \bar{x}_{i1}]} + \cdots + \beta_k \boxed{[x_{itk} - \bar{x}_{ik}]} + \boxed{[u_{it} - \bar{u}_i]}$$

Because $a_i - \bar{a}_i = 0$ (the fixed effect is removed)

- **Estimate time-demeaned equation by OLS**
  - Uses time variation within cross-sectional units (= within-estimator)

Fixed Effects Estimation with Time Fixed Effects

An omitted variable might vary **over time** but **not across states**:

- Safer cars (air bags, etc.); changes in national laws
- These produce **intercepts that change over time**

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + \beta_3 S_t + u_{it}$$

Fixed Effects Estimation with Time Fixed Effects

$$\begin{aligned} Y_{i,1982} &= \beta_0 + \beta_1 X_{i,1982} + \beta_3 S_{1982} + u_{i,1982} \\ &= (\beta_0 + \beta_3 S_{1982}) + \beta_1 X_{i,1982} + u_{i,1982} \\ &= \lambda_{1982} + \beta_1 X_{i,1982} + u_{i,1982}, \end{aligned}$$

where $\lambda_{1982} = \beta_0 + \beta_3 S_{1982}$ Similarly,

$$Y_{i,1983} = \lambda_{1983} + \beta_1 X_{i,1983} + u_{i,1983},$$

where $\lambda_{1983} = \beta_0 + \beta_3 S_{1983}$, etc.

Fixed Effects Estimation with Time Fixed Effects

1. "*T*-1 binary regressor" formulation:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \delta_2 B2_t + \ldots \delta_T BT_t + u_{it}$$

where $B2_t = \begin{cases} 1 \text{ when } t=2 \text{ (year \#2)} \\ 0 \text{ otherwise} \end{cases}$, etc.

2. "Time effects" formulation:

$$Y_{it} = \beta_1 X_{it} + \lambda_t + u_{it}$$

Fixed Effects Estimation

- **Discussion of fixed effects estimator**
  - Strict exogeneity in the original model has to be assumed
  - The *R-squared* of the demeaned equation is inappropriate
  - The effect of time-invariant variables cannot be estimated

Final Exam

- May 15, at 9am in Zoom ☺
- Exam will take place in Zoom, May 15, at 9am-11am
- ⍰ Let’s meet in the Zoom at 8:45am, to check that there are no technical issues.
- ⍰ Exam will start exactly at 9am!
- ⍰ Please make sure you have good internet connection
- ⍰ All cameras MUST be turned on
- ⍰ You can ask questions during the exam ONLY in the private chat
- ⍰ It is closed book exam, cheating on final exam can result in serious consequences for the
- student
- ⍰ Handwritings must be legible enough!
- ⍰ At 8:55am I will share protected final exam file to the class
- ⍰ At 11 am, exam is over, you will take photos of your solutions and send them to my email
- address, during the meeting. I will close the exam meeting as soon as I get all your exam
- solutions
- ⍰ Don’t forget to write your name and surname in the email, and in the SUBJECT of the email
- you must write down “Metrics Final Exam”.