ECON-C4200 - Econometrics II: Capstone

Lecture 1: Panel data

Otto Toivanen

Teachers

Professor Otto Toivanen

Economics dept. Aalto U.

otto.toivanen@aalto.fi

Office hours: on appointment.

TAs Antto Jokelainen & Tuomas Markkula

Economics dept. Aalto U.

antto.jokelainen@aalto.fi
tuomas.markkula@aalto.fi

Office hours: on appointment.

What Econometrics I was about

- Tools: economic theory + statistical tools + data + knowledge. In short: econometrics.
- Learning outcomes: Students
  1. are acquainted with the principles of empirical methods in economics.
  2. know how to perform descriptive analysis of data.
  3. are acquainted with econometrics methods for cross-section data.
  4. understand the difference between descriptive and causal analysis.
  5. have basic knowledge of the econometrics software package Stata.
  6. know the basics of how to program, how to document and how to ensure replicability of their econometric analysis.

What this course is about

- Learning outcomes: Students
  1. understand the benefits of panel data and how to make use of them
  2. are familiar with Difference-in-Difference analysis and its basic use
  3. know how to model limited dependent variables
  4. have basic knowledge of the time series econometrics, including forecasting models
  5. have basic knowledge of the VAR (**V**ector **A**uto**R**egressive) models
  6. understand what cointegration is
  7. have a basic knowledge of (G)ARCH (**G**eneralized **A**uto**R**egressive **C**onditional **H**eteroskedasticity) models and their use.

Course evaluation

- Exercises 20%
- Capstone 35%
- Exam 45%
  - Course exam 17.4.2023
  - Retake exam 2.6.2023

Lectures – current plan

28.2 Lecture 1 panel data #1, ch10

2.3 Lecture 2 panel data #2, ch10

7.3 Lecture 3 causal parameters #3.1: Difference-in-Difference, ch10

9.3 Lecture 4 causal parameters #3.2: Difference-in-Difference, examples

Lectures – current plan

14.3 Lecture 5 limited dependent variables #1, ch11

16.3 Lecture 6 limited dependent variables #2, ch11

21.3 Lecture 7 Econometrics and machine learning, ch14 ($4^t h$ ed.)

23.3 Lecture 8 applications (TBD)

Lectures – current plan

28.3 Lecture 9 time series #1: forecasting ch14 and #2: dynamic causal effects, ch15

30.3 Lecture 10 time series #2: dynamic causal effects, ch15 and #3: VAR models, ch16

4.4 Lecture 11 time series #3: VAR models, ch16 and #4: Cointegration & ARCH models, ch16

13.4 Lecture 12 recap

Exercises and Problem Sets

- 5 graded problem sets and 5 exercise sessions.
- 4 problem sets with the highest grades count towards your final grade.
- Problem sets are published a week before the deadline on Thursday. All deadlines are on Thursday (18:00 EET) a day before the exercise session.
- Problem sets have equal weight and include both analytical and empirical problems.
- You need at least 50% of points to pass the course.

Exercises and Problem Sets

- Deadlines are strict - do not email us your solutions.
- Plagiarism is strictly forbidden. Do not share your answers or code. You can discuss the exercises in small groups but all answers must be self-written.
- Detailed instructions are found on MyCourses.

Exercises and Problem Sets

Problem Set 1 - 10.03. Panel data

Problem Set 2 - 17.03. DiD

Problem Set 3 - 24.03. LDV

Problem Set 4 - 31.03. Time series

Problem Set 5 - 14.04. Time series

Panel data

Learning outcomes

- At the end of lectures 1 & 2, you

1. understand what panel data is
2. how a first-difference estimator works
3. how a least squares dummy variable estimator works
4. how a fixed effects estimator works.
5. how a random effects estimator works.
6. how to think about measurement error in a panel data context
7. why there could a need to cluster standard errors.

1. Cross-section data

- Many observation units.
- Each observed just once.
- Examples:
  1. Student grades in the n^{th} year of studies.
  2. Customer decision(s) during a single shopping trip.
  3. Firm's bids in a procurement auction.

2. Time-series data

- Same phenomenon for the same unit observed many times at different points in time.
- Examples:
    1. Inflation at the monthly level for a country.
    2. Stock market index by minute during a day.
    3. Electricity prices at 12.00 for 400 days in a row.

3. Panel data

- Observe same units several times.
- Examples:
  1. Individuals annual income and jobs for t years in the Finnish job market.
  2. Finnish firms' accounting information since 2000.
  3. Prices and sold quantities for each car type on sale in Finland 2000 - 2015.
  4. Our FLEED data.

Panel data

- Formally, one observes $Y_{it}, \boldsymbol{X}_{it}$ for
- units $i = 1, ..., n$ and
- periods $t = 1, ..., T$
- NOTE: there can be more than two dimensions, e.g., individuals, regions, time.

Panel data - Balanced vs. unbalanced

- Panel data is **balanced** if all units are observed for the same time periods.
- Panel data is **unbalanced** if this is not the case.
- Examples:
  1. Firm panel data unbalanced because firms are born and die.
  2. Customer panel data unbalanced because customers appear and disappear.

What does panel data bring to the table?

- In a cross-section, the only source of variation is across observation units.
- In time-series, the only source of variation is changes over time.
- Panel data combines these.
- FLEED: income, age and education observed for same individuals over many years.

What does panel data bring to the table?

- Consider the univariate regression

$$Y_{it} = \alpha_0 + \beta_1 X_{it} + \epsilon_{it}$$

Notice we now need also a t - index.

What does panel data bring to the table?

$$Y_{it} = \alpha_0 + \beta_1 X_{it} + u_{it}$$

- With enough time-series data, you could estimate this separately for each observation unit.

$$Y_{it} = \beta_{0i} + \beta_{1i} X_{it} + \epsilon_{it}$$

What does panel data bring to the table?

$$Y_{iy} = \alpha_0 + \beta_1 X_{it} + u_{it}$$

- With enough observation units, you could estimate this separately for each time period.

$$Y_{it} = \alpha_{0t} + \beta_{1t} X_{it} + \epsilon_{it}$$

What does panel data bring to the table?

$$Y_{it} = \alpha_0 + \beta_1 X_{it} + \epsilon_{it}$$

- Or you could decide on some combination.
- Why? To **reduce bias** & **increase precision** of your parameter estimates.
- Is there any reason to think the effect of X on Y varies over time?
- Is there reason to think the effect of X on Y varies across observation units?

What does panel data bring to the table?

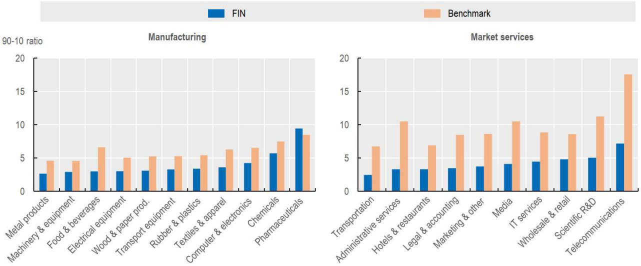
- The panel data estimator

$$Y_{it} = \alpha_{0i} + \beta_1 X_{it} + \epsilon_{it}$$

- Example: Effect of R&D (=X) on productivity (=Y).
- What is the interpretation of α_{0i}?
- Firms have different productivity levels even when they invest the same amount in R&D.

Productivity Dispersion, source: OECD. (2020). *Insights on productivity and business dynamics*.

Figure 4. Average dispersion of productivity within industries
Manufacturing and non-financial market services
Finland vs benchmark countries, 2001-12



Note: This figure reports the average dispersion in labour productivity within industries in Finland and within country-industry pairs in a set of benchmark countries. Dispersion is measured as the ratio of the 90th percentile to the 10th percentile of the firm productivity distribution. Results are presented separately for manufacturing and non-financial market services based on detailed industries, following the SNA A38 industry classification. See Box 1 for details.

What does panel data bring to the table?

- The panel data estimator

$$Y_{it} = \alpha_{0i} + \beta_1 X_{it} + \epsilon_{it}$$

- It is natural to see the panel data estimators as generalizations of the cross-section regression that you would (have) run.
- Key question: how to model α_{0i}?

4. General set-up

- Consider the following model:

$$Y_{it} = \alpha_i + \boldsymbol{X}_{it}'\beta + \epsilon_{it}$$

where α_i is a **time invariant individual effect**.

- Written in matrix form:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix} = \begin{bmatrix} \mathrm{i} & 0 & \dots & 0 \\ 0 & \mathrm{i} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathrm{i} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_N \end{bmatrix} + \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_N \end{bmatrix} \beta + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix}$$

General set-up

- Y_{it} and $\boldsymbol{X}_{it}$ are the T time observations on the outcome and on the K explanatory factors for observation unit i in period t.

General set-up

- Y_{it} and $\boldsymbol{X}_{it}$ are the T time observations on the outcome and on the K explanatory factors for observation unit i in period t.
- β is the column vector of K parameters.

General set-up

- Y_{it} and $\boldsymbol{X}_{it}$ are the T time observations on the outcome and on the K explanatory factors for observation unit i in period t.
- β is the column vector of K parameters.
- α_i is the time invariant individual effect.

General set-up

- Y_{it} and $\boldsymbol{X}_{it}$ are the T time observations on the outcome and on the K explanatory factors for observation unit i in period t.
- β is the column vector of K parameters.
- α_i is the time invariant individual effect.
- ϵ_{it} is the vector T disturbances for observation unit i.

General set-up

- Y_{it} and $\boldsymbol{X}_{it}$ are the T time observations on the outcome and on the K explanatory factors for observation unit i in period t.
- β is the column vector of K parameters.
- α_i is the time invariant individual effect.
- ϵ_{it} is the vector T disturbances for observation unit i.
- i is a T dimensional column vector with all elements equal to 1.

General set-up

- Y_{it} and $\boldsymbol{X}_{it}$ are the T time observations on the outcome and on the K explanatory factors for observation unit i in period t.
- β is the column vector of K parameters.
- α_i is the time invariant individual effect.
- ϵ_{it} is the vector T disturbances for observation unit i.
- i is a T dimensional column vector with all elements equal to 1.
- We are interested in β.

General set-up

- α_i is the time invariant individual effect. It is also called the

General set-up

- α_i is the time invariant individual effect. It is also called the
- **the unobserved component**,

General set-up

- α_i is the time invariant individual effect. It is also called the
- **the unobserved component**,
- **latent variable**,

General set-up

- α_i is the time invariant individual effect. It is also called the
- **the unobserved component**,
- **latent variable**,
- **individual** or **unobserved heterogeneity**.

5. Different estimators

- **(First) difference** - estimator.
- **Least Squares Dummy Variable** (LSDV) estimator.
- **Fixed Effects** (FE) estimator.
- **Random Effects** (RE) estimator.

5.1 First-difference estimator: 2-period example

- Imagine you observe customers in 2 time periods and know how much advertising they are subjected to.
- You are interested in the amount of sales that ads generate.
- For simplicity, let's assume you have randomized the ads.
- Let's denote quantity bought by customer i in period t by q_{it}, and the amount of advertising the customer is subjected to by a_{it}.

2-period example

- α_{0i} disappear.

$\rightarrow$ they could be correlated with u_{it}.

- Note what variation ("within variation") is left to identify the parameters.

$\rightarrow$ Needed: changes w/in an observation unit in both X and Y.

2-period example

- If no variation left, then "everything" explained by α_{0i}.
- Famous example: Firm level R&D.
- Potential problem: dummy variables.

Table: example of within-variation from FLEED

shtun	year	age	high_educ
41	11	21	0
41	12	22	0
41	13	23	0
41	14	24	0
41	15	25	0
42	1	22	
42	2	23	
42	3	24	0
42	4	25	0
42	5	26	0
42	6	27	0
42	7	28	0
42	8	29	0
42	9	30	0
42	10	31	0
42	11	32	0
42	12	33	0
42	13	34	0
42	14	35	1
42	15	36	1

1

Table

shtun	year	age	high_educ
41	11	21	0
41	12	22	0
41	13	23	0
41	14	24	0
41	15	25	0
42	1	22	
42	2	23	
42	3	24	0
42	4	25	0
42	5	26	0
42	6	27	0
42	7	28	0
42	8	29	0
42	9	30	0
42	10	31	0
42	11	32	0
42	12	33	0
42	13	34	0
42	14	35	1
42	15	36	1

2

Table

shtun	year	age	high_educ
41	11	21	0
41	12	22	0
41	13	23	0
41	14	24	0
41	15	25	0
42	1	22	
42	2	23	
42	3	24	0
42	4	25	0
42	5	26	0
42	6	27	0
42	7	28	0
42	8	29	0
42	9	30	0
42	10	31	0
42	11	32	0
42	12	33	0
42	13	34	0
42	14	35	1
42	15	36	1

3

Table

```
. sum high_educ dhigh_educ
```

Variable	Obs	Mean	Std. Dev.
high_educ	53,938	.0727131	.2596674
dhigh_educ	48,992	.0051845	.0718174

```
. tab dhigh_educ if e(sample)
```

dhigh_educ	Freq.	Percent	Cum.
0	47,497	99.48	99.48
1	249	0.52	100.00
Total	47,746	100.00	

The first difference estimator

- Consider the standard model and consider two contiguous observations for the same observation unit i:

$$\begin{aligned} Y_{it} &= \alpha_i + \boldsymbol{X}'_{it}\beta + \epsilon_{it} \\ Y_{it-1} &= \alpha_i + \boldsymbol{X}'_{it-1}\beta + \epsilon_{it-1} \end{aligned}$$

The first difference estimator

- Consider the standard model and consider two contiguous observations for the same observation unit i:

$$\begin{aligned} Y_{it} &= \alpha_i + \boldsymbol{X}'_{it}\beta + \epsilon_{it} \\ Y_{it-1} &= \alpha_i + \boldsymbol{X}'_{it-1}\beta + \epsilon_{it-1} \end{aligned}$$

- Subtracting the period $t-1$ observation from period t observation yields:

$$Y_{it} - Y_{it-1} = [\boldsymbol{X}_{it} - \boldsymbol{X}_{it-1}]'\beta + \epsilon_{it} - \epsilon_{it-1}$$

The first difference estimator

- Consider the standard model and consider two contiguous observations for the same observation unit i:

$$\begin{aligned} Y_{it} &= \alpha_i + \boldsymbol{X}'_{it}\beta + \epsilon_{it} \\ Y_{it-1} &= \alpha_i + \boldsymbol{X}'_{it-1}\beta + \epsilon_{it-1} \end{aligned}$$

- Subtracting the period $t-1$ observation from period t observation yields:

$$Y_{it} - Y_{it-1} = [\boldsymbol{X}_{it} - \boldsymbol{X}_{it-1}]'\beta + \epsilon_{it} - \epsilon_{it-1}$$

- What assumption is needed for consistency (besides a rank condition)?

$$\mathbb{E}[\epsilon_{it} - \epsilon_{it-1} \mid \boldsymbol{X}_{it} - \boldsymbol{X}_{it-1}] = 0$$

The first difference estimator

- Consider the standard model and consider two contiguous observations for the same observation unit i:

$$\begin{aligned} Y_{it} &= \alpha_i + \boldsymbol{X}'_{it}\beta + \epsilon_{it} \\ Y_{it-1} &= \alpha_i + \boldsymbol{X}'_{it-1}\beta + \epsilon_{it-1} \end{aligned}$$

- Subtracting the period $t-1$ observation from period t observation yields:

$$Y_{it} - Y_{it-1} = [\boldsymbol{X}_{it} - \boldsymbol{X}_{it-1}]'\beta + \epsilon_{it} - \epsilon_{it-1}$$

- What assumption is needed for consistency (besides a rank condition)?

$$\mathbb{E}[\epsilon_{it} - \epsilon_{it-1} \mid \boldsymbol{X}_{it} - \boldsymbol{X}_{it-1}] = 0$$

- Example: $T = 2$.

5.2 The LSDV - dummy variable approach

- Add a dummy variable for each **observation unit**.

$$Y_{it} = \alpha_1 D_1 + \alpha_2 D_2 \; ... + \alpha_N D_N + \boldsymbol{X}'_{it}\beta + \epsilon_{it}$$

- These are analogous to other dummy variables, almost.
- The differences: what happens to #variables and #parameters when n increases?

The dummy variable approach

- Number of parameters should not be a fcn of the number of observation units.
- Remedy:
  1. (First) differencing.
  2. Taking deviations from observation unit specific means (and using software do this).

5.3 The fixed effects approach

- Calculate observation unit specific means of all variables. Start from

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

- Sum up and divide by number of observations / unit:

$$\overline{Y}_i = \overline{\alpha}_{0i} + \beta_1 \overline{X}_i + \bar{\epsilon}_{it}$$

The fixed effects approach

- Substract mean equation from "base" equation.
- Substract these from each observation.

$$Y_{it} - \overline{Y}_i = \alpha_i - \overline{\alpha}_i + \beta_1 (X_{it} - \overline{X}_i) + \epsilon_{it} - \bar{\epsilon}_{it}$$

$$= \beta_1 (X_{it} - \overline{X}_i) + \epsilon_{it} - \bar{\epsilon}_{it}$$

This is often called the **within transformation**, as it takes place within each observation unit.

6. Comparison of estimators

- Let us study the effect of age and having a university degree on log income.
- We use as data all the FLEED learning sample observations.

Comparison of estimators

- Let's use our FLEED data for demonstration purposes.
- Stata has some handy commands for checking the panel dimensions.

Comparison of estimators

Stata code

```
gen high_educ     = .
replace high_educ   = 0      if  ktutk != .
replace high_educ   = 1      if  educ >= 4
xtset shtun year
xtdescribe
```

Comparison of estimators

```
. xtdescribe

 shtun:  1, 2, ..., 8444                         n =      8444
 year:  1, 2, ..., 15                          T =        15
        Delta(year) = 1 unit
        Span(year)  = 15 periods
        (shtun*year uniquely identifies each observation)
```

Comparison of estimators

```
. xtdescribe

  shtun:  1, 2, ..., 8444                          n =      8444
   year:  1, 2, ..., 15                            T =        15
          Delta(year) = 1 unit
          Span(year)  = 15 periods
          (shtun*year uniquely identifies each observation)

Distribution of T_i:   min      5%     25%      50%      75%     95%     max
                         1       2       6       13       15      15      15

     Freq.  Percent    Cum. |  Pattern
 ---------------------------+-----------------
     3680     43.58   43.58 |  111111111111111
      333      3.94   47.52 |  111............
      313      3.71   51.23 |  .............11
      305      3.61   54.84 |  ...111111111111
      259      3.07   57.91 |  .........111111
      229      2.71   60.62 |  .......11111111
      214      2.53   63.16 |  1111111111111..
      208      2.46   65.62 |  11.............
      206      2.44   68.06 |  ..1111111111111
     2697     31.94  100.00 | (other patterns)
 ---------------------------+-----------------
     8444    100.00         |  XXXXXXXXXXXXXXX
```

Comparison of estimators

```
. pwcorr lnincome age high_educ, sig

             | lnincome      age high_e~c
-------------+---------------------------
    lnincome |   1.0000 
             |
             |
         age |   0.2590   1.0000 
             |   0.0000
             |
   high_educ |   0.2284   0.0486   1.0000 
             |   0.0000   0.0000
```

Comparison of estimators

```
. tabstat lnincome age high_educ, stat(mean sd p50 n) by(high_educ)

Summary statistics: mean, sd, p50, N
  by categories of: high_educ

high_educ |  lnincome       age  high_e~c
----------+------------------------------
        0 |  9.651306  42.78691         0
          |  .7869909  13.22139         0
          |  9.798127        42         0
          |     48698     50016     50016
----------+------------------------------
        1 |  10.36468  45.24554         1
          |  .6980709  11.86615         0
          |  10.49127        43         1
          |      3724      3922      3922
----------+------------------------------
    Total |  9.701983  42.96568  .0727131
          |  .8022147  13.14298  .2596674
          |  9.852194        43         0
          |     52422     53938     53938
-----------------------------------------
```

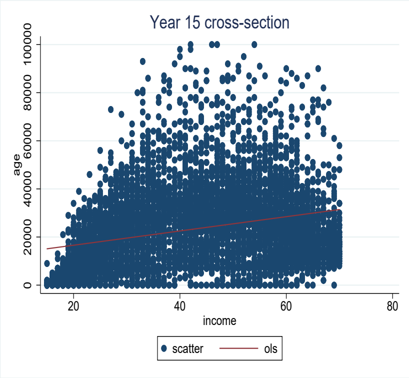
Comparison of estimators

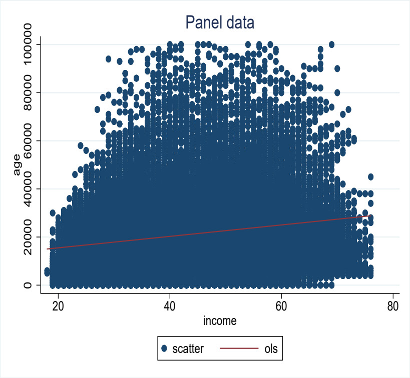
```
. ttest lnincome, by(high_educ)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |  48,698    9.651306    .0035663    .7869909    9.644316    9.658296
       1 |   3,724    10.36468    .0114392    .6980709    10.34225    10.38711
---------+--------------------------------------------------------------------
combined |  52,422    9.701983    .0035038    .8022147    9.695116     9.70885
---------+--------------------------------------------------------------------
    diff |           -.7133719    .0132786               -.7393982   -.6873457
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t = -53.7234
Ho: diff = 0                                     degrees of freedom =    52420

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

2010 cross section (LHS) vs panel data (RHS)





Comparison of estimators

Stata code

```
sort shtun year
bysort shtun: gen dlnincome    = lnincome - lnincome[_n - 1]
bysort shtun: gen dlnincome_v2  = d.lnincome
bysort shtun: gen dage       = age - age[_n - 1]
bysort shtun: gen dhigh_educ  = high_educ - high_educ[_n - 1]

regr lnincome age high_educ, robust
eststo ols
regr dlnincome dage dhigh_educ, robust
eststo fd
xtreg lnincome age high_educ , robust fe
eststo fe
xtreg lnincome age  if high_educ != ., robust fe
eststo fe_age
xtreg lnincome high_educ , robust fe
eststo fe_high_educ
estout ols fd fe*, keep(age dage high_educ dhigh_educ) cells(b(star fmt(3)) se(par fmt(2)))
stats(r2 r2_a F N, fmt(%9.5f %9.5f %9.0g))
```

Comparison of estimators

```
. xtreg lnincome age high_educ , robust fe

Fixed-effects (within) regression               Number of obs      =     52,422
Group variable: shtun                           Number of groups   =      4,921

R-sq:                                           Obs per group:
     within  = 0.2602                                         min =          1
     between = 0.1389                                         avg =       10.7
     overall = 0.1054                                         max =         15

                                                F(2,4920)          =    1855.02
corr(u_i, Xb)  = -0.7060                        Prob > F           =     0.0000

                              (Std. Err. adjusted for 4,921 clusters in shtun)
------------------------------------------------------------------------------
             |               Robust
    lnincome |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0611717   .0011271    54.27   0.000     .0589621    .0633814
   high_educ |   1.050748   .0477944    21.98   0.000     .9570499    1.144447
       _cons |   6.995669   .0482626   144.95   0.000     6.901052    7.090285
-------------+----------------------------------------------------------------
     sigma_u |  .94629399
     sigma_e |  .48687442
         rho |  .79069076   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

9

Comparison of estimators

```
. estout ols fd fe*, keep(age dage high_educ dhigh_educ) cells(b(star fmt(3)) se(par fmt(2)))
> mt(%9.5f %9.5f %9.0g))
```

	ols	fd	fe	fe_age	fe_high_educ
	b/se	b/se	b/se	b/se	b/se
age	0.016***		0.061***	0.066***	
	(0.00)		(0.00)	(0.00)	
high_educ	0.677***		1.051***		1.429***
	(0.01)		(0.05)		(0.05)
dage		-0.070			
		(0.05)			
dhigh_educ		0.496***			
		(0.05)			
r2	0.11995	0.00643	0.26025	0.22179	0.07303
r2_a	0.11992	0.00639	0.26022	0.22177	0.07301
F	3406.003	51.03453	1855.018	3016.825	914.4404
N	52422	47096	52422	52422	52422

10

Issues with first difference

- The *dhigh_educ* - dummy only takes values $0, 1$.
- More generally, the time-difference of a dummy can at most take values $-1, 0, 1$.
- Contrast this to the FE-version of *high_educ*.

Comparison of estimators

Stata code

```
bysort shtun: egen high_educ_mean = mean(high_educ) if e(sample)
gen high_educ_fe = high_educ - high_educ_mean

gen high_educ_fe_d    = 0
replace high_educ_fe_d  = 0.5   if high_educ_fe > 0 & high_educ_fe != .
replace high_educ_fe_d  = 1   if high_educ_fe == 1
tab high_educ_fe_d if e(sample)
centile high_educ_fe if e(sample), centile(0(10)100)
centile high_educ_fe if e(sample), centile(0(1)10)
centile high_educ_fe if e(sample), centile(90(1)100)
```

Tabulation of dhigh_educ and high_educ_fe

```
. tab dhigh_educ if e(sample)
```

dhigh_educ	Freq.	Percent	Cum.
0	47,497	99.48	99.48
1	249	0.52	100.00
Total	47,746	100.00	

```
. tab high_educ_fe_d if e(sample)
```

high_educ_f e_d	Freq.	Percent	Cum.
0	50,894	97.09	97.09
.5	1,528	2.91	100.00
Total	52,422	100.00	

Distribution of dhigh_educ_fe

```
. centile high_educ_fe if e(sample), centile(0(10)100)
```

Variable	Obs	Percentile	Centile	— Binom. Interp. — [95% Conf.	Interval]
high_educ_fe	52,422	0	-.9333333	-.9333333	-.9333333*
		10	0	0	0
		20	0	0	0
		30	0	0	0
		40	0	0	0
		50	0	0	0
		60	0	0	0
		70	0	0	0
		80	0	0	0
		90	0	0	0
		100	.9230769	.9230769	.9230769*

Distribution of dhigh_educ_fe

```
. centile high_educ_fe if e(sample), centile(0(1)10)
```

Variable	Obs	Percentile	Centile	— Binom. Interp. — [95% Conf.	Interval]
high_educ_fe	52,422	0	-.9333333	-.9333333	-.9333333*
		1	-.4615385	-.5	-.4444444
		2	-.1818182	-.2	-.1666667
		3	0	0	0
		4	0	0	0
		5	0	0	0
		6	0	0	0
		7	0	0	0
		8	0	0	0
		9	0	0	0
		10	0	0	0

Distribution of dhigh_educ_fe

```
. centile high_educ_fe if e(sample), centile(90(1)100)
```

Variable	Obs	Percentile	Centile	— Binom. Interp. — [95% Conf.	Interval]
high_educ_fe	52,422	90	0	0	0
		91	0	0	0
		92	0	0	0
		93	0	0	0
		94	0	0	0
		95	0	0	0
		96	0	0	0
		97	0	0	.0666667
		98	.2142857	.2	.2307692
		99	.4	.3571429	.4545454
		100	.9230769	.9230769	.9230769*

Time Fixed effects

- The **Fixed effects** panel data estimator with time FE is

$$Y_{it} = \alpha_{0i} + \beta_1 X_{it} + \beta_t + \epsilon_{it}$$

Time Fixed effects

Stata code

```
xtreg lnincome age high_educ i.year, fe
```

Time Fixed effects

```
. xtreg lnincome age high_educ i.year, fe
note: 15.year omitted because of collinearity

Fixed-effects (within) regression               Number of obs     =     52,422
Group variable: shtun                           Number of groups  =      4,921

R-sq:                                           Obs per group:
     within  = 0.2679                                         min =          1
     between = 0.1360                                         avg =       10.7
     overall = 0.1123                                         max =         15

                                                F(15,47486)       =    1158.46
corr(u_i, Xb)  = -0.6608                        Prob > F          =     0.0000

------------------------------------------------------------------------------
    lnincome |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0560064   .0008854    63.25   0.000     .0542709    .0577419
   high_educ |   1.038665   .0210499    49.34   0.000     .9974063    1.079923
             |
        year |
          2  |  -.0298515   .0123284    -2.42   0.015    -.0540154   -.0056876
          3  |  -.1419788   .0118634   -11.97   0.000    -.1652312   -.1187263
          4  |  -.1341604   .0113356   -11.84   0.000    -.1563783   -.1119425
          5  |  -.1570662   .0109702   -14.32   0.000    -.1785678   -.1355645
          6  |  -.1534954   .0106757   -14.38   0.000    -.1744199   -.1325708
          7  |  -.1547472    .010452   -14.81   0.000    -.1752334   -.1342611
          8  |  -.0841672   .0102395    -8.22   0.000    -.1042367   -.0640976
          9  |  -.0982579    .010146    -9.68   0.000    -.1181442   -.0783716
         10  |  -.0676674   .0100515    -6.73   0.000    -.0873685   -.0479663
         11  |  -.0708523   .0100877    -7.02   0.000    -.0906242   -.0510803
         12  |   -.073004   .0102014    -7.16   0.000    -.0929988   -.0530092
         13  |  -.0605306   .0103793    -5.83   0.000    -.0808742    -.040187
         14  |  -.0149915   .0105041    -1.43   0.154    -.0355797    .0055967
         15  |          0  (omitted)
             |
       _cons |   7.299694   .0389636   187.35   0.000     7.223324    7.376063
-------------+----------------------------------------------------------------
     sigma_u |  .88686083
     sigma_e |  .48441542
         rho |  .77020878   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(4920, 47486) = 15.11                Prob > F = 0.0000
```

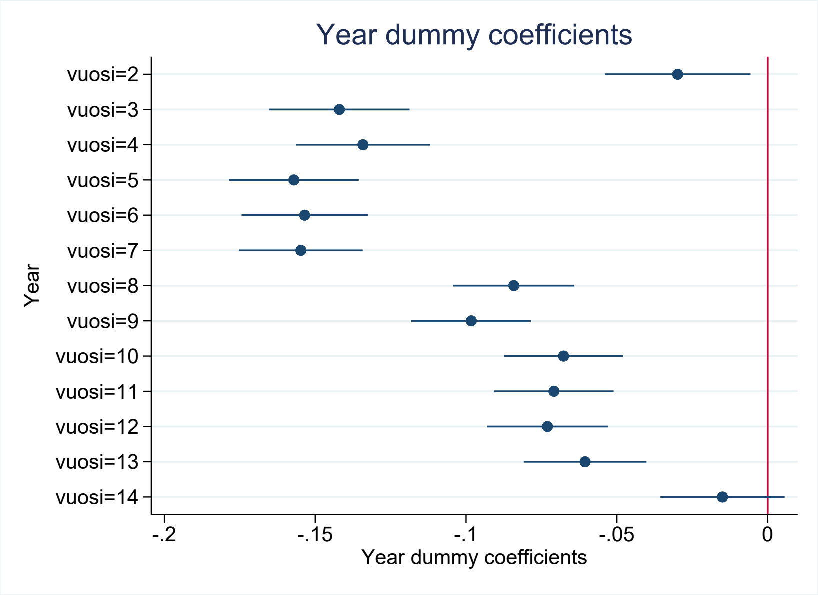
12

Time Fixed effects

Stata code

```
coefplot, drop(age high_educ _cons) ///
    xtitle("Year dummy coefficients") ///
    ytitle("coef.") ///
    title("Year dummy coefficients") ///
    xline(0) ///
    graphregion(fcolor(white))
graph export "YDcoef_fleed.pdf", replace
```

Time Fixed effects, base year = 15



7. FE assumptions

A1: conditional distribution of u has mean zero given **X**.

$$\mathbb{E}[\epsilon_{it} \mid \mathbf{X}_{it}, \alpha_i] = 0$$

this is called the **strict exogeneity** assumption.

A2: $\mathbf{X}_{it}, Y_{it}, i = 1..., n$ and $t = 1, ... T$ are i.i.d.

A3: $\mathbf{X}_{it}$ and Y_{it} have nonzero finite *fourth* moments.

FE assumptions

A4: No perfect multicollinearity.

A5: the errors for a given obs. unit are uncorrelated over time conditional on the observables.

$$corr[\epsilon_{it}, \epsilon_{is} \mid \mathbf{X}_{it}, \alpha_i] = 0 \text{ for } t \neq s.$$

FE A1 - Key benefit of the Fixed effects estimator

A1: We can rewrite the strict exogeneity assumption as

$$\mathbb{E}[\epsilon_{it}|\boldsymbol{x_{i1}}, ..., \boldsymbol{x_{iT}}, \alpha_i] = 0$$

FE A1 - Key benefit of the Fixed effects estimator

A1: We can rewrite the strict exogeneity assumption as

$$\mathbb{E}[\epsilon_{it} | \boldsymbol{x_{i1}}, ..., \boldsymbol{x_{iT}}, \alpha_i] = 0$$

- Notice this says nothing about the relationship between $\boldsymbol{X}_{i1}, ..., \boldsymbol{X}_{iT}$ and α_i.

FE A1 - Key benefit of the Fixed effects estimator

A1: We can rewrite the strict exogeneity assumption as

$$\mathbb{E}[\epsilon_{it}|\boldsymbol{x_{i1}}, ..., \boldsymbol{x_{iT}}, \alpha_i] = 0$$

- Notice this says nothing about the relationship between $\boldsymbol{X}_{i1}, ..., \boldsymbol{X}_{iT}$ and α_i.
- Thus the strict exogeneity assumption allows for **arbitrary correlation** between $\boldsymbol{X}_{it}$ and α_i.

FE A1 - Key downside of the Fixed effects estimator

A1: We can rewrite the strict exogeneity assumption as

$$\mathbb{E}[\epsilon_{it}|\boldsymbol{x_{i1}}, ..., \boldsymbol{x_{iT}}, \alpha_i] = 0$$

- Notice this says that ϵ_{it} **may not** be correlated with the previous values of $\boldsymbol{X}$ **as well as** the future values of $\boldsymbol{X}$. This feature is what gives it its name.
- As an example, the income-earnings shocks in year 5 cannot be correlated with level of education in year 1, nor in year 8.
- Think of how your current income earnings shock may be correlated with your future level of education.

FE A1 & A5 - Case R&D

A5: the errors for a given obs. unit are uncorrelated over time conditional on the observables.

- Let's use the R&D example.
- A5 implies that the "shock" that leads to high (low) productivity today disappears and the new "shock" tomorrow is uncorrelated.

Case R&D

- What could be a shock to productivity? E.g.,
  1. A new idea that gets implemented (and e.g. decreases waste).
  2. A new product that is introduced (and sells well at a high price).
- Some shocks are not transitory (i.e., they affect Y over many periods).
- In such cases A5 is violated: this period's shock is correlated with future values of the error term.

Case R&D #2

- What could be a shock to productivity? E.g.,
  1. R&D investment leads to a new idea that gets implemented (and e.g. decreases waste).
  2. A new product that is introduced (and sells well at a high price).
  3. The extra profits lead to more R&D in the future.
- In other words, this period's shock (ϵ_{it}) leads to a higher value of X_{it} in the future.
- This means that Assumption A1 is violated.

8. Measurement error and panel data

- Another way of seeing the problem with ”too little” within-unit, over-time variation: measurement error.
- Measurement error in a panel setting is more complex than in a cross-sectional setting.
- Recall that in cross-section, the noise-to-signal ratio is the source of measurement error, and we have **attenuation** bias towards zero.

Measurement error and panel data

- Now the measurement error can be
  1. **between** units and/or
  2. **within** units.
- If the measurement error is mostly between units, FE (or FD) removes it.
- If the measurement error is mostly within units **and** X is highly correlated over time , the bias due to measurement error is larger than in cross-section.
- In the R&D example, true RD is nearly constant over time and differences in reported RD are due to e.g. tax considerations or accounting issues.

9. Random effects estimator

- Think of the individual - specific constant as follows:

$$\alpha_i = \alpha + (\alpha_i - \alpha)$$

- That is, there is a common constant α and deviations from it.
- The FE estimator assumes that the deviations are "fixed". What if they were part of the stochastic error term? That is what the **random effects** estimator does.
- In the RE model the error term has two components: The within-unit constant η_i and the "regular" error term ϵ_{it}.
- The first one, η_i captures the permanent observation-unit specific shocks.
- The second one, ϵ_{it}, captures the observation-unit - time - period - specific shocks, just as before.

RE estimator

- Both η_i and ϵ_{it} need to be uncorrelated with $\boldsymbol{x_{it}}$.
- No autocorrelation in ϵ_{it} is allowed.
- No correlation across random effects η_i (across observation units) is allowed.
- Under the above assumptions, we can write:

$$
\begin{aligned}
y_{it} &= \alpha + \boldsymbol{x_{it}}'\beta + \eta_i + \epsilon_{it} \\
y_{it} &= \alpha + \boldsymbol{x_{it}}'\beta + w_{it}
\end{aligned}
$$

RE estimator

- If the RE assumptions hold, it is the efficient estimator and FE is inefficient.
- However, the RE assumptions are stricter as the explanatory variables are not allowed to be correlated with the random effect η_i whereas the fixed effects α_i are.

10. Clustering of standard errors

- Examples of clusters:

1. observation units in panel data.
2. individuals from a given firm in a cross-section or panel.
3. individuals in a family in a cross-section or panel.
4. firms in a multi-country cross-section or panel.

Key worry / insight

- Given a cluster-structure, errors may be correlated in a particular way.
- Errors may be correlated within clusters.
- Using (group) FE does not necessarily do away with the problem.
- In the presence of w/in-cluster correlation, se's are downward biased (Moulton 1986).
- Applies in particular to se's of regressors that are at a higher level of aggregation (=same value for each member in group g).
- Example: Using region dummies when estimating the effect of education on income in the FLEED data.

1. Clustering

- With clustering, one assumes that errors are uncorrelated across clusters, but may be correlated within clusters.
- This means that $\mathbb{E}[\epsilon_i \epsilon_j] = 0$ unless i and j are in the same cluster, but can be non-zero within a cluster.

The bias

- It can then be shown that the following regular standard errors are biased if there is within-cluster correlation.
- The size of bias depends on other things, too.

The remedy

- Do not use the standard (even heterosk. robust) standard errors.
- Use cluster-robust standard errors.
- Most packages calculate them.

What level of clustering?

- We face a traditional bias-variance trade-off: larger and fewer clusters have less bias, but more variability.
- The consensus is to be conservative and avoid bias and to use **bigger and more aggregate clusters** when possible, up to and including the point at which there is a concern about having too few clusters.
- One should keep in mind that the art and science of clustering is developing.