Pooled Cross-Sections and Panel Data Models

Applied Econometrics for Researchers, PhD

Vera Rocha, CBS-SI, vr.si@cbs.dk

30th November 2022

Agenda for today

1. Pooled cross-sections vs. panel data
2. The value of pooled cross-sections & panel data for policy analysis *(only brief intro, more on this next week)*
3. Two-period panel data and extensions
4. First-differences model
5. Fixed and random effects estimators
6. Examples along the way (including in Stata)

Key readings:

- Wooldridge, *Introductory econometrics, a modern approach*, **Chapters 13 & 14**
- For guidance on R, check Cunningham, *Mixtape*, **Chapter 8 on Panel Data**

Introducing a time dimension:

Cross-sectional data *(so far)*

- Observing many subjects drawn *independently* from the population
- Observing them at the same point in time

Time-series data *(not our focus)*

- Observing the same ”subject” (e.g., country)
- Over multiple periods

Pooled cross-sections

- Observing many subjects drawn *independently* from the population
- Observed at a number of different points in time; only one observation per subject

Panel/Longitudinal data

- Observing many subjects drawn *independently* from the population
- Observed at a number of different points in time; following subjects over time

Cross-sections, Pooled Cross-sections & Panels

Cross-Section: Observations in a given period (*t*) for subjects (e.g., firms, individuals, households, countries, … – ideally a random sample from the population):

$$(y_{it}, x_{it1}, x_{it2}, x_{it3}, \ldots, x_{itk})\ with\ i = 1,2, \ldots, n$$ *[t could be omitted]*

Extending this to a **2-period case**:

Period 1: $(y_{i1}, x_{i11}, x_{i12}, x_{i13}, \ldots, x_{i1k})$

Period 2: $(y_{i2}, x_{i21}, x_{i22}, x_{i23}, \ldots, x_{i2k})$

- If individuals in Period 1 **≠** individuals in Period 2: **Independent & Pooled Cross-Sections**
- If individuals in Period 1 **=** individuals in Period 2: **Panel Data**

Pooling Cross-sections across Time

Individuals are not the same in periods 1 and 2!

Pooling **independent cross-sections** for two periods :

Period 1: $(y_{i1}, x_{i11}, x_{i12}, x_{i13}, \ldots, x_{i1k})$; Period 2: $(y_{i2}, x_{i21}, x_{i22}, x_{i23}, \ldots, x_{i2k})$

Three possible approaches:

1. Estimate a joint model, ignoring the time dimension (pooling)

Increased sample size

$$Y = X\beta + \mathrm{u} \rightarrow \hat{\beta} pooled$$

2. Use the cross-sections separately (no pooling)

We could test: $\widehat{\beta_1} = \widehat{\beta_2}$

$t = 1$: $Y = X\beta_1 + \mathrm{u} \rightarrow \widehat{\beta_1}$ $\quad$ $t = 2$: $Y = X\beta_2 + \mathrm{u} \rightarrow \widehat{\beta_2}$

3. Combine cross-sections but allow certain coefficients to differ between periods (partial pooling) – possibly convenient in case of structural change

Possibly most interesting

Adding Time Dummies to the Regression

If 2 periods: a **dummy variable $\boldsymbol{d2} = \mathbf{1}$** if period = 2 (and 0 otherwise) captures specificities of each period:

$$y_i = \beta_0 + \delta_0 d2_i + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + u_i,$$

$i = 1, 2, \ldots, n$ (and not necessarily the same in the two periods)

- The constant term is then allowed to change between period 1 and 2
- Other coefficients can also change between periods if interacted with the time dummy *d2*
- *Example: Has the return to education changed over time?*

$$\log(wage)_i = \beta_0 + \delta_0 d2_i + \beta_1 educ_i + \delta_1 d2_i \cdot educ_i + \cdots + u_i$$

- Test H_0: $\delta_1 = 0$

Note: If multiple periods, we could add dummies for all of them, and use the first period as reference.

Change in Coefficients across Time: Example

```
. tab y85
```

y85	Freq.	Percent	Cum.
0	550	50.74	50.74
1	534	49.26	100.00
Total	1,084	100.00	

Two pooled cross-sections: different individuals surveyed in 1978 & 1985

```
. reg lwage y85 educ y85educ female y85fem exper expersq union
```

Source	SS	df	MS
Model	135.992074	8	16.9990092
Residual	183.099094	1,075	.170324738
Total	319.091167	1,083	.29463635

Number of obs	=	1,084
F(8, 1075)	=	99.80
Prob > F	=	0.0000
R-squared	=	0.4262
Adj R-squared	=	0.4219
Root MSE	=	.4127

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
y85	.1178062	.1237817	0.95	0.341	-.125075	.3606874
educ	.0747209	.0066764	11.19	0.000	.0616206	.0878212
y85educ	.0184605	.0093542	1.97	0.049	.000106	.036815
female	-.3167086	.0366215	-8.65	0.000	-.3885663	-.244851
y85fem	.085052	.051309	1.66	0.098	-.0156251	.185729
exper	.0295843	.0035673	8.29	0.000	.0225846	.036584
expersq	-.0003994	.0000775	-5.15	0.000	-.0005516	-.0002473
union	.2021319	.0302945	6.67	0.000	.1426888	.2615749
_cons	.4589329	.0934485	4.91	0.000	.2755707	.642295

- The return to education in 1978 is estimated to be ≈ 7.5%, being 1.8% **higher** in 1985 (i.e. return to education in 1985 is ≈ 0.075+0.018 ≈ 9%).
- Gender wage gap of ≈ 27% in 1978 [100*(exp(-0.317)-1)%], with a tendency to decrease later (marginally in significance). GWG = 21% in 1985.

Dataset available here: https://ideas.repec.org/p/boc/bocins/cps78_85.html

Structural Change across Time

- If we allow all coefficients to change over time:

$$y_i = \beta_0 + \delta_0 d2_i + \beta_1 x_{i1} + \delta_1 d2_i x_{i1} + \beta_2 x_{i2} + \delta_2 d2_i x_{i2}$$
$$+ \ldots + \beta_k x_{ik} + \delta_k d2_i x_{ik} + u_i$$

- F-test for $\delta_0 = \delta_1 = \cdots = \delta_k = 0$
- Same logic of a Chow-test to determine whether a multiple regression function differs across two groups
- Maybe better to keep cross-sections separate?

Policy Analysis with Pooled Cross-sections

- Pooled cross-sections can be useful for evaluating the impact of certain events or policy interventions
- Two cross-sectional datasets may be enough if collected before and after the occurrence of the event/intervention
- Event or policy intervention must be a ”natural (or a quasi-) experiment”
    - **Control Group** – not affected by the policy change/event/intervention
    - **Treatment Group** – affected by the policy change/event/intervention
- Key ingredients for a difference-in-differences analysis
    - Treated vs. Control group difference
    - Time difference

More details on this next week!

Example: Diff-in-Diff with Pooled Cross-sections

Warm up for next week

$$\log(durat) = \beta_0 + \delta_0 after + \beta_1 highearn + \boldsymbol{\delta_1} after \cdot highearn + controls + u$$

- 2 pooled cross-sections, different years
- $\boldsymbol{y}$ **[$\boldsymbol{log(durat)}$]:** length of time (in weeks) that an injured worker receives workers' compensation
- $\boldsymbol{d2}$ **[$\boldsymbol{after}$]:** dummy=1 once the cap on weekly earnings covered by workers' compensation was raised
- $\boldsymbol{dT}$ **[$\boldsymbol{highearn}$]:** dummy=1 for high-income workers; 0 for low-income workers
- To what extent more generous workers' compensation causes people to stay out of work longer (everything else constant)?

Example: Diff-in-Diff with Pooled Cross-sections

- $afchange$ is statistically insignificant → the increase in the earnings cap has no effect on duration for low-income workers (δ_0)
- High-income earners had already ≈ 15% longer time on workers' compensation before the policy change (β_1)
- The average length of time on workers' compensation <u>for high earners</u> increased by further ≈ 22% due to the increase in earnings cap. ($\boldsymbol{\delta_1}$)

```
. reg ldurat afchnge highearn afhigh_highearn male married manuf construc head neck

      Source |       SS           df       MS      Number of obs   =     6,824
-------------+----------------------------------   F(9, 6814)      =     23.33
       Model |  346.049473         9  38.4499415   Prob > F        =    0.0000
    Residual |  11227.9331     6,814  1.64777416   R-squared       =    0.0299
-------------+----------------------------------   Adj R-squared   =    0.0286
       Total |  11573.9826     6,823  1.69631871   Root MSE        =    1.2837

------------------------------------------------------------------------------
      ldurat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     afchnge |   .0226262    .039981     0.57   0.571     -.055749    .1010013
    highearn |   .1512806    .047027     3.22   0.001      .059093    .2434682
afhigh_highearn | .2237872   .0637635     3.51   0.000     .0987909    .3487836
        male |  -.1208571   .0405547    -2.98   0.003    -.2003569   -.0413573
     married |   .1227717   .0351384     3.49   0.000     .0538895    .1916539
       manuf |  -.1110681   .0359758    -3.09   0.002    -.1815919   -.0405443
    construc |   .2110412    .046122     4.58   0.000     .1206276    .3014547
        head |    -.47542   .0824976    -5.76   0.000    -.6371411   -.3136989
        neck |   .2774732   .1215091     2.28   0.022     .0392774    .5156691
       _cons |   1.240535   .0458285    27.07   0.000     1.150697    1.330373
------------------------------------------------------------------------------
```

Note: Control variables can significantly affect the "treatment effect"! (Recall matching lecture)

Dataset available here: https://ideas.repec.org/p/boc/bocins/injury.html

Panel Data: Two-Period Case

Individuals are now the same in periods 1 and 2!

Same *n* subjects observed in period 1 and period 2.

- Period 1: $(y_{i1}, x_{i11}, x_{i12}, x_{i13}, \ldots, x_{i1k})$
- Period 2: $(y_{i2}, x_{i21}, x_{i22}, x_{i23}, \ldots, x_{i2k})$

N subjects = *2*N* observations (*balanced panel*)

Period 2 can be years (months, weeks, …) after period 1

i can correspond to persons, firms, households, countries, regions…

Also known as *longitudinal data.*

Consider simplest case: One regressor. We want to estimate the **causal effect of *x* on *y***, keeping all else equal.

Back to *Unobserved Effects*

Model (example): $y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it} + \boldsymbol{a_i} + u_{it}$

Time dummy: $d2_t$ takes the same value for all subjects (”macro effect”)

Composite error term: $v_{it} = a_i + u_{it}$

Unobserved ”fixed effect” a_i (unobserved heterogeneity):

- Time invariant
- Specific to each subject (firm, person, county…)
- All factors that affect y_{it} that do not change over time

Idiosyncratic error u_{it}:

- Varies randomly across both subjects and time: The ”usual” error term

Is (pooled) OLS an option?

Note that we now have a composite error term: $v_{it} = a_i + u_{it}$

In order for **pooled OLS** to produce a consistent estimate of β_1, it is required that:

$$cov(x_{it}, v_{it}) = 0$$

and this now implies $\boldsymbol{cov(x_{it}, a_i) = 0}$ too!

It is anyway recommended to use clustered-robust s.e. when using pooled OLS!

It is no longer enough to satisfy $cov(x_{it}, u_{it}) = 0$.

Back to **”omitted variable bias”** or **”heterogeneity bias”**

Solution: Panel Structure!

Repeated observations of the same subjects offers a solution:

- First-differences (FD) estimator
- $T = 1$: $y_{i1} = \beta_0 + \beta_1 x_{i1} + \boldsymbol{a_i} + u_{i1}$
- $T = 2$: $y_{i2} = (\beta_0 + \delta_0) + \beta_1 x_{i2} + \boldsymbol{a_i} + u_{i2}$
- Subtracting the first from the second:
- $\Delta y_i = \delta_0 + \beta_1 \Delta x_i + \Delta u_i$ → we get rid of $\boldsymbol{a_i}$; OLS should now be consistent, as soon as $cov(\Delta x_i, \Delta u_i) = 0$

Recall the returns to education example: a_i can capture unobserved (permanent) ability

If not, find an instrument for *x*; recall IV lecture

Any problems with this approach?

- We **cannot estimate the parameters of time invariant variables** (e.g., gender; and even time-varying variables that do not vary within the time period covered, e.g., education?)
- An alternative might be to **interact these variables with time dummies or other time-varying variables**.
- Panel data might also be difficult to obtain – especially through surveys.
- Country-level registers might be an (amazing) alternative.
- **Still:** with a **2-period panel** we can **control for unobserved effects** → not possible in standard cross-sections (**omitted variable bias**)

Policy Analysis (again) with Panel Data

- Panel data can be (even more) useful for policy analysis than repeated cross-sections
- Individuals/firms often **self-select** into the treatment/program (*recall endogeneity lecture – bias due to self-selection into treatment*)
- …or they are assigned to the program based on **(unobserved) characteristics that could be related to the outcome variable**.
- Assume a number of individuals go through the ”program” in period 2, but some do not.
- Define a ”treatment” dummy: $treat_{i1} = 1/0$, for treated/control individuals

$$\boldsymbol{y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 treat_{it} + a_i + u_{it}}$$

- $\boldsymbol{a_i}$ captures time invariant characteristics, including those that could affect self-selection (or assignment) into the treatment/program

Policy Analysis (now) with First-Differences

First-difference the previous model and use OLS to estimate:

$$\Delta y_i = \delta_0 + \beta_1 \Delta treat_i + \Delta u_i$$

But because the treatment happens only in period 2:

$$\Delta treat_i = treat_{i2}$$

$$\widehat{\beta_1} = \overline{\Delta y}_{Treated} - \overline{\Delta y}_{Control}$$

Difference over time is now **within the same individual**.

OLS estimator is consistent as soon as $cov(treat_i, \Delta u_i) = 0$

Policy Analysis using FD: Example

$$\boldsymbol{scrap_{it} = \beta_0 + \delta_0 y88_t + \beta_1 grant_{it} + a_i + u_{it}}, \qquad t = 1987{,}1988$$

- ***Scrap*** rate of firm i in year t refers to the % items that must be scrapped due to defects (if lower ⬄ ”higher productivity”)
- ***Grant*** is a binary indicator = 1 for firms that received a job training grant in 1988
- $\boldsymbol{a_i}$ contains factors such as average employees’ ability, capital, managerial skills: roughly constant over a 2-year period and **possibly correlated with both “grant” and “scrap rate”**
- Taking the first differences (1988-1987) we eliminate $\boldsymbol{a_i}$

$$\Delta scrap_i = \delta_0 + \beta_1 \Delta grant_i + \Delta u_i$$

& note that in this case $\Delta grant_i = grant_{i,1988}$

Data available here: https://ideas.repec.org/p/boc/bocins/jtrain.html

Policy Analysis using FD: Example

```
sort fcode year
by fcode: gen fd_lscrap = lscrap[_n]-lscrap[_n-1]
by fcode: gen fd_grant = grant[_n]-grant[_n-1]
reg fd_lscrap fd_grant if year == 1988
```

Not strictly necessary since for 1987 these variables will be missing

```
. reg fd_lscrap fd_grant if year == 1988

      Source |       SS           df       MS      Number of obs   =        54
-------------+----------------------------------   F(1, 52)        =      3.74
       Model |  1.23795555         1  1.23795555   Prob > F        =    0.0585
    Residual |  17.1971834        52  .330715065   R-squared       =    0.0672
-------------+----------------------------------   Adj R-squared   =    0.0492
       Total |  18.4351389        53  .34783281    Root MSE        =    .57508

------------------------------------------------------------------------------
   fd_lscrap |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    fd_grant |  -.3170579   .1638751    -1.93   0.058    -.6458974    .0117816
       _cons |  -.0574357    .097206    -0.59   0.557    -.2524938    .1376223
------------------------------------------------------------------------------
```

```
Summary for variables: fd_lscrap
     by categories of: grant

   grant |      mean
---------+-----------
       0 |  -.0574357
       1 |  -.3744936
---------+-----------
   Total |  -.1689931
---------------------
```

Note: Controls could matter!

$$\widehat{\beta_1} = \overline{\Delta y}_{Treated} - \overline{\Delta y}_{Control}$$

$$-0.3171 = -0.3745 - (-0.0574)$$

Because the treatment occurs in t=2!

Comparing FD with Pooled OLS

Running a "pooled OLS" instead:

$$scrap_{it} = \beta_0 + \delta_0 d2_t + \beta_1 grant_{it} + a_i + u_{it}$$

Source	SS	df	MS			
				Number of obs	=	108
				F(2, 105)	=	0.18
Model	.810536031	2	.405268016	Prob > F	=	0.8378
Residual	240.098945	105	2.28665662	R-squared	=	0.0034
				Adj R-squared	=	-0.0156
Total	240.909481	107	2.25149048	Root MSE	=	1.5122

lscrap	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
d2	-.1889081	.3281441	-0.58	0.566	-.8395572	.461741
grant	.0566004	.43091	0.13	0.896	-.7978145	.9110152
_cons	.597434	.2057802	2.90	0.005	.1894099	1.005458

- Can we trust OLS estimates? Remember that a_i is **included in the error term, and possibly correlated with *grant***.
- The difference between OLS and FD estimates indicate this correlation is indeed present, and that firms with lower-ability workers (or firms of lower quality on average) are more likely to receive the grant/subsidy (as expected)

Panel Data with T > 2

First-differences can still be applied as before – we could **difference adjacent periods** and with that eliminate a_i… [note that we lose the 1st period!]

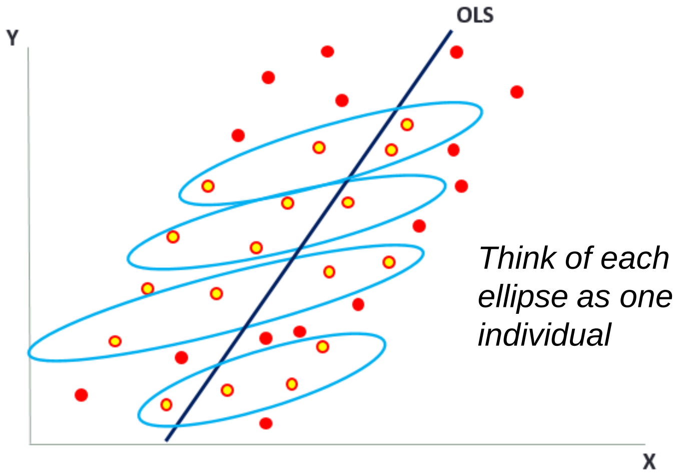
…but there are **more efficient approaches for this case**: **Within-estimation** (also called *Least Squares Dummy Variable (LSDV) estimation*)

Logic: A dummy for each individual

- Omit the intercept in the regression
- Allow each individual to have a different intercept

How to implement the **fixed-effects model**?

- Including a ”thousand” dummy variables
  - Drawback of this approach?
  - Alternative way? Data transformation…



Think of each ellipse as one individual

Data Transformation for FE Estimation

- Like in FD estimation, we transform the data to remove the unobserved effect $\boldsymbol{a_i}$ prior to any estimation.
- Note that **any constant explanatory variables (e.g., gender), will also be lost** along with $\boldsymbol{a_i}$.
- Suppose that for each "individual" i in our data:

$$y_{it} = a_i + \beta x_{it} + u_{it} \qquad (1)$$

- For each "individual" i, we can average this equation over time:

$$\bar{y}_i = a_i + \beta \bar{x}_i + \bar{u}_i \qquad (2)$$

- Subtracting (2) from (1): $y_{it} - \bar{y}_i = \beta(x_{it} - \bar{x}_i) + (u_{it} - \bar{u}_i)$ $\boldsymbol{a_i}$ is dropped!
- OLS model on: $\boldsymbol{\ddot{y}_{it} = \beta \ddot{x}_{it} + \ddot{u}_{it}}$ [no intercept!]
- OLS on time-demeaned data → **fixed effects (or within) estimator**; consistent and unbiased if all the x variables are exogenous

Fixed Effects versus Random Effects

Fixed Effects

- **Allows** for arbitrary **correlation between a_i and the explanatory variables** in any time period
- Does not affect consistency of the estimates since the data transformation eliminates a_i
- **Drawback:** no estimates for time-invariant regressors (they are included in a_i)
- Intercepts capture individual heterogeneity (a single intercept refers to the average across all $\hat{a}_i$)

Random Effects

- **DOES NOT allow** for arbitrary **correlation between a_i and the explanatory variables** in any time period
- **Why?** a_i keeps being part of the error term!
- Consistent only if $cov(a_i, x_{it}) = 0$
- **Advantage:** coefficients of time-invariant regressors can be estimated; also more efficient (saving degrees of freedom)

Hausman Test compares the two estimators. A rejection under Hausman test implies that RE key assumption is false, and we then use FE.

Random Effects Model

Remember that RE assumes no correlation between x and a_i

$$y_{it} = \beta_0 + (time\ dummies) + \beta_1 x_{1it} + \cdots + \beta_k x_{kit} + \boldsymbol{a_i} + u_{it}$$

Back to the **composite error term**: $v_{it} = a_i + u_{it}$

- Key assumption: $cov(a_i, x_{kij}) = 0, \qquad \mathrm{t} = 1, 2, \dots, \mathrm{T}; \qquad \mathrm{j} = 1, 2, \dots, \mathrm{k}$
- Data are still transformed, such that the equation being estimated is:

$$y_{it} - \theta\bar{y}_i = \beta_0(1 - \theta) + \beta_1(x_{1it} - \theta\bar{x}_1) + \cdots + (v_{it} - \theta\bar{v}_i)$$

where $$\theta = 1 - \frac{\sigma_u}{\sqrt{\sigma_u^2 + T\sigma_a^2}}$$

- The closer is $\hat{\theta}$ to zero (one), the closer will RE estimates be to pooled OLS (FE).
- The closer is $\hat{\theta}$ to zero, the less important is the unobserved effect $\boldsymbol{a_i}$.

Between Estimator vs. Within Estimator

Note that in panel data we have **two kinds of variations**

- Variation from observation to observation within a single ellipse **(within individuals)**
- Variation in observations from ellipse to ellipse **(between individuals)**

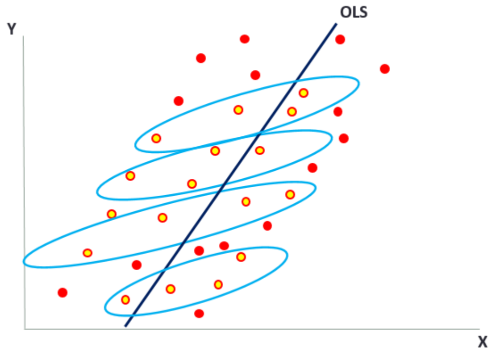
FE estimator: within estimator

OLS: between estimator (cross-sectional equation)

RE estimator: a (matrix-) weighted average of the two

- **Random Effects** = Fixed effects (within) + λ*between

$$\lambda = \frac{\sigma_u^2}{\sigma_u^2 + T\sigma_a^2}$$



Random Effects' Take-Aways:

☑ **Extra efficiency** associated with the RE estimator

- Using information from **both the within and the between estimators**

☑ Estimating coefficients of **time-invariant variables**

- Varying between ellipses (recall previous diagram)

✘ **Biased when there is correlation between explanatory variables and the composite error term**

– Between estimator is biased

Panel Data in Stata

What do panel data look like? How to reshape them?

- Long data vs. wide data
- `reshape` command

Set your data for panel data analysis

- `xtset id time`
- Describe with `xtsum`

(Linear) Panel data estimation commands

- Regression: xtreg (options fe or re)
- Note that there are other commands for binary dependent variables! *(out of our scope)*
- <u>Can be combined with other methods (e.g., IV or run panel data models on a matched sample)</u>

Example: Back to wages!

nlswork.dta

US women, 14-46 during the 1968-1988

Let’s see the effects of schooling, age, experience, tenure, race, and area of residence on their wages

- Data structure and description
- Compare different estimators

Panel Data Format

```
. list idcode year ln_wage grade age ttl_exp in 1/24
```

	idcode	year	ln_wage	grade	age	ttl_exp
1.	1	70	1.451214	12	18	1.083333
2.	1	71	1.02862	12	19	1.275641
3.	1	72	1.589977	12	20	2.25641
4.	1	73	1.780273	12	21	2.314102
5.	1	75	1.777012	12	23	2.775641
6.	1	77	1.778681	12	25	3.775641
7.	1	78	2.493976	12	26	3.852564
8.	1	80	2.551715	12	28	5.294872
9.	1	83	2.420261	12	31	5.294872
10.	1	85	2.614172	12	33	7.160256
11.	1	87	2.536374	12	35	8.98718
12.	1	88	2.462927	12	37	10.33333
13.	2	71	1.360348	12	19	.7115384
14.	2	72	1.206198	12	20	1.134615
15.	2	73	1.549883	12	21	1.461538
16.	2	75	1.832581	12	23	2.211539
17.	2	77	1.726721	12	25	3.211539
18.	2	78	1.68991	12	26	4.211538
19.	2	80	1.726964	12	28	6.096154
20.	2	82	1.808289	12	30	7.666667
21.	2	83	1.863417	12	31	8.583333
22.	2	85	1.789367	12	33	10.17949
23.	2	87	1.84653	12	35	12.17949
24.	2	88	1.856449	12	37	13.62179

- Data are already in the **long form** (desirable for estimation)
  - i.e., first observation is for individual 1 in year 1; second observation is for individual 1 in year 2
- **Unbalanced panel!** Why?
- Any **time-invariant variables**? Implications?
- Remember to declare your data as a panel: `xtset idcode year`

Panel Data Format

- 4711 individuals
- 15 years
- But not everyone appears is all years (**unbalanced panel**)
- In principle it has to do with random missing data (testable)
  - Let's assume so here
  - In the future take potential attrition bias into consideration if needed

```
. xtdescribe

  idcode:  1, 2, ..., 5159                                   n =       4711
    year:  68, 69, ..., 88                                   T =         15
           Delta(year) = 1 unit
           Span(year)  = 21 periods
           (idcode*year uniquely identifies each observation)

Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                         1       1       3         5         9      13      15

     Freq.  Percent    Cum. |  Pattern
 ---------------------------+-----------------------
      136      2.89    2.89 |  1....................
      114      2.42    5.31 |  ....................1
       89      1.89    7.20 |  .................1.11
       87      1.85    9.04 |  ...................11
       86      1.83   10.87 |  111111.1.11.1.11.1.11
       61      1.29   12.16 |  ..............11.1.11
       56      1.19   13.35 |  11...................
       54      1.15   14.50 |  ...............1.1.11
       54      1.15   15.64 |  .......1.11.1.11.1.11
     3974     84.36  100.00 | (other patterns)
 ---------------------------+-----------------------
     4711    100.00         |  XXXXXX.X.XX.X.XX.X.XX
```

Within vs. Between Variation

```
. xtsum idcode year ln_wage grade age ttl_exp

Variable         |      Mean   Std. Dev.       Min        Max |    Observations
-----------------+--------------------------------------------+----------------
idcode   overall |  2601.284   1487.359          1       5159 |     N =   28534
         between |              1487.57          1       5159 |     n =    4711
         within  |                    0   2601.284   2601.284 | T-bar = 6.05689
                 |                                            |
year     overall |  77.95865   6.383879         68         88 |     N =   28534
         between |             5.156521         68         88 |     n =    4711
         within  |             5.138271   63.79198   92.70865 | T-bar = 6.05689
                 |                                            |
ln_wage  overall |  1.674907   .4780935          0   5.263916 |     N =   28534
         between |              .424569          0   3.912023 |     n =    4711
         within  |               .29266  -.4077221    4.78367 | T-bar = 6.05689
                 |                                            |
grade    overall |  12.53259   2.323905          0         18 |     N =   28532
         between |             2.566536          0         18 |     n =    4709
         within  |                    0   12.53259   12.53259 | T-bar = 6.05904
                 |                                            |
age      overall |  29.04511   6.700584         14         46 |     N =   28510
         between |             5.485756         14         45 |     n =    4710
         within  |              5.16945   14.79511   43.79511 | T-bar = 6.05308
                 |                                            |
ttl_exp  overall |  6.215316   4.652117          0   28.88461 |     N =   28534
         between |             3.724221          0    24.7062 |     n =    4711
         within  |             3.484133  -9.642671   20.38091 | T-bar = 6.05689
```

- Total variation can be decomposed into **within variation** (over time for each individual) and **between variation** (across individuals)
- **Time-invariant variables have zero within variation**: consequences for FE estimation?

Within-estimation

Different s.e.! → note the difference in the degrees of freedom of the 2 models (why?)

OLS on the demeaned data vs FE on the original data

```
. * Run an OLS on the demeaned data:
. reg new_ln_wage new_grade new_age new_age_sq new_tenure new_ten_sq new_race2
note: new_grade omitted because of collinearity
note: new_race2 omitted because of collinearity
note: new_race3 omitted because of collinearity

      Source |       SS           df       MS      Number of obs   =    28,091
-------------+----------------------------------   F(6, 28084)     =    817.54
       Model |  356.514127         6  59.4190212   Prob > F        =    0.0000
    Residual |  2041.15265    28,084  .072680268   R-squared       =    0.1487
-------------+----------------------------------   Adj R-squared   =    0.1485
       Total |  2397.66678    28,090  .085356596   Root MSE        =    .26959

------------------------------------------------------------------------------
 new_ln_wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   new_grade |          0  (omitted)
     new_age |   .0419553   .0026101    16.07   0.000     .0368394    .0470711
  new_age_sq |  -.0004994   .0000432   -11.55   0.000    -.0005841   -.0004147
  new_tenure |   .0409226   .0014907    27.45   0.000     .0380008    .0438444
  new_ten_sq |  -.0014496   .0000968   -14.97   0.000    -.0016394   -.0012599
   new_race2 |          0  (omitted)
   new_race3 |          0  (omitted)
new_not_smsa |  -.0933711   .0088227   -10.58   0.000     -.110664   -.0760782
   new_south |  -.0618201   .0101196    -6.11   0.000     -.081655   -.0419853
       _cons |   .0004568   .0016085     0.28   0.776     -.002696    .0036097
------------------------------------------------------------------------------
```

Note the difference in the constant term – meaning?
Missing coefficients for some variables – why?

```
. * compare with FE on the original data:
. xtreg ln_wage grade age age_sq tenure ten_sq race2 race3 not_smsa south, fe
note: grade omitted because of collinearity
note: race2 omitted because of collinearity
note: race3 omitted because of collinearity

Fixed-effects (within) regression               Number of obs     =     28,091
Group variable: idcode                          Number of groups  =      4,697

R-sq:                                           Obs per group:
     within  = 0.1492                                         min =          1
     between = 0.3051                                         avg =        6.0
     overall = 0.2266                                         max =         15

                                                F(6,23388)        =     683.34
corr(u_i, Xb)  = 0.2232                         Prob > F          =     0.0000

------------------------------------------------------------------------------
     ln_wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       grade |          0  (omitted)
         age |   .0419768   .0028583    14.69   0.000     .0363743    .0475793
      age_sq |  -.0004996   .0000473   -10.56   0.000    -.0005924   -.0004069
      tenure |    .040913   .0016304    25.09   0.000     .0377173    .0441088
      ten_sq |  -.0014494   .0001059   -13.69   0.000    -.0016569   -.0012419
       race2 |          0  (omitted)
       race3 |          0  (omitted)
    not_smsa |  -.0934057    .009664    -9.67   0.000    -.1123477   -.0744637
       south |   -.062689   .0110836    -5.66   0.000    -.0844137   -.0409643
       _cons |   .8601684   .0416079    20.67   0.000     .7786143    .9417225
-------------+----------------------------------------------------------------
     sigma_u |  .37179939
     sigma_e |  .29477901
         rho |  .61402356   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(4696, 23388) = 6.88                 Prob > F = 0.0000
```

Individual (dummies) FE are jointly significant

Random Effects & Hausman Test

```
. xtreg ln_wage grade age age_sq tenure ten_sq race2 race3 not_smsa south, re

Random-effects GLS regression                   Number of obs     =     28,091
Group variable: idcode                          Number of groups  =      4,697

R-sq:                                           Obs per group:
     within  = 0.1472                                         min =          1
     between = 0.4484                                         avg =        6.0
     overall = 0.3461                                         max =         15

                                                Wald chi2(9)      =    7996.86
corr(u_i, X)   = 0 (assumed)                    Prob > chi2       =     0.0000

------------------------------------------------------------------------------
     ln_wage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       grade |   .0697135    .001805    38.62   0.000     .0661757    .0732512
         age |   .0414789   .0027294    15.20   0.000     .0361294    .0468284
      age_sq |  -.0005217   .0000452   -11.55   0.000    -.0006102   -.0004332
      tenure |    .046826   .0015529    30.15   0.000     .0437824    .0498696
      ten_sq |  -.0015455   .0001017   -15.20   0.000    -.0017448   -.0013461
       race2 |  -.0520281   .0102026    -5.10   0.000    -.0720249   -.0320314
       race3 |   .0697214   .0412066     1.69   0.091    -.0110421    .1504849
    not_smsa |  -.1335169   .0073019   -18.29   0.000    -.1478283   -.1192054
       south |  -.0882823   .0074406   -11.86   0.000    -.1028656   -.0736989
       _cons |   .0243543   .0446233     0.55   0.585    -.0631057    .1118144
-------------+----------------------------------------------------------------
     sigma_u |  .26251702
     sigma_e |  .29477901
         rho |  .44230314   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

```
. hausman FE RE

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |       FE           RE         Difference          S.E.
-------------+----------------------------------------------------------------
         age |    .0419768     .0414789        .0004979        .0008489
      age_sq |   -.0004996    -.0005217        .0000221        .0000142
      tenure |     .040913      .046826       -.005913        .0004969
      ten_sq |   -.0014494    -.0015455        .0000961        .0000294
    not_smsa |   -.0934057    -.1335169        .0401112        .0063305
       south |    -.062689    -.0882823        .0255933        .0082149
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from xtreg
            B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(6) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =      280.37
                Prob>chi2 =      0.0000
```

Note that the comparison can only be made for coefficients that are estimated with BOTH estimators. H0: The coefficients are the same is rejected.

⇔

We reject the hypothesis that RE provides consistent estimates

Just one more week to go!