GV300 - Quantitative Political Analysis

University of Essex - Department of Government

Lorenzo Crippa

Week 23 – 2 March, 2020

NSS survey


NSS
National Student Survey
WHAT
DO YOU
THINK?
Department of Government
www.thestudentsurvey.com
University of Essex
£1
ESSEX
STUDENTS!
FOR EVERY SURVEY
YOU COMPLETE,
WE'LL GIVE £1
TO CHARITY

Today's activity

- Today we work on panel data. We use the dataset `panelModelsData_rFile.dta` that you find on Moodle.

Today's activity

- Today we work on panel data. We use the dataset `panelModelsData_rFile.dta` that you find on Moodle.
- Download the dataset and open it into R or Stata

Today's activity

- Today we work on panel data. We use the dataset `panelModelsData_rFile.dta` that you find on Moodle.
- Download the dataset and open it into R or Stata
- Data on 595 U.S. workers, measured at different points in time

Today's activity

- Today we work on panel data. We use the dataset `panelModelsData_rFile.dta` that you find on Moodle.
- Download the dataset and open it into R or Stata
- Data on 595 U.S. workers, measured at different points in time
- Consider only the first 100 workers, get rid of the rest of the data

Analysis

1. Causal identification

Analysis

1. Causal identification
   (a) We want to explain $log(wage)$ (`lwage`) as a function of at least experience (`exp`) and education (`ed`)

Analysis

1. Causal identification
   (a) We want to explain *log*(*wage*) (`lwage`) as a function of at least experience (`exp`) and education (`ed`)
   (b) What problem of identification do panel data help solving?

Analysis

1. Causal identification
   (a) We want to explain *log*(*wage*) (`lwage`) as a function of at least experience (`exp`) and education (`ed`)
   (b) What problem of identification do panel data help solving?
2. Data description

Analysis

1. Causal identification
   (a) We want to explain *log*(*wage*) (`lwage`) as a function of at least experience (`exp`) and education (`ed`)
   (b) What problem of identification do panel data help solving?
2. Data description
   (a) Summary statistics and plots as usual

Analysis

1. Causal identification
   (a) We want to explain *log*(*wage*) (`lwage`) as a function of at least experience (`exp`) and education (`ed`)
   (b) What problem of identification do panel data help solving?
2. Data description
   (a) Summary statistics and plots as usual
   (b) Today we have lots of information: we also want to summarize variation of data between units, within units and overall

Analysis

1. Causal identification
   (a) We want to explain *log*(*wage*) (`lwage`) as a function of at least experience (`exp`) and education (`ed`)
   (b) What problem of identification do panel data help solving?
2. Data description
   (a) Summary statistics and plots as usual
   (b) Today we have lots of information: we also want to summarize variation of data between units, within units and overall
3. Data analysis

Analysis

1. Causal identification
   (a) We want to explain *log*(*wage*) (`lwage`) as a function of at least experience (`exp`) and education (`ed`)
   (b) What problem of identification do panel data help solving?
2. Data description
   (a) Summary statistics and plots as usual
   (b) Today we have lots of information: we also want to summarize variation of data between units, within units and overall
3. Data analysis
   (a) Run a pooled, a fixed-effect and a random-effect model

Analysis

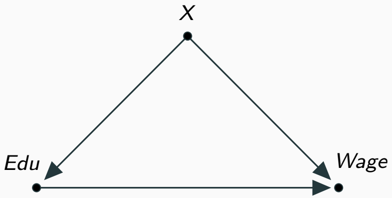
1. Causal identification
   (a) We want to explain *log*(*wage*) (`lwage`) as a function of at least experience (`exp`) and education (`ed`)
   (b) What problem of identification do panel data help solving?
2. Data description
   (a) Summary statistics and plots as usual
   (b) Today we have lots of information: we also want to summarize variation of data between units, within units and overall
3. Data analysis
   (a) Run a pooled, a fixed-effect and a random-effect model
   (b) Remember to use robust standard errors. Are robust standard errors enough or should we do something more?

Analysis

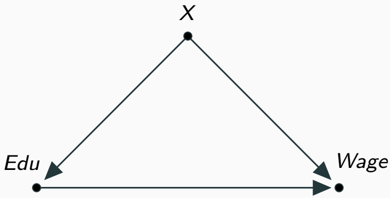
1. Causal identification
   (a) We want to explain *log*(*wage*) (`lwage`) as a function of at least experience (`exp`) and education (`ed`)
   (b) What problem of identification do panel data help solving?
2. Data description
   (a) Summary statistics and plots as usual
   (b) Today we have lots of information: we also want to summarize variation of data between units, within units and overall
3. Data analysis
   (a) Run a pooled, a fixed-effect and a random-effect model
   (b) Remember to use robust standard errors. Are robust standard errors enough or should we do something more?
   (c) Compare the results and discuss which model best fits data based on point 2(b).

1. Causal identification

1(a) and (b) – Causal model

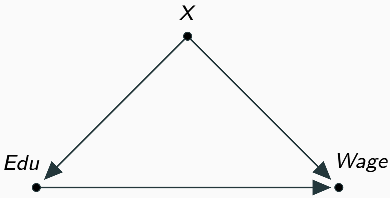


1(a) and (b) – Causal model



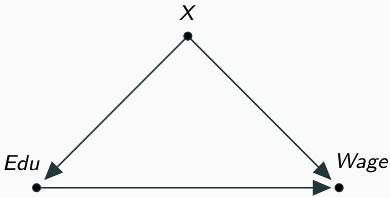
- Suppose X is a set of confounders that make each specific unit *idiosyncratic* both in treatment and outcome variables

1(a) and (b) – Causal model



- Suppose X is a set of confounders that make each specific unit *idiosyncratic* both in treatment and outcome variables
- Then panel data help us identify the causal effect of Edu on $Wage$ because they allow us to include a fixed (random) effect and control for it.

1(a) and (b) – Causal model



- Suppose X is a set of confounders that make each specific unit *idiosyncratic* both in treatment and outcome variables
- Then panel data help us identify the causal effect of *Edu* on *Wage* because they allow us to include a fixed (random) effect and control for it.
- **BUT** fixed/random effects are no solution for all other types of confounders (non-idiosyncratic)!

2. Data description

2(a) – Summary statistics

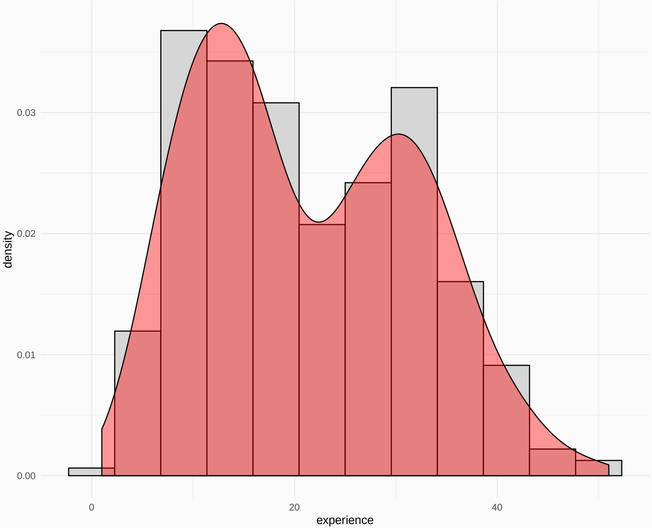
Variable	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
`id`	700	50.500	28.887	1	25.8	75.2	100
`t`	700	4.000	2.001	1	2	6	7
`occ`	700	0.486	0.500	0	0	1	1
`lwage`	700	6.704	0.492	5.165	6.397	7.003	8.049
`ed`	700	13.040	2.995	4	12	16	17
`exp`	700	21.280	10.788	1	12	30	51
`exp2`	700	569.060	508.066	1	144	900	2,601
`south`	700	0.327	0.470	0	0	1	1
`smsa`	700	0.627	0.484	0	0	1	1
`fem`	700	0.130	0.337	0	0	0	1
`union`	700	0.321	0.467	0	0	1	1

2(a) – Usual ways of describing data

We can use plots as usual to describe our data. For instance:

2(a) – Usual ways of describing data

We can use plots as usual to describe our data. For instance:

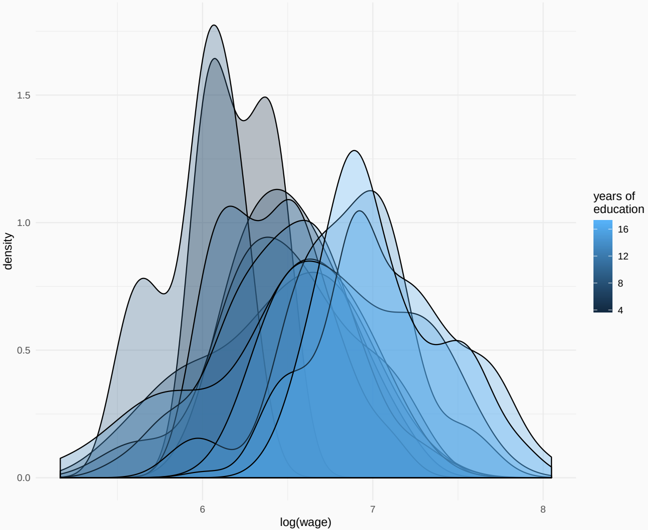


With `ggplot2` you can do basically everything

With `ggplot2` the limit is your imagination (sorry Stata users):

With `ggplot2` you can do basically everything

With `ggplot2` the limit is your imagination (sorry Stata users):

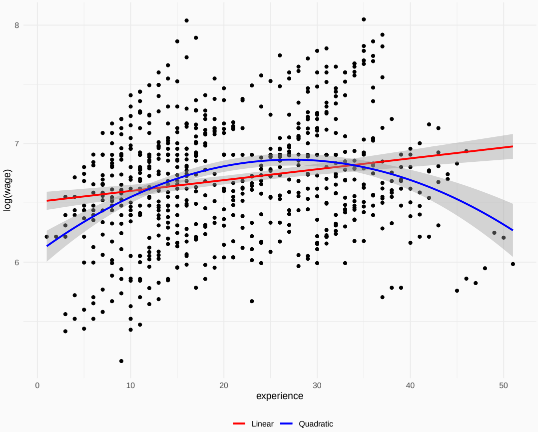


2(a) – Scatterplots

What does the following scatterplot tell you?

2(a) – Scatterplots

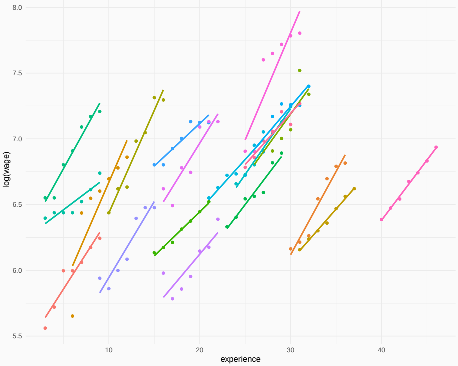
What does the following scatterplot tell you?



Our usual linear model doesn't really seem to fit this cloud. Why?

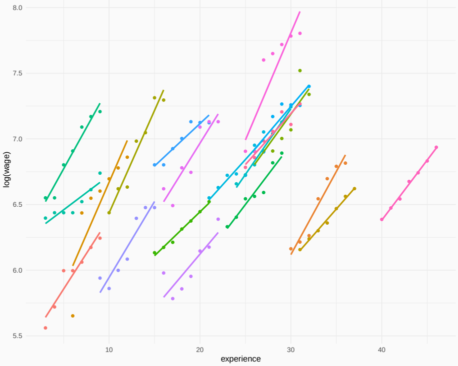
2(b) – Fixed effect

Take a look at this graph (only first 20 individuals). What does it suggest about the relationship we are studying?



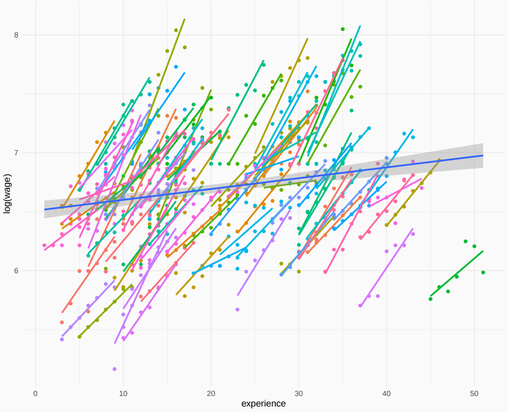
2(b) – Fixed effect

Take a look at this graph (only first 20 individuals). What does it suggest about the relationship we are studying?



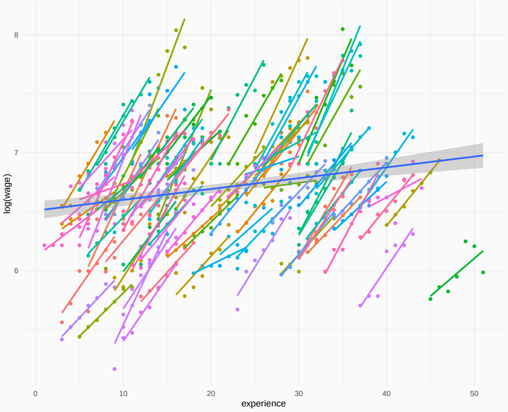
The slopes of the regression lines by individual seem similar, only between-units intercepts are different!

Now look at the following (all 100 individuals, plus pooled):



2(b) – Fixed effect

Now look at the following (all 100 individuals, plus pooled):



A typical example of a fixed effect: slopes and intercepts of lines by individuals are different from those of pooled line.

Fixed and random effect made intuitive

- You can have both unit- and time-fixed (random) effect

Fixed and random effect made intuitive

- You can have both unit- and time-fixed (random) effect
- Use fixed effect when intercepts **and** slopes of reg lines by units (or time) $\neq$ intercept and slope of the pooled reg line

Fixed and random effect made intuitive

- You can have both unit- and time-fixed (random) effect
- Use fixed effect when intercepts **and** slopes of reg lines by units (or time) $\neq$ intercept and slope of the pooled reg line
- Use random effect when **only** intercepts of reg lines by units (or time) $\neq$ intercept of the pooled reg line. Slopes $\approx$ to the slope of the pooled reg (with some variation)

Fixed and random effect made intuitive

- You can have both unit- and time-fixed (random) effect
- Use fixed effect when intercepts **and** slopes of reg lines by units (or time) $\neq$ intercept and slope of the pooled reg line
- Use random effect when **only** intercepts of reg lines by units (or time) $\neq$ intercept of the pooled reg line. Slopes $\approx$ to the slope of the pooled reg (with some variation)
- **Both** fixed and random have different intercepts of the reg lines by units (or time)!

Fixed and random effect made intuitive

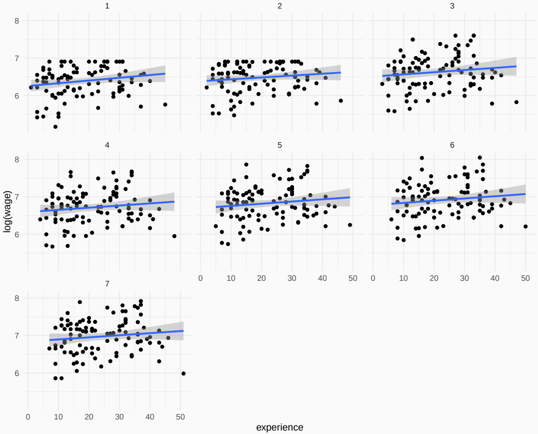
- You can have both unit- and time-fixed (random) effect
- Use fixed effect when intercepts **and** slopes of reg lines by units (or time) $\neq$ intercept and slope of the pooled reg line
- Use random effect when **only** intercepts of reg lines by units (or time) $\neq$ intercept of the pooled reg line. Slopes $\approx$ to the slope of the pooled reg (with some variation)
- **Both** fixed and random have different intercepts of the reg lines by units (or time)!
- Intuitively: you include a fixed effect in a pooled model $\rightarrow$ you have no clue of what's the difference between units, so you get rid of it.

Fixed and random effect made intuitive

- You can have both unit- and time-fixed (random) effect
- Use fixed effect when intercepts **and** slopes of reg lines by units (or time) $\neq$ intercept and slope of the pooled reg line
- Use random effect when **only** intercepts of reg lines by units (or time) $\neq$ intercept of the pooled reg line. Slopes $\approx$ to the slope of the pooled reg (with some variation)
- **Both** fixed and random have different intercepts of the reg lines by units (or time)!
- Intuitively: you include a fixed effect in a pooled model $\rightarrow$ you have no clue of what's the difference between units, so you get rid of it. Random effect $\rightarrow$ you can explain part of it.

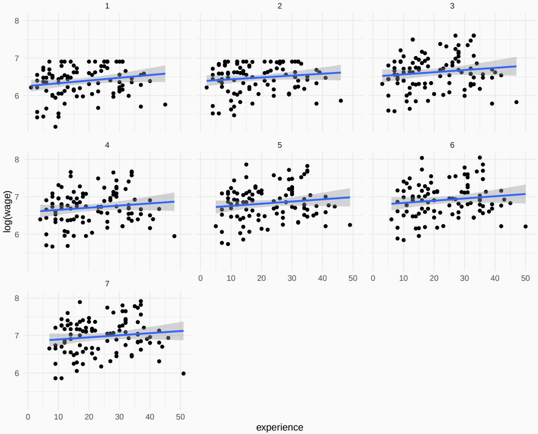
2(b) – What about time?

Are the time-points we have absolutely different from each others?



2(b) – What about time?

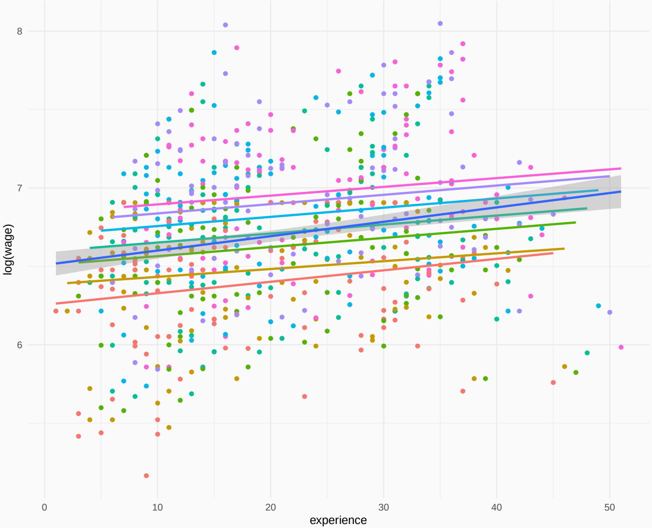
Are the time-points we have absolutely different from each others?



In R we can even animate our graphs using `gganimate`!

2(b) – Time fixed effect?

Look at the following lines: is a time-fixed effect appropriate here?



2(b) – Within, between and overall variation

We can get statistics about within, between and overall variation of a selected number of variables (see R script):

2(b) – Within, between and overall variation

We can get statistics about within, between and overall variation of a selected number of variables (see R script):

```
within.variation
   t     lwage   ed     exp    exp2   wks
   2.00  0.235   0      2.00   95.2   3.85

between.variation
   t     lwage   ed     exp    exp2   wks
   0     0.434   3.01   10.6   501.   3.24

overall.variation
   t     lwage   ed     exp    exp2   wks
   2.00  0.492   3.00   10.8   508.   5.02
```

3. Data analysis

Fixed effect: dummies

You introduce a fixed effect to perform a within-unit (or time) estimation.

Fixed effect: dummies

You introduce a fixed effect to perform a within-unit (or time) estimation. Suppose you have k different units. Mathematically:

$$Y_{it} = a_0 + \mathbf{X}'_{\mathbf{it}}\mathbf{b} + a_1 D_1 + a_2 D_2 + \ldots + a_{k-1} D_{k-1} + e_{it}$$

$$\mathbf{D} = D_1, D_2, \ldots, D_k = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \\ 0 & 0 & \ldots & 1 \end{bmatrix}$$

Fixed effect: dummies

You introduce a fixed effect to perform a within-unit (or time) estimation. Suppose you have k different units. Mathematically:

$$Y_{it} = a_0 + \mathbf{X_{it}'b} + a_1 D_1 + a_2 D_2 + \ldots + a_{k-1} D_{k-1} + e_{it}$$

$$\mathbf{D} = D_1, D_2, \ldots, D_k = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \\ 0 & 0 & \ldots & 1 \end{bmatrix}$$

The matrix has size $N \times k$. Your software will exclude one dummy (thus $k-1$) or the intercept a_0 to avoid multicollinearity

Fixed effect: entity demeaned

Problems in estimating a fixed effect using dummies:

Fixed effect: entity demeaned

Problems in estimating a fixed effect using dummies:

1. You easily run out of degrees of freedom

Fixed effect: entity demeaned

Problems in estimating a fixed effect using dummies:

1. You easily run out of degrees of freedom
2. Computationally heavy

Fixed effect: entity demeaned

Problems in estimating a fixed effect using dummies:

1. You easily run out of degrees of freedom
2. Computationally heavy

Solution: entity-demeaned FE (what your softwares do)

$$Y_{it} = a_0 + \mathbf{X}'_{\mathbf{it}}\mathbf{b} + u_{it}$$

Fixed effect: entity demeaned

Problems in estimating a fixed effect using dummies:

1. You easily run out of degrees of freedom
2. Computationally heavy

Solution: entity-demeaned FE (what your softwares do)

$$Y_{it} = a_0 + \mathbf{X}_{\mathbf{it}}'\mathbf{b} + u_{it} \rightarrow Y_{it} = a_0 + \mathbf{X}_{\mathbf{it}}'\mathbf{b} + a_i + e_{it}$$

Fixed effect: entity demeaned

Problems in estimating a fixed effect using dummies:

1. You easily run out of degrees of freedom
2. Computationally heavy

Solution: entity-demeaned FE (what your softwares do)

$$Y_{it} = a_0 + \mathbf{X}'_{\mathbf{it}}\mathbf{b} + u_{it} \rightarrow Y_{it} = a_0 + \mathbf{X}'_{\mathbf{it}}\mathbf{b} + a_i + e_{it}$$

$$\overline{Y}_i = a_0 + \overline{\mathbf{X}}'_{\mathbf{i}}\mathbf{b} + a_i + \overline{e}_i$$

Fixed effect: entity demeaned

Problems in estimating a fixed effect using dummies:

1. You easily run out of degrees of freedom
2. Computationally heavy

Solution: entity-demeaned FE (what your softwares do)

$$Y_{it} = a_0 + \mathbf{X_{it}'b} + u_{it} \rightarrow Y_{it} = a_0 + \mathbf{X_{it}'b} + a_i + e_{it}$$

$$\overline{Y}_i = a_0 + \overline{\mathbf{X}}_{\mathbf{i}}'\mathbf{b} + a_i + \overline{e}_i$$

We can now do the group demeaning:

$$(Y_{it} - \overline{Y}_i) = (\mathbf{X_{it}'} - \overline{\mathbf{X}}_{\mathbf{i}}')\mathbf{b} + (a_i - a_i) + (e_{it} - \overline{e}_i)$$

Fixed effect: entity demeaned

Problems in estimating a fixed effect using dummies:

1. You easily run out of degrees of freedom
2. Computationally heavy

Solution: entity-demeaned FE (what your softwares do)

$$Y_{it} = a_0 + \mathbf{X'_{it}}\mathbf{b} + u_{it} \rightarrow Y_{it} = a_0 + \mathbf{X'_{it}}\mathbf{b} + a_i + e_{it}$$

$$\overline{Y}_i = a_0 + \mathbf{\overline{X}'_i}\mathbf{b} + a_i + \overline{e}_i$$

We can now do the group demeaning:

$$(Y_{it} - \overline{Y}_i) = (\mathbf{X'_{it}} - \mathbf{\overline{X}'_i})\mathbf{b} + (a_i - a_i) + (e_{it} - \overline{e}_i)$$

Thus we get rid of between-unit variation $(a_i - a_i = 0)$

Maths on random effect

Random effect:

$$(Y_{it} - \hat{\theta}_i \overline{Y}_i) = (1 - \hat{\theta}_i)a_0 + (\mathbf{X}_{it} - \hat{\theta}_i \overline{\mathbf{X}_i})'\mathbf{b} + [(1 - \hat{\theta}_i)a_i + (e_{it} - \hat{\theta}_i \overline{e}_i)]$$

Maths on random effect

Random effect:

$$(Y_{it} - \hat{\theta}_i \overline{Y}_i) = (1 - \hat{\theta}_i)a_0 + (\mathbf{X}_{it} - \hat{\theta}_i \overline{\mathbf{X}_i})'\mathbf{b} + [(1 - \hat{\theta}_i)a_i + (e_{it} - \hat{\theta}_i \overline{e}_i)]$$

where $\hat{\theta}_i$ is an estimate of

$$\theta_i = 1 - \sqrt{\frac{\sigma_e^2}{T_i \sigma_a^2 + \sigma_e^2}}$$

Maths on random effect

Random effect:

$$(Y_{it} - \hat{\theta}_i \overline{Y}_i) = (1 - \hat{\theta}_i) a_0 + (\mathbf{X}_{it} - \hat{\theta}_i \overline{\mathbf{X}_i})' \mathbf{b} + [(1 - \hat{\theta}_i) a_i + (e_{it} - \hat{\theta}_i \overline{e}_i)]$$

where $\hat{\theta}_i$ is an estimate of

$$\theta_i = 1 - \sqrt{\frac{\sigma_e^2}{T_i \sigma_a^2 + \sigma_e^2}}$$

Assumption: $E(e_{it} | \mathbf{x}_{it}) = 0$

Maths on random effect

Random effect:

$$(Y_{it} - \hat{\theta}_i \overline{Y}_i) = (1 - \hat{\theta}_i) a_0 + (\mathbf{X}_{it} - \hat{\theta}_i \overline{\mathbf{X}_i})' \mathbf{b} + [(1 - \hat{\theta}_i) a_i + (e_{it} - \hat{\theta}_i \overline{e}_i)]$$

where $\hat{\theta}_i$ is an estimate of

$$\theta_i = 1 - \sqrt{\frac{\sigma_e^2}{T_i \sigma_a^2 + \sigma_e^2}}$$

Assumption: $E(e_{it} | \mathbf{x}_{it}) = 0$

Notice that, for each covariate in $\mathbf{X_{it}}$, you have a (slightly) different slope for each unit i, depending on $\hat{\theta}_i$!

From random effect to fixed effect and pooled model

Now, consider:

$$(Y_{it} - \hat{\theta}_i \overline{Y}_i) = (1 - \hat{\theta}_i) a_0 + (\mathbf{X}_{it} - \hat{\theta}_i \overline{\mathbf{X}_i})' \mathbf{b} + [(1 - \hat{\theta}_i) a_i + (e_{it} - \hat{\theta}_i \overline{e}_i)]$$

$$\theta_i = 1 - \sqrt{\frac{\sigma_e^2}{T_i \sigma_a^2 + \sigma_e^2}}$$

From random effect to fixed effect and pooled model

Now, consider:

$$(Y_{it} - \hat{\theta}_i \overline{Y}_i) = (1 - \hat{\theta}_i)a_0 + (\mathbf{X}_{it} - \hat{\theta}_i \overline{\mathbf{X}_i})'\mathbf{b} + [(1 - \hat{\theta}_i)a_i + (e_{it} - \hat{\theta}_i \overline{e}_i)]$$

$$\theta_i = 1 - \sqrt{\frac{\sigma_e^2}{T_i \sigma_a^2 + \sigma_e^2}}$$

1. what if $\sigma_a^2 \gg \sigma_e^2$ (huge between-unit variation)?

From random effect to fixed effect and pooled model

Now, consider:

$$(Y_{it} - \hat{\theta}_i \overline{Y}_i) = (1 - \hat{\theta}_i) a_0 + (\mathbf{X}_{it} - \hat{\theta}_i \overline{\mathbf{X}_i})' \mathbf{b} + [(1 - \hat{\theta}_i) a_i + (e_{it} - \hat{\theta}_i \overline{e}_i)]$$

$$\theta_i = 1 - \sqrt{\frac{\sigma_e^2}{T_i \sigma_a^2 + \sigma_e^2}}$$

1. what if $\sigma_a^2 \gg \sigma_e^2$ (huge between-unit variation)? $\theta_i \to 1$: we are back to the entity-demeaned fixed effect!

From random effect to fixed effect and pooled model

Now, consider:

$$(Y_{it} - \hat{\theta}_i \overline{Y}_i) = (1 - \hat{\theta}_i) a_0 + (\mathbf{X}_{it} - \hat{\theta}_i \overline{\mathbf{X}_i})' \mathbf{b} + [(1 - \hat{\theta}_i) a_i + (e_{it} - \hat{\theta}_i \overline{e}_i)]$$

$$\theta_i = 1 - \sqrt{\frac{\sigma_e^2}{T_i \sigma_a^2 + \sigma_e^2}}$$

1. what if $\sigma_a^2 \gg \sigma_e^2$ (huge between-unit variation)? $\theta_i \to 1$: we are back to the entity-demeaned fixed effect!
2. what if $\sigma_a^2 \ll \sigma_e^2$ (minimal between-unit variation)?

From random effect to fixed effect and pooled model

Now, consider:

$$(Y_{it} - \hat{\theta}_i \overline{Y}_i) = (1 - \hat{\theta}_i) a_0 + (\mathbf{X}_{it} - \hat{\theta}_i \overline{\mathbf{X}_i})' \mathbf{b} + [(1 - \hat{\theta}_i) a_i + (e_{it} - \hat{\theta}_i \overline{e}_i)]$$

$$\theta_i = 1 - \sqrt{\frac{\sigma_e^2}{T_i \sigma_a^2 + \sigma_e^2}}$$

1. what if $\sigma_a^2 \gg \sigma_e^2$ (huge between-unit variation)? $\theta_i \to 1$: we are back to the entity-demeaned fixed effect!
2. what if $\sigma_a^2 \ll \sigma_e^2$ (minimal between-unit variation)? $\theta_i \to 0$: we are back to the pooled model!

From random effect to fixed effect and pooled model

Now, consider:

$$(Y_{it} - \hat{\theta}_i \overline{Y}_i) = (1 - \hat{\theta}_i) a_0 + (\mathbf{X}_{it} - \hat{\theta}_i \overline{\mathbf{X}_i})' \mathbf{b} + [(1 - \hat{\theta}_i) a_i + (e_{it} - \hat{\theta}_i \overline{e}_i)]$$

$$\theta_i = 1 - \sqrt{\frac{\sigma_e^2}{T_i \sigma_a^2 + \sigma_e^2}}$$

1. what if $\sigma_a^2 \gg \sigma_e^2$ (huge between-unit variation)? $\theta_i \to 1$: we are back to the entity-demeaned fixed effect!
2. what if $\sigma_a^2 \ll \sigma_e^2$ (minimal between-unit variation)? $\theta_i \to 0$: we are back to the pooled model!
3. in-between these extremes: random effect

From random effect to fixed effect and pooled model

Now, consider:

$$(Y_{it} - \hat{\theta}_i \overline{Y}_i) = (1 - \hat{\theta}_i)a_0 + (\mathbf{X}_{it} - \hat{\theta}_i \overline{\mathbf{X}_i})'\mathbf{b} + [(1 - \hat{\theta}_i)a_i + (e_{it} - \hat{\theta}_i \overline{e}_i)]$$

$$\theta_i = 1 - \sqrt{\frac{\sigma_e^2}{T_i \sigma_a^2 + \sigma_e^2}}$$

1. what if $\sigma_a^2 \gg \sigma_e^2$ (huge between-unit variation)? $\theta_i \to 1$: we are back to the entity-demeaned fixed effect!
2. what if $\sigma_a^2 \ll \sigma_e^2$ (minimal between-unit variation)? $\theta_i \to 0$: we are back to the pooled model!
3. in-between these extremes: random effect

Your software will usually provide to you an estimate of θ_i (or of the variance of slopes) and related tests. Check the output!

3(c) – Model comparison

	Pooled	Pooled (clustered SE)	Fixed Effect	Random Effect
(Intercept)	4.72***	4.72***		3.20***
	(0.15)	(0.24)		(0.33)
exp	0.06***	0.06***	0.11***	0.09***
	(0.01)	(0.01)	(0.01)	(0.01)
exp2	−0.00***	−0.00***	−0.00	−0.00*
	(0.00)	(0.00)	(0.00)	(0.00)
wks	−0.00	−0.00	0.00*	0.00
	(0.00)	(0.00)	(0.00)	(0.00)
ed	0.10***	0.10***		0.14***
	(0.01)	(0.01)		(0.02)
Adj. R^2	0.43	0.43	0.70	0.52
Num. obs.	700	700	700	700

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Fixed effect or random effect?

We can perform a Hausman test to decide between FE and RE

Fixed effect or random effect?

We can perform a Hausman test to decide between FE and RE

- H_0: difference between coefficients estimated in a fixed and random effect is not systematic

Fixed effect or random effect?

We can perform a Hausman test to decide between FE and RE

- H_0: difference between coefficients estimated in a fixed and random effect is not systematic
- Under H_0, RE will provide consistent **and efficient** estimates

Fixed effect or random effect?

We can perform a Hausman test to decide between FE and RE

- H_0: difference between coefficients estimated in a fixed and random effect is not systematic
- Under H_0, RE will provide consistent **and efficient** estimates
- Under H_1, RE will provide inconsistent estimates

Fixed effect or random effect?

We can perform a Hausman test to decide between FE and RE

- H_0: difference between coefficients estimated in a fixed and random effect is not systematic
- Under H_0, RE will provide consistent **and efficient** estimates
- Under H_1, RE will provide inconsistent estimates

In R (Stata is very similar, see script):

```
phtest(fe, re)

##############################################
#   Hausman Test
#
# data:  lwage ~ exp + exp2 + wks + ed
# chisq = 3805.5, df = 3, p-value < 2.2e-16
# alternative hypothesis: one model is inconsistent
##############################################
```

Final remarks

Remember:

- Heteroskedasticity-robust standard errors are not enough anymore with panel data

Final remarks

Remember:

- Heteroskedasticity-robust standard errors are not enough anymore with panel data
- We should use **clustered** SEs (on our units) when we have panel data, otherwise we are not fixing serial correlation between units

Final remarks

Remember:

- Heteroskedasticity-robust standard errors are not enough anymore with panel data
- We should use **clustered** SEs (on our units) when we have panel data, otherwise we are not fixing serial correlation between units
- These SEs will **also** be robust to heteroskedasticity

Final remarks

Remember:

- Heteroskedasticity-robust standard errors are not enough anymore with panel data
- We should use **clustered** SEs (on our units) when we have panel data, otherwise we are not fixing serial correlation between units
- These SEs will **also** be robust to heteroskedasticity
- Look at the education variable. Why is it omitted from the fixed effect model?

Final remarks

Remember:

- Heteroskedasticity-robust standard errors are not enough anymore with panel data
- We should use **clustered** SEs (on our units) when we have panel data, otherwise we are not fixing serial correlation between units
- These SEs will **also** be robust to heteroskedasticity
- Look at the education variable. Why is it omitted from the fixed effect model?
- Because for each respondent (within unit) education does not change! Thus there is no within-variation, and the within-variation is the only thing a fixed-effect looks at.

Conclusion

All clear? More questions?

Thanks and see you next week!