EC 3303: Econometrics I

Fixed Effects

Kelvin Seah

AY 2022/2023, Semester 2

Outline

1. Fixed Effects

2. Differencing

Limitations of Regression

- Multiple regression is a powerful tool to estimate causal effects if we can observe ***all*** the variables which are correlated with both the regressor of interest X_i & the dependent variable Y_i.
  - Including these variables as controls will guarantee that the estimator of the effect of interest is unbiased.
  - But in practice, data may not be available on these variables.

Fixed Effects Regression

- If the ***unobserved variables***:
  - vary from entity to entity but are constant over time
  - can use regression with ***fixed effects*** to obtain an unbiased estimator of the causal effect of interest.
- Fixed effects regression requires panel data.
  - Data on multiple entities in which each entity is observed at two or more time periods.
  - Balanced panel: has no missing observations. Values of variables are all observed for each entity and each time period.
  - Unbalanced panel: some observations are missing. There is missing data for at least one time period for at least one entity.

Does being taught by a native teacher affect student achievement?

- In the U.S., students taught by native teachers typically score higher on achievement tests.

- Is this because of the nativity of the teacher, or because students taught by native teachers & students taught by immigrant teachers are dissimilar?

- Assuming that teacher nativity is as good as randomly assigned, *conditional on* student gender, race, family income, parental background, home language, family size, teacher education & experience, & student motivation

- & that the causal effect of teacher nativity is additive & constant (=ρ for each student i), then:

A regression of testscore on teacher nativity with these variables as controls will give an *unbiased* estimator of the causal effect of teacher nativity

$$Y_i = \alpha + \rho Native_i + \boldsymbol{X'}_i \boldsymbol{\beta} + u_i \quad (1)$$

where

Y_i: testscore of student *i*

$Native_i$: dummy for teacher nativity (=1 if teacher of student *i* is native, =0 otherwise)

$\boldsymbol{X}_i$: vector of control variables: student gender, race, family income, parental background, home language, family size, & teacher education & experience

u_i: other factors influencing student *i*'s testscore

- In practice, researchers cannot observe student motivation.
    - So cannot include "student motivation" as a control variable.
- Running equation (1) without "motivation" as a control variable will lead to OVB in the OLS estimator of ρ.
- The model actually estimated is

$$Y_i = \alpha + \rho Native_i + \boldsymbol{X'}_i \boldsymbol{\beta} + [\delta Motiv_i + error_i]$$

    - since motivation levels of students taught by native teachers may differ from those taught by non-native teachers

$$corr(Motiv, Native) \neq 0$$

    - and since motivation level is a determinant of testscore Y_i, so

Conditional mean independence is violated.

However if,

1. motivation remains *constant over time* for a given student &

2. panel data is available, then

we can control for motivation even though we cannot measure it.

$$Y_{it} = \alpha + \rho Native_{it} + \boldsymbol{X'}_{it}\boldsymbol{\beta} + \delta Z_i + u_{it} \quad (2)$$

where i indexes student $i = 1, \ldots, n$; while t indexes time-period $t = 1, \ldots, T$.

- Let Z_i be a variable which determines the testscore of student i, but that does not change over time (e.g. motivation level)
  - Z_i only varies across students i, but does not vary over time – so it will not produce any change in Y between the time periods considered.

- Since we cannot observe Z_i, we control for its effect by *eliminating* it from the model:

- Suppose we observe students for $T = 2$ time periods (say 2010 & 2015), then

$$Y_{i,2015} = \alpha + \rho Native_{i,2015} + \boldsymbol{X'}_{i,2015}\boldsymbol{\beta} + \delta Z_i + u_{i,2015} \quad (3)$$

$$Y_{i,2010} = \alpha + \rho Native_{i,2010} + \boldsymbol{X'}_{i,2010}\boldsymbol{\beta} + \delta Z_i + u_{i,2010} \quad (4)$$

$(3) - (4)$:

$$Y_{i,2015} - Y_{i,2010} = \rho\left[Native_{i,2015} - Native_{i,2010}\right] + \left[\boldsymbol{X'}_{i,2015} - \boldsymbol{X'}_{i,2010}\right]\boldsymbol{\beta} + u_{i,2015} - u_{i,2010} \quad (5)$$

- Removing Z_i this way is called "*differencing*".

Differencing

- We can create a differenced variable like $(Y_{i,2015} - Y_{i,2010})$ in Stata by generating a new variable equal to the difference in the testscore variables for the years 2015 & 2010.

- Intercept is usually included

$$\widehat{Y_{i,2015} - Y_{i,2010}} = \widehat{\alpha_1} + \hat{\rho}\left[Native_{i,2015} - Native_{i,2010}\right] + \left[\boldsymbol{X'}_{i,2015} - \boldsymbol{X'}_{i,2010}\right]\widehat{\boldsymbol{\beta}} \qquad (6)$$

  - to allow for the possibility that the mean change in testscore (over the two years for the entities) is non-zero, even if there is no change in any of the included regressors (over the two years for the entities).

Running Differenced Regressions

- National Longitudinal Survey of Youth.
  - Subset of 581 teenagers
  - Interviewed in 1990, 1992, 1994
  - Numbers at the end of variable names reflect time period in which the variable was measured
  - Variables without numbers at the end do not vary over time.
  - Consider variables:
    - id: subject id, same for each teenager across every wave
    - anti: measure for antisocial behavior (0-6)
    - pov: 1 if family in poverty; 0 if not
    - self: self esteem (6-24)
    - and later, gender: 1 if female, 0 if male

- set more off
- use http://www3.nd.edu/~rwilliam/statafiles/nlsy.dta, clear
- des anti* self* pov* gender

```
. des anti* self* pov* gender

              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
anti90          byte   %8.0g                  child antisocial behavior in 1990
anti92          byte   %8.0g                  child antisocial behavior in 1992
anti94          byte   %8.0g                  child antisocial behavior in 1994
self90          byte   %8.0g                  child self-esteem in 1990
self92          float  %9.0g                  child self-esteem in 1992
self94          byte   %8.0g                  child self-esteem in 1994
pov90           byte   %8.0g                  family poverty status in 1990
pov92           byte   %8.0g                  family poverty status in 1992
pov94           byte   %8.0g
gender          byte   %8.0g                  child's gender
```

- keep anti* self* pov* gender

- Ignore observations from 1994 so we can use differencing
- Create unique ID for each entity (girl)
  - `gen id=_n`
- Generate differenced variables
  - `gen d_anti= anti92 - anti90`
  - `gen d_pov = pov92 - pov90`
  - `gen d_self = self92 - self90`

```
. list in 1/5
```

	anti90	anti92	anti94	gender	self90	self92	self94	pov90	pov92	pov94	d_anti	d_pov	d_self	id
1.	1	1	1	1	21	24	23	1	1	1	0	0	3	1
2.	0	0	0	1	20	24	24	0	0	0	0	0	4	2
3.	5	5	5	0	21	24	24	0	0	0	0	0	3	3
4.	2	3	1	0	23	21	21	0	0	0	1	0	-2	4
5.	1	0	0	1	22	23	24	0	0	0	-1	0	1	5

- Examine how antisocial behavior is associated with poverty and self-esteem.
  - `reg d_anti d_pov d_self, cluster(id)`

```
. reg d_anti d_pov d_self, cluster(id)

Linear regression                                      Number of obs =     581
                                                       F(  2,   580) =    3.94
                                                       Prob > F      =  0.0200
                                                       R-squared     =  0.0169
                                                       Root MSE      =  1.2829

                                   (Std. Err. adjusted for 581 clusters in id)
```

d_anti	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
d_pov	.1969039	.145378	1.35	0.176	-.0886277	.4824355
d_self	-.0391292	.014996	-2.61	0.009	-.0685823	-.0096762
_cons	.0403031	.0539599	0.75	0.455	-.0656775	.1462837

- data is in "wide" format here

	anti90	anti92	anti94	gender	self90	self92	self94	pov90	pov92	pov94	d_anti	d_pov	d_self	id
1.	1	1	1	1	21	24	23	1	1	1	0	0	3	1
2.	0	0	0	1	20	24	24	0	0	0	0	0	4	2
3.	5	5	5	0	21	24	24	0	0	0	0	0	3	3
4.	2	3	1	0	23	21	21	0	0	0	1	0	-2	4
5.	1	0	0	1	22	23	24	0	0	0	-1	0	1	5

- We can also difference variables when we structure the dataset from “wide” to long”.
  - In “long” format, each row contains variables for an entity measured at a single time period.
  - Typical “panel form”.

- Convert the dataset from “wide” to “long” using “reshape” command.
  - `reshape long anti pov self, i(id) j(year)`
  - The variable list following the reshape command is the list of variables that varies across time. Since these variables are measured over 3 years (1990, 1992, 1994), STATA creates 3 records (rows) for each entity.
  - The option `j(year)` creates a year variable. option `i(id)` tells STATA what the entity is.

reshaped data:

```
. list in 1/21
```

	id	year	anti	gender	self	pov
1.	1	90	1	1	21	1
2.	1	92	1	1	24	1
3.	1	94	1	1	23	1
4.	2	90	0	1	20	0
5.	2	92	0	1	24	0
6.	2	94	0	1	24	0
7.	3	90	5	0	21	0
8.	3	92	5	0	24	0
9.	3	94	5	0	24	0
10.	4	90	2	0	23	0
11.	4	92	3	0	21	0
12.	4	94	1	0	21	0
13.	5	90	1	1	22	0
14.	5	92	0	1	23	0
15.	5	94	0	1	24	0
16.	6	90	1	0	19	0
17.	6	92	1	0	21	0
18.	6	94	1	0	24	0
19.	7	90	3	1	24	0
20.	7	92	3	1	16	0
21.	7	94	4	1	13	0

- Generate differenced variables
  - `drop if year >=94` *do this so that we only have 2 years of data
  - `gen d_anti= anti - anti[_n-1] if year==92`
  - `gen d_pov = pov - pov[_n-1] if year==92`
  - `gen d_self = self - self[_n-1] if year==92`
- Run the differenced regression
  - `reg d_anti d_pov d_self, cluster(id)`

```
. reg d_anti d_pov d_self, cluster(id)

Linear regression                                      Number of obs =     581
                                                       F(  2,   580) =    3.94
                                                       Prob > F      =  0.0200
                                                       R-squared     =  0.0169
                                                       Root MSE      =  1.2829

                                   (Std. Err. adjusted for 581 clusters in id)
------------------------------------------------------------------------------
             |               Robust
      d_anti |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       d_pov |   .1969039    .145378     1.35   0.176    -.0886277    .4824355
      d_self |  -.0391292    .014996    -2.61   0.009    -.0685823   -.0096762
       _cons |   .0403031   .0539599     0.75   0.455    -.0656775    .1462837
------------------------------------------------------------------------------
```

Fixed Effects Regression

- Differencing works when entities are observed for only *2 different time periods*.

- If $T > 2$, we use the more general *fixed effects regression*.

- Consider again:

$$Y_{it} = \alpha + \rho Native_{it} + \boldsymbol{X'}_{it}\boldsymbol{\beta} + \delta Z_i + u_{it} \quad (7)$$

- Can get rid of Z_i by entity-demeaning.

Step 1: average each variable over all *t* for each *i*

$$\bar{Y}_i = \alpha + \rho\overline{Native}_i + \bar{\boldsymbol{X}}'_i\boldsymbol{\beta} + \delta Z_i + \bar{u}_i \quad (8)$$

Where $\bar{Y}_i = \frac{1}{T}\sum_{t=1}^{T} Y_{it}$, other variables defined similarly

Step 2: (7)-(8)

$$(Y_{it} - \bar{Y}_i) = \rho(Native_{it} - \overline{Native}_i) + (\boldsymbol{X}'_{it} - \bar{\boldsymbol{X}}'_i)\boldsymbol{\beta} + (u_{it} - \bar{u}_i) \ (9)$$

- This transformation yields “entity-demeaned” variables

- $(Y_{it} - \bar{Y}_i) = \rho(Native_{it} - \overline{Native}_i) + (\boldsymbol{X'}_{it} - \boldsymbol{\bar{X}'}_i)\boldsymbol{\beta} + (u_{it} - \bar{u}_i)$ (9)

- Can rewrite

$\tilde{Y}_{it} = (Y_{it} - \bar{Y}_i)$; $\widetilde{Native}_{it} = (Native_{it} - \overline{Native}_i)$; $\tilde{X}'_{it} = (\boldsymbol{X'}_{it} - \boldsymbol{\bar{X}'}_i)$
$\tilde{u}_{it} = (u_{it} - \bar{u}_i)$

- Can then run a regression of $\tilde{Y}_{it}$ on $\widetilde{Native}$ and $\tilde{X}'$ to get an unbiased estimator of ρ.

- This is equivalent to:

$$Y_{it} = \alpha + \rho Native_{it} + \boldsymbol{X'}_{it}\boldsymbol{\beta} + \gamma_2 W2_i + \gamma_3 W3_i + \cdots + \gamma_n Wn_i + u_{it} \quad (10)$$

where

$W2_i$ is a dummy variable (=1 if *i* is student 2, = 0 otherwise),

$W3_i$ is a dummy variable (=1 if *i* is student 3, = 0 otherwise)….

- So there are *n-1 dummy variables*, each indicating an individual student.

- This is known as regression with student (entity) fixed effects.

How to run FE regressions: xtreg

- To use "`xtreg`", need to have the dataset in "long" format & declare the data to be a panel.
  - `xtset id year`
  - The general format is `xtset panelvar timevar`
    - "`xtset`" tells STATA to treat id as the entity / panel variable & year as the time variable.

- Run the demeaned regression
  - `xtreg anti pov self, fe cluster(id)`

* `fe` tells STATA to run a fixed effects model. `cluster(id)` tells STATA to cluster standard errors by entity.

```
. xtreg anti pov self, fe cluster(id)

Fixed-effects (within) regression               Number of obs      =      1743
Group variable: id                              Number of groups   =       581

R-sq:  within  = 0.0212                         Obs per group: min =         3
       between = 0.0418                                        avg =       3.0
       overall = 0.0327                                        max =         3

                                                F(2,580)           =     11.07
corr(u_i, Xb)  = 0.0789                         Prob > F           =    0.0000

                                   (Std. Err. adjusted for 581 clusters in id)
------------------------------------------------------------------------------
             |               Robust
        anti |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         pov |   .1048989   .0992739     1.06   0.291    -.0900813    .2998791
        self |  -.0514953   .0113174    -4.55   0.000    -.0737234   -.0292672
       _cons |   2.650289   .2336169    11.34   0.000      2.19145    3.109127
-------------+----------------------------------------------------------------
     sigma_u |  1.3228744
     sigma_e |  1.0023447
         rho |  .63527864   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Standard errors should be clustered

- LSA #2: (X_i, Y_i), $i = 1, \ldots, n$ are independently & identically distributed (i.i.d).
- arises if the entity is randomly selected by simple random sampling.
- However, for panel data, observations are not i.i.d.; Y, X, u tend to be correlated over time for a given entity ("autocorrelated").
- To make standard errors valid, we cluster standard errors by entity so as to allow for correlation of regression errors within an entity.

More Practice

- use fatality.dta   /* Load data */

/***

* (1) running first differenced regressions

***/

- * Regress fatality rate in 1982 on real beer tax in 1982, using OLS. What is the coefficient on real beer tax?
- regress fatalityrate beertax if year==1982, robust
- * Regress fatality rate in 1988 on real beer tax in 1988, using OLS. What is the coefficient on real beer tax?
- regress fatalityrate beertax if year==1988, robust
- * Here is how you create the first differenced variables using data from 1988 and 1982.

- preserve
- gsort state -year /*Sort dataset by ascending values of state and descending values of year*/
- keep if year==1988 | year==1982 /*keep only data from 1982 and 1988*/
- bysort state: generate d_fatalityrate= fatalityrate[_n] - fatalityrate[_n-1] /*create first differenced fatality rate variable*/
- bysort state: generate d_beertax= beertax[_n] - beertax[_n-1] /*create first differenced beer tax variable*/
- regress d_fatalityrate d_beertax, cluster(state) /*run first differenced regression. We cluster standard errors by state so as to allow for correlation of regression errors within a state*/
- restore

- /***
- * running fixed effects regressions
- ***/
- preserve
- xtset state year /*declare the dataset to be a panel. xtset tells Stata to treat state as the entity variable & year as the time variable*/
- xtreg fatalityrate beertax, fe cluster(state) /*fe tells STATA to run a fixed effects regression. cluster(state) tells Stata to cluster standard errors by state*/
- restore