Panel Data Methods

Taro Mieno

AECN 896-003: Applied Econometrics

Panel Data

The Focus

What can we do with panel data to mitigate bias, which is not possible with cross-sectional data?

As it will turn out,

Panel data can be very powerful compared to cross-sectional data

Data type

Independently pooled cross sectional (IPCS) data

Data obtained by sampling randomly from a large population at different points in time without deliberate effort to collect observations of the same cross-sectional units across time

- sample of hourly wage, education, experience, and so on from the population of working people each year

Data type

Panel (longitudinal) data

Data follow the same individuals, families, firms, cities, states or whatever, across time

- randomly selecting people from a population at a given point in time.
- then the same people are reinterviewed at several subsequent points in time, which would results in data on wages, hours, education, and so on, for the same group of people in different years.

Panel Data

```
  year  fcode employ    sales
1 1987 410032    100 47000000
2 1988 410032    131 43000000
3 1989 410032    123 49000000
4 1987 410440     12  1560000
5 1988 410440     13  1970000
6 1989 410440     14  2350000
7 1987 410495     20   750000
8 1988 410495     25   110000
9 1989 410495     24   950000
```

New way of writing a model

$$y_{i,t} = \beta_0 + \beta_1 x_{1,i,t} + \beta_2 x_{2,i,t} + \ldots \beta_k x_{k,i,t} + u_{i,t}$$

- i: indicates cross-sectional unit
- t: indicates time

Pooled OLS estimation

Proceed exactly the same way as before with OLS on a cross-sectional dataset

Analysis of IPCS data

- Analytical framework and method we have used can be used: there is not much you can do differently than with cross-sectional data
- One reason for using IPCS is to increase sample size for estimation accuracy
- It can be used to do program (impact) evaluation more credibly than using cross-sectional data

Program evaluation (policy analysis)

Randomized experiment

Changes in the environment in which agents operate, which are assigned randomly to agents by the investigators

- financial aid to start up a new business provided randomly to some people
- Randomly change school/teach ratio to measure its impact on education quality

Program evaluation (policy analysis)

Randomized experiment

Changes in the environment in which agents operate, which are assigned randomly to agents by the investigators

- financial aid to start up a new business provided randomly to some people
- Randomly change school/teach ratio to measure its impact on education quality

Evaluation of a randomized experiment

Great! Not really much to worry about!

$$y \text{ (income)} = \beta_0 + \beta_1 program \text{ (financial aid)} + u$$

, where $E[u|program] = 0$. OLS is just fine.

Program evaluation (policy analysis)

Natural Experiment (Quasi-experiment)

An event or policy change (often a change in government policy) that happens outside of the control of investigators, which changes the environment in which agents (individuals, families, firms, or cities) operate.

- availability of livestock insurance program
- job training program
- water use limits

Program evaluation (policy analysis)

Natural Experiment (Quasi-experiment)

An event or policy change (often a change in government policy) that happens outside of the control of investigators, which changes the environment in which agents (individuals, families, firms, or cities) operate.

- availability of livestock insurance program
- job training program
- water use limits

Evaluation of a natural experiment

Gotta be more careful!

$$y \text{ (land price)} = \beta_0 + \beta_1 program \text{ (livestock insurance)} + u$$

Program evaluation (policy analysis)

Natural Experiment (Quasi-experiment)

An event or policy change (often a change in government policy) that happens outside of the control of investigators, which changes the environment in which agents (individuals, families, firms, or cities) operate.

- availability of livestock insurance program
- job training program
- water use limits

Evaluation of a natural experiment

Gotta be more careful!

$$y \text{ (land price)} = \beta_0 + \beta_1 program \text{ (livestock insurance)} + u$$

Is $E[u|program] = 0$?

Evaluation of a natural experiment using IPCS data

If you have IPCS data, instead of cross-sectional data, you may be able to evaluate the impact of a program more credibly!

Natural experiment data structure

Observations on two groups before and after the program placement:

- Treated (treated at $t = 2$)
  - observed at $t = 1$ (before)
  - observed at $t = 2$ (after)
- Control (untreated)
  - observed at $t = 1$ (before)
  - observed at $t = 2$ (after)

Difference-in-Differences (DID) Estimation

- DID can be conducted using either IPCS or panel data, but not using cross-sectional data
- DID can be a very useful strategy to estimate the impact of policy changes

DID by example: The impact of incinerator

Incinerator construction

- rumored about the incinerator being built in North Andover, Massachusetts, began in 1978
- construction started in 1981

Data collected

Housing prices in 1978 and 1981, and other variables

Naive analysis (sort of what you did in assignment 1)

Run regression on the following model using the 1981 data (cross-sectional data)

$$rpice = \gamma_0 + \gamma_1 nearinc + u$$

- $rprice$: house price in real terms (inflation-corrected)
- $nearinc$: 1 if the house is near the incinerator, and 0 otherwise
- γ_1: the difference between the mean house price of houses nearby the incinerator and the rest (not nearby) in 1981

Naive analysis (sort of what you did in assignment 1)

Run regression on the following model using the 1981 data (cross-sectional data)

$$rpice = \gamma_0 + \gamma_1 nearinc + u$$

- $rprice$: house price in real terms (inflation-corrected)
- $nearinc$: 1 if the house is near the incinerator, and 0 otherwise
- γ_1: the difference between the mean house price of houses nearby the incinerator and the rest (not nearby) in 1981

What is the problem with this approach?

Is $nearinc$ endogenous?

R code: OLS using the 1981 data

```r
data <- read.dta13('KIELMC.dta') %>%
  mutate(rprice=rprice/1000)
reg_81 <-lm(rprice~nearinc,data=filter(data,year==1981))
```

Table

	Dependent variable:	
	rprice	
nearinc	-30.688^{***}	(5.828)
Constant	101.308^{***}	(3.093)

Naive analysis

Run regression on the following model using the 1978 data

$$rpice = \lambda_0 + \lambda_1 nearinc + u$$

R code: OLS using the 1978 data

```
reg_78 <-lm(rprice~nearinc,data=filter(data,year==1978))
```

Table

	Dependent variable:
	rprice
nearinc	-18.824^{***} (4.745)
Constant	82.517^{***} (2.654)

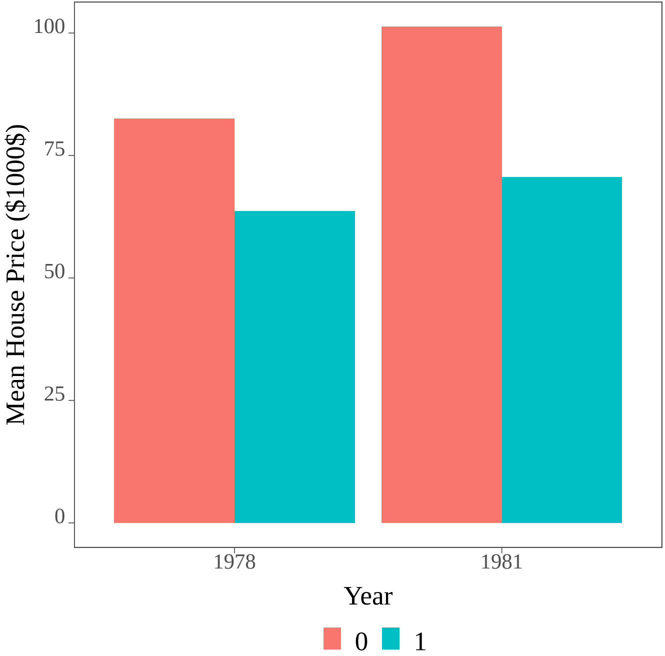
So,

Houses nearby the incinerator were already lower than those houses that are not nearby...

R code: Visualization

```r
data_mean <- data %>%
  group_by(year,nearinc) %>%
  summarize(m_rprice=mean(rprice))

g_dif <- ggplot(data=data_mean) +
  geom_bar(aes(y=m_rprice,x=factor(year),fill=factor(nearinc)),
  stat='identity',position='dodge') +
  ylab('Mean House Price ($1000$)') +
  xlab('Year') +
  scale_fill_discrete(name='') +
  theme(
    legend.position='bottom'
  )
```


Mean House Price (1000)
100
75
50
25
0
1978
1981
Year
0
1

Estimated Model

(1981) $rpice = \gamma_0 + \gamma_1 nearinc + u$

(1978) $rpice = \lambda_0 + \lambda_1 nearinc + u$

According to the models,

$$E[rprice|year = 1981, nearinc = 0] = \gamma_0$$

$$E[rprice|year = 1981, nearinc = 1] = \gamma_0 + \gamma_1$$

$$E[rprice|year = 1978, nearinc = 0] = \lambda_0$$

$$E[rprice|year = 1978, nearinc = 1] = \lambda_0 + \lambda_1$$

According to the models,

$$E[rprice|year = 1981, nearinc = 0] = \gamma_0$$

$$E[rprice|year = 1981, nearinc = 1] = \gamma_0 + \gamma_1$$

$$E[rprice|year = 1978, nearinc = 0] = \lambda_0$$

$$E[rprice|year = 1978, nearinc = 1] = \lambda_0 + \lambda_1$$

Interpretation

$$\begin{aligned} \gamma_1 = & E[rprice|year = 1981, nearinc = 1] \\ & - E[rprice|year = 1981, nearinc = 0] \\ \lambda_1 = & E[rprice|year = 1978, nearinc = 1] \\ & - E[rprice|year = 1978, nearinc = 0] \end{aligned}$$

Interpretation

$$\begin{aligned}\gamma_1 =& E[rprice|year = 1981, nearinc = 1]\\ &- E[rprice|year = 1981, nearinc = 0]\\ \lambda_1 =& E[rprice|year = 1978, nearinc = 1]\\ &- E[rprice|year = 1978, nearinc = 0]\end{aligned}$$

DID

$$\begin{aligned}DID =& \gamma_1 - \lambda_1\\ =& \Big(E[rprice|year = 1981, nearinc = 1]\\ &- E[rprice|year = 1978, nearinc = 1]\Big)\\ &- \Big(E[rprice|year = 1981, nearinc = 0]\\ &- E[rprice|year = 1978, nearinc = 0]\Big)\end{aligned}$$

DID

$$
\begin{aligned}
DID =& \gamma_1 - \lambda_1 \\
=& \Big(E[rprice|year = 1981, nearinc = 1] \\
& - E[rprice|year = 1978, nearinc = 1] \Big) \\
& - \Big(E[rprice|year = 1981, nearinc = 0] \\
& - E[rprice|year = 1978, nearinc = 0] \Big)
\end{aligned}
$$

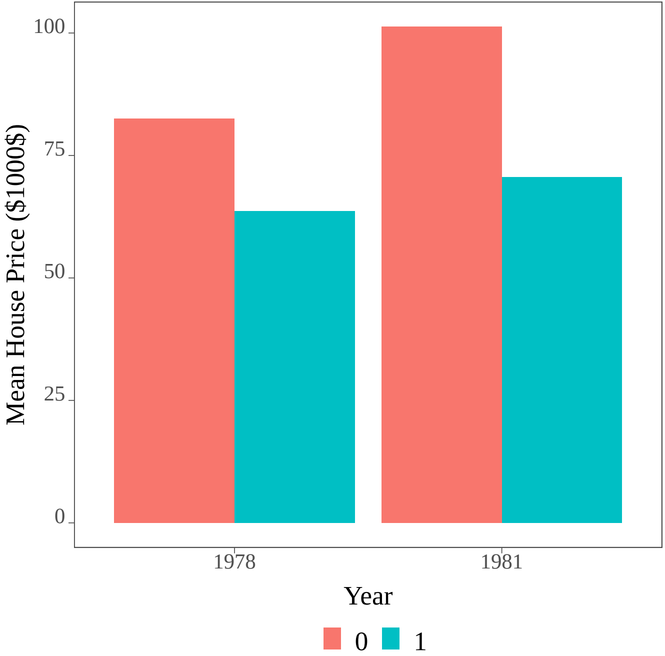

Mean House Price (1000)
100
75
50
25
0
1978
1981
Year
0
1

Table: Expected House Price

	before	after
nearinc=0	γ_0	$\gamma_0 + \alpha_0 + 0$
nearinc=1	γ_1	$\gamma_1 + \alpha_1 + \beta$

- γ_j is the expected house price of those that are $nearinc = j$
- α_j is any macro shocks other than the incinerator event that happened between the before and after period to the houses that are $nearinc = j$
- β is the true causal impact of the incinerator placement

Table: Expected House Price

	before	after
nearinc=0	γ_0	$\gamma_0 + \alpha_0 + 0$
nearinc=1	γ_1	$\gamma_1 + \alpha_1 + \beta$

OLS on the 1981 data

$$\begin{aligned} E[\hat{\beta}] =& E[rpice|nearinc == 1, after] \\ & - E[rpice|nearinc == 0, after] \\ =& (\gamma_1 + \alpha_1 + \beta) - (\gamma_0 + \alpha_0) \\ =& [(\gamma_1 - \gamma_0) + (\alpha_1 - \alpha_0)] + \beta \end{aligned}$$

Bias

$$E[bias] = [(\gamma_1 - \gamma_0) + (\alpha_1 - \alpha_0)]$$

Table: Expected House Price

	before	after
nearinc=0	γ_0	$\gamma_0 + \alpha_0 + 0$
nearinc=1	γ_1	$\gamma_1 + \alpha_1 + \beta$

OLS on the IPCS data without the control group

$$\begin{aligned} E[\hat{\beta}] =& E[rpice|nearinc == 1, after] \\ & - E[rpice|nearinc == 1, before] \\ =& (\gamma_1 + \alpha_1 + \beta) - (\gamma_1) \\ =& [\alpha_1] + \beta \end{aligned}$$

Bias

$$E[bias] = \alpha_1$$

Table: Expected House Price

	before	after
nearinc=0	γ_0	$\gamma_0 + \alpha_0 + 0$
nearinc=1	γ_1	$\gamma_1 + \alpha_1 + \beta$

DID

$$
\begin{aligned}
E[\hat{\beta}] =& \Big(E[rprice|year = 1981, nearinc = 1] \\
& - E[rprice|year = 1978, nearinc = 1]\Big) \\
& - \Big(E[rprice|year = 1981, nearinc = 0] \\
& - E[rprice|year = 1978, nearinc = 0]\Big) \\
& = \Big((\gamma_1 + \alpha_1 + \beta) - (\gamma_1)\Big) - \Big((\gamma_0 + \alpha_0 + \beta) - (\gamma_0)\Big) \\
& = (\alpha_1 + \beta) - (\alpha_0) = (\alpha_1 - \alpha_0) + \beta
\end{aligned}
$$

Bias

$$E[bias] = \alpha_1 - \alpha_0$$

Condition under which DID works (Unbiased)

$$\alpha_1 = \alpha_0$$

The dependent variable of the treatment ($nearinc = 1$) group would have grown by the same amount as the control group ($nearinc = 0$) if it were not for the treatment (incinerator construction)

Important

- The above condition is untestable
- If there were any significant changes in policy that happen to only one of them other than the treatment of interest, DID confound the impact of the other policy and the treatment of interest
- It is a common practice to present the trend of the dependent variable prior to the year before the treatment happened

DID in single regression

$$rprice = \beta_0 + \sigma_0 y81 + \beta_1 nearinc + \sigma_1 y81 \cdot nearinc + u$$

where $\hat{\sigma_1}$ is the DID estimate. You can include other variables as controls.

Two-Period Panel Data Analysis

Panel (longitudinal) data

Data follow the same individuals, families, firms, cities, states or whatever, across time

- randomly selecting people from a population at a given point in time.
- then the same people are reinterviewed at several subsequent points in time, which would results in data on wages, hours, education, and so on, for the same group of people in different years.

Two-period panel data

For a cross section of individuals, schools, firms, cities, or whatever, we have two years of data; call these $t = 1$ and $t = 2$. These years need not be adjacent, but $t = 1$ corresponds to the earlier year.

An example we use

Crime and unemployment rates data

- 46 cities (cross-sectional unit is city)
- 1982 $(t = 1)$ and 1987 $(t = 2)$

OLS

Regression analysis using only the 1987 data (cross-sectional);

$$crmrte = \beta_0 + \beta_1 unem + u$$

R code: Crime Data

```r
data <- read.dta13('CRIME2.dta')
reg_87 <-lm(crmrte~unem,data=filter(data,year==87))
```

Table

	Dependent variable:
	crmrte
unem	-4.161 (3.416)
Constant	128.378*** (20.757)

Table

	Dependent variable:
	crmrte
unem	-4.161 (3.416)
Constant	128.378^{***} (20.757)

Question

- Is the negative sign of the coefficient estimate on $unem$ what you expected?
- Any problem with this regression?
  - we could include other controls like age distribution, gender distribution, education levels, law enforcement efforts, etc
  - there could be many other factors that cannot be observed

Taking advantage of the panel data structure

You can take advantage of the panel data structure to mitigate the omitted variable bias

View the unobserved factors differently

$$y_{i,t} = \beta_0 + \sigma_0 d2_t + \beta_1 x_{i,t} + v_{i,t}$$

- subscripts i and t indicate cross-sectional unit (city) and time (year), respectively
- $d2_t$: 0 when $t = 1$ and 1 when $t = 2$ (no i subscript because it does not change across i)

The error term: $v_{i,t}$

The error term consists of two parts: α_i and $u_{i,t}$

$$v_{i,t} = \alpha_i + u_{i,t}$$

- α_i: all unobserved, time-invariant factors of i that affect $y_{i,t}$ (referred to as fixed effects and unobserved heterogeneity)
- $u_{i,t}$: the rest of the error that is time-varying (often referred to as idiosyncratic error)

Fixed (Unobserved) Effects Model Example

$$crmrte_{i,t} = \beta_0 + \sigma_0 d87_t + \beta_1 unem_{i,t} + \alpha_i + u_{i,t}$$

- Since i denotes different cities, we call α_i an unobserved city effect or a city fixed effect

city fixed effect: α_i

- It represents all factors affecting city crime rates that do not change over time. (Geographical features, such as the city's location in the United States, are included in α_i)
- Many other factors may not be exactly constant, but they might be roughly constant over a five-year period (slow to change)
    - certain demographic features of the population (age, race, and education)
    - different cities may have their own methods for reporting crimes
    - the people living in the cities might have different attitudes toward crime

Estimation of the model: Pooled OLS (POLS)

Pools the data and run OLS on the following model

$$y_{i,t} = \beta_0 + \sigma_0 d2_t + \beta_1 x_{i,t} + v_{i,t}$$

where $v_{i,t} = \alpha_i + u_{i,t}$

Condition under which POLS is unbiased

$$\begin{aligned} & E[v_{i,t}|x_{i,t}] = 0 \\ \Rightarrow & E[u_{i,t}|x_{i,t}] = 0 \text{ and } E[\alpha_i|x_{i,t}] = 0 \end{aligned}$$

R code: Crime Data

```
data <- read.dta13('CRIME2.dta')
reg_pols <-lm(crmrte˜unem+d87,data=data)
```

Table

	Dependent variable:
	crmrte
unem	0.427 (1.188)
d87	7.940 (7.975)
Constant	93.420*** (12.739)

Taking advantage of the panel structure

The Idea

Difference α_i out!!

$$\begin{aligned}(t=2)\ \ y_{i,2} &= \beta_0 + \sigma_0 + \beta_1 x_{i,2} + a_i + u_{i,2}\\ (t=1)\ \ y_{i,1} &= \beta_0 + \beta_1 x_{i,1} + a_i + u_{i,1}\end{aligned}$$

Subtracting the second from the first,

$$\begin{aligned}(y_{i,2} - y_{i,1}) &= \sigma_0 + \beta_1 (x_{i,2} - x_{i,1}) + (u_{i,2} - u_{i,1}) \ \text{ or}\\ \Delta y_i &= \sigma_0 + \beta_1 \Delta x_i + \Delta u_i\end{aligned}$$

We call this first-differenced equation.

First-differenced (FD) estimator

Run OLS on the first differenced model:

$$\Delta y_i = \sigma_0 + \beta_1 \Delta x_i + \Delta u_i$$

First-differenced (FD) estimator

Run OLS on the first differenced model:

$$\Delta y_i = \sigma_0 + \beta_1 \Delta x_i + \Delta u_i$$

Strict Exogeneity: conditions for unbiasedness

$$E[\Delta u_{i,t} | \Delta x_{i,t}] = 0,$$

which is satisfied if

$$E[u_{i,t} | x_{i,s}] = 0 \ ^{\forall} s \text{ and } t,$$

meaning

$$\begin{aligned} &E[u_{i,1} | x_{i,1}] = 0 \\ &E[u_{i,2} | x_{i,1}] = 0 \\ &E[u_{i,1} | x_{i,2}] = 0 \\ &E[u_{i,2} | x_{i,2}] = 0 \end{aligned}$$

Strict Exogeneity: conditions for unbiasedness

$$E[\Delta u_{i,t}|\Delta x_{i,t}] = 0,$$

which is satisfied if

$$E[u_{i,t}|x_{i,s}] = 0 \ {}^{\forall}s \ \text{ and } \ t,$$

meaning

$$\begin{aligned} &E[u_{i,1}|x_{i,1}] = 0 \\ &E[u_{i,2}|x_{i,1}] = 0 \\ &E[u_{i,1}|x_{i,2}] = 0 \\ &E[u_{i,2}|x_{i,2}] = 0 \end{aligned}$$

Important

It is okay that $x_{i,t}$ is correlated with α_i unlike POLS (we sometimes say correlation between $x_{i,t}$ with α_i is allowed).

Implementation of FD estimation in R

plm package

- $pdata.frame()$: create a data.frame that is aware of the cross-sectional units and time
- plm: implement a variety of panel data estimation methods

R code: Crime Data

```
  #--- library plm ---#
  library(plm)

  #--- take a look at the portion of the data ---#
  data %>% dplyr::select(crmrte,unem,year) %>% head()

    crmrte unem year
1 74.65756  8.2   82
2 70.11729  3.7   87
3 92.93487  8.1   82
4 89.97221  5.4   87
5 83.61113  9.0   82
6 77.19476  5.9   87

  dplyr::select(data,city_id,crmrte,unem,year)

Error in .f(.x[[i]], ...): object 'city_id' not found
  #--- create city (cross-sectional unit) id ---#
  # you normally do not have to do this
  data <- mutate(data,city_id=rep(seq(1,46),each=2))

  #--- convert the dataset ---#
  pdata_crm <- pdata.frame(data,index=c('city_id','year'))
```

R code: FD estimation

```
  #--- FD estimation ---#
  reg_fd <- plm(crmrte~unem,data=pdata_crm,model='fd')

  #--- summary of the estimation ---#
  summary(reg_fd)$coef
```

```
            Estimate Std. Error  t-value    Pr(>|t|)
(Intercept)  15.4022  4.7021169 3.275589 0.002060469
unem          2.2180  0.8778658 2.526581 0.015189318
```

R code: stargaze the FD and POLS results

```r
#--- stargazer ---#
stargazer(reg_fd,reg_pols,type='latex',single.row=TRUE,
omit.stat=c('all'),table.layout='-ldc-t-o-n',
column.labels=c('FD','POLS'),omit='Constant',
omit.labels='Intercept?',omit.yes.no=c('Yes','Yes'))
```

Table

	Dependent variable:	
	crmrte	
	FD	POLS
unem	2.218** (0.878)	0.427 (1.188)
d87		7.940 (7.975)
Intercept?	Yes	Yes
Note:	*p<0.1; **p<0.05; ***p<0.01	

Some Drawbacks

- time-constant variables are dropped by first-differencing
- you typically lose variation in explanatory variables (sometimes substantially)
  - residential water price
  - education

Another example

$$log(wage)_{i,1} = \beta_0 + \beta_1 educ_{i,1} + \beta_2 ability_i + u_{i,1}$$
$$log(wage)_{i,2} = \beta_0 + \beta_1 educ_{i,2} + \beta_2 ability_i + u_{i,2}$$

First-differencing

$$\Delta log(wage)_i = \beta_1 \Delta educ_i + \Delta u_i$$

Note

- So, as long as ability is indeed time-invariant, you do not have to worry about correlation between education and ability
- Even if ability is time-variant, it should change very slowly over time. Consequently, most of the correlation between ability and education should be gone.

Program (impact) evaluation with two-period panel data

- Panel data sets are very useful for policy analysis and, in particular, program evaluation
- Note that the important difference from the incinerator example is that the same cross-sectional units appear in each time period

Example: Michigan job training program

- improve the worker productivity of manufacturing firms
- grants were awarded on a first-come, first-served basis
- you are interested in evaluating the impact of this program on products scrap rate

A Model

$$log(scrap)_{i,t} = \beta_0 + \sigma_0 y88_t + \beta_1 grant_{i,t} + \alpha_i + u_{i,t}$$

- $scrap_{i,t}$: products scrap rate for firm i at time t
- $y88_t$: year dummy that takes 1 if in 1988, 0 if in 1987
- $grant_{i,t}$: 1 if you received a grant to have workers participaite in the program at t, 0 otherwise.
- α_i: unobserved time-invariant firm characteristics
- $u_{i,t}$ idiosyncratic error

Question

What would be in α_i?

-
-
-

Question

What would be in α_i?

- employee ability
-
-

Question

What would be in α_i?

- employee ability
- capital
-

Question

What would be in α_i?

- employee ability
- capital
- management skill

Question

What would be in α_i?

- employee ability
- capital
- management skill

These factors should be slow to change in the 2-year period

Question

What would be in α_i?

- employee ability
- capital
- management skill

These factors should be slow to change in the 2-year period

Question

Is $grant_{i,t}$ endogenous? In other words, is $grant_{i,t}$ correlated with any of the above firm characteristics? (pay attention to the way the grants are granted!)

Question

What would be in α_i?

- employee ability
- capital
- management skill

These factors should be slow to change in the 2-year period

Question

Is $grant_{i,t}$ endogenous? In other words, is $grant_{i,t}$ correlated with any of the above firm characteristics? (pay attention to the way the grants are granted!)

Answer

Likely. Why?

A model

$$log(scrap_{i,t}) = \beta_0 + \sigma_0 y88_t + \beta_1 grant_{i,t} + \alpha_i + u_{i,t}$$

two-periods

$$\begin{aligned} t = 2: \quad log(scrap_{i,2}) =& \beta_0 + \sigma_0 y88_t (= 1) \\ & + \beta_1 grant_{i,2} + \alpha_i + u_{i,2} \\ t = 1: \quad log(scrap_{i,1}) =& \beta_0 + \sigma_0 y88_t (= 0) \\ & + \beta_1 grant_{i,1} + \alpha_i + u_{i,1} \end{aligned}$$

First-differencing

$$\Delta log(scrap)_i = \sigma_0 + \beta_1 \Delta grant_i + \Delta u_i$$

R code: POLS and FD Estimation

```r
#--- preparation ---#
data_tr <- readRDS('jtrain.rds') %>% # load the data
  mutate(avgsal=avgsal/1000) %>%
  filter(year!=1989)

#--- POLS ---#
reg_pols <- lm(log(scrap)~grant+d88,data=data_tr)

#--- FD ---#
pdata_tr <- pdata.frame(
  data_tr,
  index=c('fcode','year')
  )
reg_fd <- plm(log(scrap)~grant,data=pdata_tr,model='fd')
```

```r
#--- stargaze ---#
stargazer(reg_fd,reg_pols,type='latex',
  single.row=TRUE,omit.stat=c('all'),table.layout='-ldc-t-o-n',
  column.labels=c('FD','POLS'),omit='Constant',
  omit.labels='Intercept?',omit.yes.no=c('Yes','Yes'))
```

Table

	Dependent variable:	
	log(scrap)	
	FD	POLS
grant	-0.317^{*} (0.164)	0.057 (0.431)
d88		-0.189 (0.328)
Intercept?	Yes	Yes
Note:		*p<0.1; **p<0.05; ***p<0.01

Table

	Dependent variable:	
	log(scrap)	
	FD	POLS
grant	-0.317^{*} (0.164)	0.057 (0.431)
d88		-0.189 (0.328)
Intercept?	Yes	Yes
Note:		$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Bias direction

What seems to be the sign of the bias of not controlling for unobserved firm characteristics?

General Framework

$$y_{i,t} = \beta_0 + \sigma_0 d2_t + \beta_1 prog_{i,t} + \alpha_i + u_{i,t}$$

General Framework

$$y_{i,t} = \beta_0 + \sigma_0 d2_t + \beta_1 prog_{i,t} + \alpha_i + u_{i,t}$$

First-Differencing

$$\Delta y_i = \sigma_0 + \beta_1 \Delta prog_i + \Delta u_i$$

First-Differencing

$$\Delta y_i = \sigma_0 + \beta_1 \Delta prog_i + \Delta u_i$$

Interpretation of β_1

$$E[\Delta y_i | treated] = \sigma_0 + \beta_1$$

$$E[\Delta y_i | control] = \sigma_0$$

So,

$$\beta_1 = E[\Delta y_i | treated] - E[\Delta y_i | control]$$

Interpretation of β_1

$$E[\Delta y_i|treated] = \sigma_0 + \beta_1$$
$$E[\Delta y_i|control] = \sigma_0$$

So,

$$\beta_1 = E[\Delta y_i|treated] - E[\Delta y_i|control]$$

DID

FD estimator is DID for two-period panel data

DID

FD estimator is DID for two-period panel data

Adding other controls

$$\begin{aligned} y_{i,t} =& \beta_0 + \sigma_0 d2_t + \beta_1 prog_{i,t} \\ & + \alpha_1 x1_{i,t} + \cdots + \alpha_k xk_{i,t} + \alpha_i + u_{i,t} \end{aligned}$$

where $x1, \ldots, xk$ are other controls

R code: FD Estimation with other controls

```
  #--- FD ---#
  reg_fd <- plm(log(scrap)~grant+avgsal+tothrs,
    data=pdata_tr,model='fd')

  #--- summary ---#
  summary(reg_fd)$coef
```

```
                 Estimate  Std. Error    t-value  Pr(>|t|)
(Intercept) -0.034625099 0.131343608 -0.2636223 0.7935337
grant       -0.114918942 0.200242877 -0.5738978 0.5695095
avgsal      -0.090541318 0.077611281 -1.1666000 0.2508363
tothrs      -0.005094339 0.004251918 -1.1981271 0.2384861
```

Fixed Effects (FE) and Random Effects (RE) estimation methods

Idea

First-difference is not the only data transformation that eliminate individual fixed effects (time-invariant individual characteristics)

Within-transformation

Consider the following general model:

$$y_{i,t} = \beta_1 x_{i,t} + \alpha_i + u_{i,t}$$

For each i, average this equation over time, we get

$$\frac{\sum_{t=1}^{T} y_{i,t}}{T} = \frac{\sum_{t=1}^{T} x_{i,t}}{T} + \alpha_i \ \ (\frac{\sum_{t=1}^{T} \alpha_i}{T}) + \frac{\sum_{t=1}^{T} u_{i,t}}{T}$$

Subtracting the second equation from the first one,

$$(y_{i,t} - \frac{\sum_{t=1}^{T} y_{i,t}}{T}) = \beta_1(x_{i,t} - \frac{\sum_{t=1}^{T} x_{i,t}}{T}) + (u_{i,t} - \frac{\sum_{t=1}^{T} u_{i,t}}{T})$$

Within-transformation

$$(y_{i,t} - \frac{\sum_{t=1}^{T} y_{i,t}}{T}) = \beta_1(x_{i,t} - \frac{\sum_{t=1}^{T} x_{i,t}}{T}) + (u_{i,t} - \frac{\sum_{t=1}^{T} u_{i,t}}{T})$$

Alternatively, we sometimes write as follows (at least the book does):

$$\ddot{y}_{i,t} = \beta_1 \ddot{x}_{i,t} + \ddot{u}_{i,t}$$

Important

α_i is gone!!

R code: Within-transformed data

```
   id year income educ mean_income mean_educ wt_income wt_educ
1:  1 2015     77   12          81        13        -4      -1
2:  1 2016     82   13          81        13         1       0
3:  1 2017     84   14          81        13         3       1
4:  2 2015    110   18         122        19       -12      -1
5:  2 2016    125   19         122        19         3       0
6:  2 2017    131   20         122        19         9       1
```

Model with multiple variables in general

$$y_{i,t} = \beta_0 + \beta_1 x_{1,i,t} + \beta_2 x_{2,i,t} + \cdots + \beta_k x_{k,i,t} + \alpha_i + u_{i,t}$$

Within-transformed model

$$\ddot{y}_{i,t} = \beta_0 + \beta_1 \ddot{x}_{1,i,t} + \beta_2 \ddot{x}_{2,i,t} + \cdots + \beta_k \ddot{x}_{k,i,t} + \ddot{u}_{i,t}$$

Fixed Effect Estimation

Run OLS on the within-transformed model:

$$\ddot{y}_{i,t} = \beta_0 + \beta_1 \ddot{x}_{1,i,t} + \beta_2 \ddot{x}_{2,i,t} + \cdots + \beta_k \ddot{x}_{k,i,t} + \ddot{u}_{i,t}$$

Fixed Effect Estimation

Run OLS on the within-transformed model:

$$\ddot{y}_{i,t} = \beta_0 + \beta_1 \ddot{x}_{1,i,t} + \beta_2 \ddot{x}_{2,i,t} + \cdots + \beta_k \ddot{x}_{k,i,t} + \ddot{u}_{i,t}$$

Conditions under which FE is unbiased

$$E[\ddot{u}_{i,t} | \ddot{x}_{j,i,t}] = 0 \ \ ^{\forall} j = 1, \ldots, k$$

, which is satisfied if

$$E[u_{i,s} | x_{j,i,t}] = 0 \ \ ^{\forall} s, \ \ t, \ \ \text{and} \ \ j$$

(ex. $E[\ddot{u}_{i,1} | \ddot{x}_{1,i,2}] = 0$)

Conditions under which FE is unbiased

$$E[\ddot{u}_{i,t}|\ddot{x}_{j,i,t}] = 0 \;\; ^{\forall} j = 1, \ldots, k$$

, which is satisfied if

$$E[u_{i,s}|x_{j,i,t}] = 0 \;\; ^{\forall} s, \;\; t, \;\; \text{and} \;\; j$$

(ex. $E[\ddot{u}_{i,1}|\ddot{x}_{1,i,2}] = 0$)

Note

- If an independent variable is a function of the unobserved factors in the previous period, the above condition is violated
- If an independent variable is a function of the dependent variable (, which includes the error term) in the previous period, the above condition is violated

Note

- If an independent variable is a function of the unobserved factors in the previous period, the above condition is violated
- If an independent variable is a function of the dependent variable (, which includes the error term) in the previous period, the above condition is violated

Example

$$crmrte_{i,t} = \beta_0 + \beta_1 lawenf + \alpha_i + u_{i,t}$$

- $crmrte$: crime rate
- $lawenf$: law enforcement efforts

Example

$$crmrte_{i,t} = \beta_0 + \beta_1 lawenf + \alpha_i + u_{i,t}$$

- $crmrte$: crime rate
- $lawenf$: law enforcement efforts

Question

Would $lawenf$ be a function of $crmrte$ in the previous periods?

Implementation of FE estimation in R

Model

$$log(scrap_{i,t}) = \beta_0 + \beta_1 grant_{i,t} + \beta_2 avgsal_{i,t} + \beta_3 tothrs_{i,t} \\ + \beta_4 union_{i,t} + \alpha_i + u_{i,t}$$

R code: FE estimation

```
  #--- create pdata.frame ---#
  pdata_tr <- pdata.frame(data_tr,index=c('fcode','year'))

  #--- FE estimation ---#
  # just tell R that your method is 'within'
  reg_fe <- plm(log(scrap)~grant+avgsal+tothrs,
    data=pdata_tr,model='within')

  #--- summary of the estimation ---#
  summary(reg_fe)$coef

           Estimate  Std. Error    t-value   Pr(>|t|)
grant  -0.130650521 0.188789519 -0.6920433 0.49311737
avgsal -0.104048854 0.057575558 -1.8071706 0.07865804
tothrs -0.005279497 0.004141846 -1.2746726 0.21016467
```

Least square dummy variable (LSDV) regression

- Instead of within-transforming the model, include individual dummy variable and run OLS (We call this least square dummy variable regression).
- It turns out FE estimation is nothing but LSDV (they are mathematically equivalent)

R code: Example

```
   id year income educ fe_1 fe_2
1:  1 2015     77   12    1    0
2:  1 2016     82   13    1    0
3:  1 2017     84   14    1    0
4:  2 2015    110   18    0    1
5:  2 2016    125   19    0    1
6:  2 2017    131   20    0    1
```

FE and LSDV

They will give you identical coefficient and standard error estimates

R code: LSDV vs FE

```
    #--- lsdv ---#
    reg_lsdv_ex <- lm(income˜educ+fe_1+fe_2,data=raw_data)

    #--- fe ---#
    pdata <- pdata.frame(raw_data,index=c('id','year'))
    reg_fe_ex <- plm(income˜educ,data=pdata)
```

R code: stargazer

```r
stargazer(reg_lsdv_ex,reg_fe_ex,type='latex',
  single.row=TRUE,omit.stat=c('all'),table.layout='-ldc-t-o-n',
  column.labels=c('LSDV','FE'),omit='Constant',
  omit.labels='Intercept?',omit.yes.no=c('Yes','Yes'))
```

Table

	Dependent variable:	
	income	
	LSDV	FE
educ	7.000* (2.309)	7.000* (2.309)
fe_1	1.000 (14.360)	
fe_2		
Intercept?	Yes	Yes
Note:	*p$<$0.1; **p$<$0.05; ***p$<$0.01	

FE and LSDV: differences

- LSDV is more computationally demanding because you have so many more variables (individual dummies) included in LSDV
- Estimating LSDV with many individuals was computationally too demanding before
- But, the problem was solved (the $felm$ function from the lfe package)

Implementation of (LSDV) FE estimation in R

$felm$ function from the lfe package

$$felm(y \sim x_1 + \cdots + x_k | fe\_vars | instrument | cluster$$

$$, data = data)$$

- fe_vars: variable(s) that indicates which observations belongs in the same group
- $instrument$: more on this later
- $cluster$: variable that you desire to cluster the error around

Implementation of LSDV (FE) estimation in R

Model

$$log(scrap_{i,t}) = \beta_0 + \beta_1 grant_{i,t} + \beta_2 avgsal_{i,t} + \beta_3 tothrs_{i,t} + \beta_4 union_{i,t} + \alpha_i + u_{i,t}$$

R code: FE estimation

```r
#=== FE estimation ===#
reg_lsdv <- felm(log(scrap)~grant+avgsal+tothrs|fcode|0|0,
  data=data_tr)

#=== summary of the estimation ===#
summary(reg_lsdv)$coef
```

Equivalence of FE and LSDV

Table

	Dependent variable:	
	log(scrap)	
	FE	LSDV
grant	−0.131 (0.189)	−0.131 (0.189)
avgsal	-0.104^{*} (0.058)	-0.104^{*} (0.058)
tothrs	−0.005 (0.004)	−0.005 (0.004)
Note:	*p<0.1; **p<0.05; ***p<0.01	

Fixed Effects (FE) or First-difference (FD)

- both transformations eliminate α_i, which makes your estimation robust to the correlation between α_i and independent variables (This is the most important reason why we use FD and FE)
- FE is more/less efficient that FD depending on the characteristics of the error term (look at the Wooldridge textbook for more detail if you are interested)
- The standard practice is to use FE

Unbalanced Panel

Panel data which do not record the observations for exactly the same periods for all the individuals

R code: stargazer

```
   id year income educ
1:  1 2015     77   12
2:  1 2016     82   13
3:  1 2017     84   14
4:  2 2015    110   18
5:  2 2016     NA   NA
6:  2 2017    131   20
```

Consequence

As long as attrition of the observations (missing observations) is random, there is no consequence to your FE or FD estimation

Random Effects (RE) model

- Can be more efficient than FE and POLS
- Correlation between α_i and independent variables are NOT allowed, meaning RE estimators are biased due to the correlation
- Unless α_i and independent variables are not correlated (which does not hold most of the time unless you got data from controlled experiments), RE is not an attractive option
- You almost never see this methods used in papers that use real world data (non-experimental data)

Hausman Test

Hausman test

$H_0:$ α_i is not correlated with any indepdent variables

$H_1:$ α_i is correlated with at least one of indepdent variables

Hausman test

$H_0 :$ α_i is not correlated with any indepdent variables

$H_1 :$ α_i is correlated with at least one of indepdent variables

Idea

- If α_i is not correlated with any independent variables, both RE and FE are unbiased and produce similar coefficient estimates
- If α_i is correlated with any independent variables, only FE is unbiased, and RE and FE should produce substantially different coefficient estimates

Hausman test

$H_0:\ \alpha_i$ is not correlated with any indepdent variables

$H_1:\ \alpha_i$ is correlated with at least one of indepdent variables

Idea

- If α_i is not correlated with any independent variables, both RE and FE are unbiased and produce similar coefficient estimates
- If α_i is correlated with any independent variables, only FE is unbiased, and RE and FE should produce substantially different coefficient estimates

Procedure

- FE estimation
- RE estimation
- Compare the coefficient estimates between the two

Model

$$log(scrap_{i,t}) = \beta_0 + \beta_1 grant_{i,t} + \beta_2 avgsal_{i,t} + \beta_3 tothrs_{i,t} + \beta_4 union_{i,t} + \alpha_i + u_{i,t}$$

R code: Hausman test

```
  #--- FE ---#
  reg_fe <- plm(log(scrap)˜grant+avgsal+tothrs,
    data=pdata_tr,model='within')

  #--- RE ---#
  # tell R that you want the model to be ''random''
  reg_re <- plm(log(scrap)˜grant+avgsal+tothrs,
    data=pdata_tr,model='random')

  #--- Hausman test ---#
  # phtest() is from the plm package
  phtest(reg_fe, reg_re)


Hausman Test

data:  log(scrap) ˜ grant + avgsal + tothrs
chisq = 3.2055, df = 3, p-value = 0.361
alternative hypothesis: one model is inconsistent
```

Hausman test

- If you reject the null, that means α_i is likely to be correlated with some independent variables $\Rightarrow$ RE estimation is biased
- If you do not reject the null, that means α_i is not likely to be correlated with any of the independent variables $\Rightarrow$ both FE and RE estimations are unbiased

Year Fixed Effects

Year Fixed Effects (FE)

Just a collection of year dummies, which takes 1 if in a specific year, 0 otherwise.

Example

```
   id year income educ FE_2015 FE_2016 FE_2017
1:  1 2015     77   12       1       0       0
2:  1 2016     82   13       0       1       0
3:  1 2017     84   14       0       0       1
4:  2 2015    110   18       1       0       0
5:  2 2016    120   19       0       1       0
6:  2 2017    131   20       0       0       1
7:  3 2015     56   10       1       0       0
8:  3 2016     60   11       0       1       0
9:  3 2017     61   12       0       0       1
```

What do year FEs do?

capture anything that happened to all the individuals for a specific year relative to the base year

What do year FEs do?

capture anything that happened to all the individuals for a specific year relative to the base year

Example

Education and wage data from 2012 to 2014,

$$\begin{aligned} log(income) = & \beta_0 + \beta_1 educ + \beta_2 exper \\ & + \sigma_1 FE_{2012} + \sigma_2 FE_{2013} \end{aligned}$$

- σ_1: captures whatever the difference in $log(income)$ between 2012 and 2014
- σ_2: captures whatever the difference in $log(income)$ between 2013 and 2014

Example

Education and wage data from 2012 to 2014,

$$\begin{aligned} log(income) =& \beta_0 + \beta_1 educ + \beta_2 exper \\ & + \sigma_1 FE_{2012} + \sigma_2 FE_{2013} \end{aligned}$$

- σ_1: captures whatever the difference in $log(income)$ between 2012 and 2014
- σ_2: captures whatever the difference in $log(income)$ between 2013 and 2014

Interpretation

$\sigma_1 = 0.05$ means that $log(income)$ is greater in 2012 than 2014 by 5% on average for whatever reasons with everything else fixed.

Year FE when panel or IPCS datasets are used

It is almost always a good practice to include year FE.

Year FE when panel or IPCS datasets are used

It is almost always a good practice to include year FE.

Why?

- Remember year FEs capture anything that happened to all the individuals for a specific year relative to the base year
- In other words, all the unobserved factors that are common to all the individuals in a specific year is controlled for (taken out of the error term)

Why?

- Remember year FEs capture anything that happened to all the individuals for a specific year relative to the base year
- In other words, all the unobserved factors that are common to all the individuals in a specific year is controlled for (taken out of the error term)

Example

Economic trend in:

$$\begin{aligned} log(income) =& \beta_0 + \beta_1 educ + \beta_2 exper \\ & + \sigma_1 FE_{2012} + \sigma_2 FE_{2013} \end{aligned}$$

- Education is non-decreasing through time
- Economy might have either been going down or up during the observed period

Other Examples

- energy price (as long as everybody's facing the same price each year)
- crop price (as long as everybody's facing the same price each year)

Caveats

- year FEs would be perfectly collinear with a variable that changes only across time, but not across individuals
- if your variable of interest is such a variable, you cannot include year FEs, which would then make your estimation subject to omitted variable bias due to other unobserved yearly-changing factors

Inference using panel data

Heteroskedasticity

Just like we saw for OLS using cross-sectional data, heteroskedasticity leads to biased estimation of the standard error of the coefficient estimators if not taken into account

Heteroskedasticity

Just like we saw for OLS using cross-sectional data, heteroskedasticity leads to biased estimation of the standard error of the coefficient estimators if not taken into account

Serial correlation

Correlattion of errors over time, which we call serial correlation

Heteroskedasticity

Just like we saw for OLS using cross-sectional data, heteroskedasticity leads to biased estimation of the standard error of the coefficient estimators if not taken into account

Serial correlation

Correlattion of errors over time, which we call serial correlation

Consequences of serial correlation

- just like heteroskedasticity, serial correlation could lead to biased estimation of the standard error of the coefficient estimators if not taken into account
- do not affect the unbiasedness and consistency property of your estimators

Important

- Taking into account the potential of serial correlation when estimating the standard error of the coefficient estimators can dramatically change your conclusions about the statistical significance of some independent variables!!
- When serial correlation is ignored, you tend to underestimate the standard error (why?), inflating t-statistic, which in turn leads to over-rejection that you should.

Bertrand, Duflo, and Mullainathan (2004)

- Examined how problematic serial correlation is in terms of inference via Monte Carlo simulation
  - generate a fake treatment dummy variable in a way that it has no impact on the outcome (dependent variable) in the dataset of women's wages from the Current Population Survey (CPS)
  - run regression of the oucome on the treatment variable
  - test if the treatment variable has statistically significant effect via t-test
- They rejected the null 67.5% at the 5% significance level!!

When only heteroskedasticity is present

You can estimate the heteroskedasticity-consistent standard error using the $felm$ function

R code: heteroskedasticity-consistent se estimation when using panel data

```
#--- FE estimation ---#
reg_fe <- felm(log(scrap)~grant+avgsal+tothrs|fcode+year|0|0,
  data=data_tr)

#=== robust se ===#
se_robust <- summary(reg_fe,robust=TRUE)$coef[,2]

#=== non-robust se ===#
se_non_robust <- summary(reg_fe)$coef[,2]
```

R code: stargazer with robust se

```r
#--- stargazer ---#
stargazer(reg_fe,reg_fe,se=list(se_robust,se_non_robust),
  column.labels=c('Het-robust','Non-robust'),type='latex')
```

Table

	Dependent variable:	
	log(scrap)	
	Het-robust	Non-robust
grant	-0.115 (0.202)	-0.115 (0.200)
avgsal	-0.091 (0.070)	-0.091 (0.078)
tothrs	-0.005^{**} (0.003)	-0.005 (0.004)

When both heteroskedasticity and serial correlation are present

- You can take into account both heteroskedasticity and serial correlation by clustering by individual (whatever the unit of individual is: state, county, farmer)
- cluster by individual allows correlation within individuals (over time)

R code: Cluster by individual

```r
#--- FE estimation ---#
reg_fe <- felm(log(scrap)~grant+avgsal+tothrs|fcode+year|0|fcode,
  data=data_tr)

#=== extract se ===#
se_cluster <- summary(reg_fe)$coef[,2]
```

R code: stargazing three regressions results

```
#--- stargazer ---#
stargazer(list(reg_fe,reg_fe),se=list(se_robust,se_cluster),
  type='latex',table.layout='-ldc-t-',
  column.labels=c('Het-robust','Cluster (indiv)'),no.space=TRUE)
```

Table

	Dependent variable:	
	log(scrap)	
	Het-robust	Cluster (indiv)
grant	-0.115	-0.115
	(0.202)	(0.288)
avgsal	-0.091	-0.091
	(0.070)	(0.099)
tothrs	-0.005^{**}	-0.005
	(0.003)	(0.004)