MQE: Economic Inference from Data: Module 2: Fixed Effects

Claire Duquennois

6/9/2020

Module 2: Fixed Effects

- Data Structures
- Fixed Effects

-A simulation

-Fixed effects as demeaned data

-Thinking about variation

-Example: Crime and Unemployment

Controlling for unobservables

We saw with AGG(2006) that even with many covariates, unobservables are a problem.

Certain types of data allow us to control for more of these unobservables by using fixed effects.

Example:

$$Income_i = \beta_0 + \beta_1 Schooling_i + \epsilon$$

β_1 cannot be interpreted as causal: big OVB problems, even with lots of control variables. Unlikely to have good measures of 'ability', 'enthusiasm', 'grit'...

What if I can control for unchanging individual characteristics?

Data Structures: Cross-Section

Individual	Income	Schooling	Female
1	22000	12	1
2	57000	16	1
. . .	. . .	. . .	. . .
N	15000	12	0

Each individual is observed once.

Data Structures: Panel Data

Individual	Income	Schooling	Female	Year
1	22000	12	1	2001
1	23000	12	1	2002
2	57000	16	1	2001
2	63000	17	1	2002
...	...	...	...	...
N	15000	12	0	2001
N	13000	12	0	2002

Each individual is observed multiple times.

Data Structures: Panel Data Subscripts

Unique observations must be identified by both the individual and time dimensions… notice the new subscripts:

$$Income_{it} = \beta_0 + \beta_1 Schooling_{it} + \epsilon.$$

Data Structures: Panel Data

Panel Data can be

-**balanced**: same number of observations for each unit

-**unbalanced**: some units are observed more often then others (probably good to look into why)

Review: Indicator (Dummy) Variables

If I have multiple Female observation and multiple non-female observations I can control for the effect of being female on wages:

$$Income_{it} = \beta_0 + \beta_1 Schooling_{it} + \beta_2 Female_i + \epsilon.$$

Fixed Effects as Individual Indicator Variables

Indiv	Income	School	Female	Year	Indiv1	Indiv2	. . .	IndivN
1	22000	12	1	2007	1	0	0	0
1	23000	12	1	2008	1	0	0	0
2	57000	16	1	2007	0	1	0	0
2	63000	17	1	2008	0	1	0	0
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
N	15000	12	0	2007	0	0	0	1
N	13000	12	0	2008	0	0	0	1

Fixed Effects as Individual Indicator Variables

I can estimate:

$$Inc_{it} = \beta_0 + \beta_1 School_{it} + \beta_2 Fem_i + \beta_{a1} Ind1_i + \beta_{a2} Ind2_i + ... + \beta_{aN-1} Ind(N-1)_i + \epsilon.$$

What do the β_{ak} coefficients tell me?

Also:

-Why do the *IndN* indicators only have an *i* subscript?

-What is the implied assumption if *Fem* only has an *i* subscript?

-Why are there only (N-1) individual dummies?

Fixed Effects as Individual Indicator Variables

What will these individual controls control for?

-β_{a1} will control for the effect of being individual 1 on income that is not explained by that person's gender or schooling.

-Any **time invariant** characteristic that affects individual 1's income, such as ability, grit, enthusiasm... will be controlled for by adding this individual dummy variable.

-These controls are known as individual **fixed effects**.

For notational convenience:

$$Income_{it} = \beta_0 + \beta_1 Schooling_{it} + \beta_2 Female_i + \gamma_i + \epsilon.$$

Fixed Effects

With my panel data, what else can I control for?

$$Income_{it} = \beta_0 + \beta_1 Schooling_{it} + \beta_2 Female_i + \gamma_i + \tau_t + \epsilon.$$

-What is τ_t?

-What is this estimation equivalent to?

A Simulation:

You are a principle of a small school composed of four classrooms. You have just implemented a new option available to teachers for students to spend some small group reading time with a para-educator. You would like to know how this reading time is affecting reading scores.

You have data for ten students in each class that tells you:

-the class the student is in

-whether they participated in small group reading

-their reading score.

Generating Simulated Data

I will work with a simulated dataset to show how the use of fixed effects can help us recover the true treatment effect.

I start by loading the dplyr package and "setting the seed":

```
#install.packages("dplyr")
#install.packages("lfe")
#install.packages("stargazer")

library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.6.3

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(lfe)
```

```
## Loading required package: Matrix
```

```
library(stargazer)
```

```
##
## Please cite as:
```

A Simulation:

I generate a vector of class identifiers and a random error term.

```r
class<-c(1,2,3,4)
scores<-as.data.frame(class)
scores<-rbind(scores,scores,scores,scores,scores,scores,scores,scores,scores,scores)
scores$error<-rnorm(40, mean=0, sd=5)

#note: if you are not working in markdown you would just write head(scores)
knitr::kable(head(scores))
```

class	error
1	3.6633624
2	-0.1891486
3	6.0150457
4	7.3490101
1	0.6684515
2	2.5991362

A Simulation:

I simulate some selection into treatment. The probability of getting treated is

-0.8 for students in classrooms 3 and 4

-0.2 in classrooms 1 and 2.

```r
scores$treat1<-rbinom(40,1,0.2)
scores$treat2<-rbinom(40,1,0.8)
scores$treat[scores$class%in%c(1,2)]<-scores$treat1[scores$class%in%c(1,2)]
scores$treat[scores$class%in%c(3,4)]<-scores$treat2[scores$class%in%c(3,4)]

knitr::kable(head(scores))
```

class	error	treat1	treat2	treat
1	3.6633624	0	1	0
2	-0.1891486	0	1	0
3	6.0150457	0	1	1
4	7.3490101	1	0	0
1	0.6684515	0	1	0
2	2.5991362	0	1	0

A Simulation:

I drop unneeded variables and generate a dummy variable for each classroom

```
scores<-scores%>%select(class,error,treat)
scores <- fastDummies::dummy_cols(scores, select_columns = "class")

knitr::kable(head(scores))
```

class	error	treat	class_1	class_2	class_3	class_4
1	3.6633624	0	1	0	0	0
2	-0.1891486	0	0	1	0	0
3	6.0150457	1	0	0	1	0
4	7.3490101	0	0	0	0	1
1	0.6684515	0	1	0	0	0
2	2.5991362	0	0	1	0	0

A Simulation:

Finally! I simulate the DGP (Data Generating Process):

-The true treatment effect = 15

-students in classrooms 1 and 2 have higher reading scores

-students in classrooms 3 and 4 have lower reading scores.

```
scores$score<-80+15*scores$treat+10*scores$class_2+-30*scores$class_3+
  -35*scores$class_4+scores$error

knitr::kable(head(scores))
```

class	error	treat	class_1	class_2	class_3	class_4	score
1	3.6633624	0	1	0	0	0	83.66336
2	-0.1891486	0	0	1	0	0	89.81085
3	6.0150457	1	0	0	1	0	71.01505
4	7.3490101	0	0	0	0	1	52.34901
1	0.6684515	0	1	0	0	0	80.66845
2	2.5991362	0	0	1	0	0	92.59914

A Simulation:

I estimate three specifications. The first:

$$Score_{ci} = \beta_0 + \beta_1 Treat_{ci} + \epsilon$$

```
nofe<-felm(score~treat,scores)
```

A Simulation:

The second:

$$Score_{ci} = \beta_0 + \beta_1 Treat_{ci} + \beta_2 Class2_c + \beta_3 Class3_c + \beta_4 Class4_c + \epsilon$$

```r
dummies<-felm(score~treat+class_2+class_3+class_4, scores)
```

A Simulation:

The third:

$$\mathit{Score}_{ci} = \beta_0 + \beta_1 \mathit{Treat}_{ci} + \kappa_c + \epsilon$$

where κ_i is a classroom fixed effect.

```r
fe<-felm(score~treat|class,scores)
```

A Simulation:

```r
stargazer(nofe, dummies, fe,header=FALSE, type='latex')
```

Table 8

	Dependent variable:		
	score		
	(1)	(2)	(3)
treat	2.280	15.998***	15.998***
	(5.822)	(1.771)	(1.771)
class_2		11.678***	
		(2.275)	
class_3		$-$27.916***	
		(2.435)	
class_4		$-$32.332***	
		(2.376)	
Constant	72.557***	78.527***	
	(3.905)	(1.643)	
Observations	40	40	40
R^2	0.004	0.930	0.930
Adjusted R^2	$-$0.022	0.922	0.922
Residual Std. Error	18.318 (df = 38)	5.072 (df = 35)	5.072 (df = 35)
Note:		*p<0.1; **p<0.05; ***p<0.01	

A Simulation:

Recall: $\beta_1 = 15$ (the true treatment effect) $\Rightarrow \hat{\beta}_1^{nofe}$ is very biased!

Why?

-The classes are an important omitted variable: $cor(Score, Class3/4) < 0$ and $cor(Treat, Class3/4) > 0$ creating substantial downward bias.

We can correct for this in two (equivalent) ways:

-adding the dummy variables for the class to the regression,

-adding a class fixed effect.

Either approach returns an identical unbiased estimate such that $E[\hat{\beta}_1] = \beta_1$.

Fixed Effects as Demeaned Data:

Fixed effect estimates are also known as the **within estimator**, because it identifies β using within-unit variation.

$\Rightarrow$ we only using the variation that exists **within the classroom** to estimate the treatment effect.

This is the equivalent of “correcting” our data by demeaning each observation using it’s classroom mean, so that the corrected data represents deviations from the classroom mean.

Fixed Effects as Demeaned Data:

Our fixed effect estimation is

$$y_{ci} = \beta_1 x_{ci} + \kappa_c + \epsilon_{ci}$$

For each class, the average across the students is

$$\bar{y}_i = \beta_1 \bar{x}_i + \kappa_c + \bar{\epsilon}_i$$

Subtracting this from the fixed effect model gives

$$y_{ic} - \bar{y}_i = \beta_1 (x_{ic} - \bar{x}_i) + (\epsilon_{ic} - \bar{\epsilon}_i)$$