Week 6: Panel data & methods

Advanced Econometrics 4EK608

Vysoká škola ekonomická v Praze

Content

Pooled cross sections

- **Pooled cross sections data:** Random sampling from a large population at different time periods. For example: 400 randomly selected respondents in a survey, 5 consecutive years. For each year, we have a different - randomly chosen - set of respondents.
- Pooled cross sections should not be confused with "actual" panel data (where we would follow individual respondents across time).
- Pooled cross sections: sampling from a changing population at different points in time generates **independent, not identically distributed** (*inid*) observations.
- Pooled cross sections are easy to deal with, simply by allowing the intercept (and perhaps some selected slopes) in a LRM to vary across time.
- Can be used for policy analysis (difference-in-differences estimator).

Pooled cross sections

Pooled cross sections - model example

$$\log(wage_{it}) = \theta_0 + \theta_1 d91_t + \theta_2 d92_t + \delta_1 female_{it} + \delta_2 educ_{it} +$$
$$+ \gamma_1 exper_{it} + \gamma_2(female \times d91)_{it} + \gamma_3(female \times d92)_{it} + u_{it}$$

where $t = 1990, 1991, 1992;$ $i = 1, 2, \ldots, 500$ ←

> Each year, we draw 500 individuals at random. Individual respondents are not followed. Total observations: $N \times T = 1.500$

$d91_t$ and $d92_t$ are time dummies,

$female_{it}$, $educ_{it}$ and $exper_{it}$ describe the gender, education and work experience of the i-th individual at time t,

$(female \times d91)_{it}$ is an interaction element, may be used to describe whether changes in wages over time are statistically different for man and woman.

Pooled cross sections: Chow test

Pooled cross sections - model example contd.

$$\log(wage_{it}) = \beta_0 + \beta_1 d91_t + \beta_2 d92_t + \beta_3 female_{it} + \\ + \beta_4 educ_{it} + \beta_5 exper_{it} + u_{it}$$

Chow test for structural changes across time

Basically an F-test for linear restrictions, can be used to determine whether the estimated slope coefficients change across time.

In our $\log(wage)$ equation, we would test the H_0 of "time-invariant" β_3, β_4 and β_5 coefficients, while allowing for time dummies (time-specific intercepts).

Pooled cross sections: Chow test

SSR_r: restricted model – pooled regression, allowing for different time intercepts.

SSR_{ur}:run a regression for each of the time periods. $SSR_{ur} = SSR_1 + SSR_2 + \cdots + SSR_T$
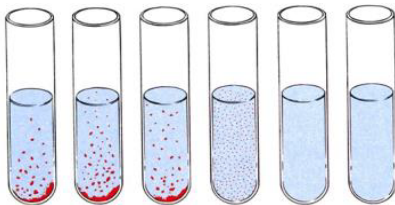
$T + Tk$ parameters estimated in the unrestricted model

$$F = \frac{(SSR_r - SSR_{ur})}{SSR_{ur}} \cdot \frac{(n - T - Tk)}{(T-1)k};$$

under H_0 of no structural break, $F \sim F((T-1)k, (n - T - Tk))$

Note: This test is not robust to heteroskedasticity (including changing variance across time). Robust variants of the test exist, based on interaction terms.

Policy analysis with pooled cross sections

Scientific experiment

- Test tubes identical except for catalyst
- Measure: Effect at different catalyst volumes (reaction speed, product volume, . . .)
- Perform the experiment n-times
- Control for other factors (heat, . . .)
- Estimate average effect (& standard error)

Natural experiment (quasi-experiment)


- Garbage incinerator is built in one given suburban area over time
- How do we estimate the effect on individual house-prices?
- Identical control group does not exist. . .
- DiD: Difference-in-Differences estimator (assumptions apply!)

Policy analysis with pooled cross sections

DiD example: In-house employee training for women returning from maternal leave & its wage effect
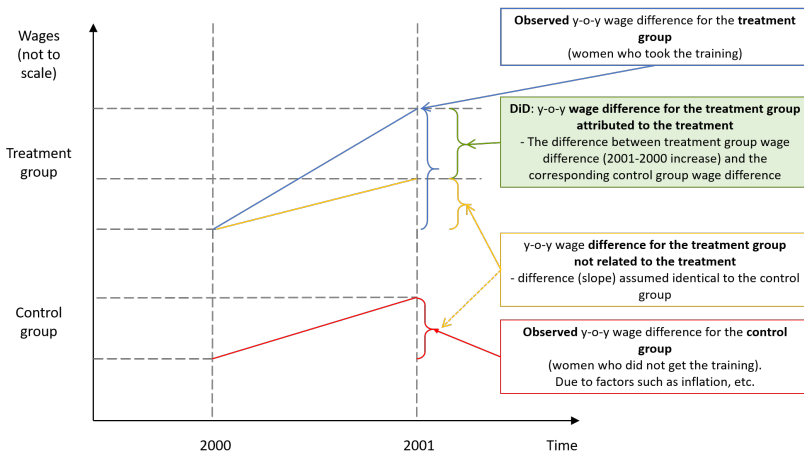
Policy analysis with pooled cross sections

DiD estimator: we can use LRMs to compare the changes in conditional means for the treatment and control groups...

- Group specific and time specific effects are allowed (controlled for)

Assumptions:

- Unbiased DiD estimates require that the treatment (being subject to economic policy change...) is not systematically related to factors affecting the outcome (dependent variable) that are not accounted for explicitly in our model and thus are "hidden" in the random element.

- DiD **attributes all differences in trends** between the treatment and control groups **to the intervention** (treatment). We assume there are no other factors that affect the difference in trends between the two groups.

Example: policy analysis with pooled cross sections

$$y_{it} = \beta_0 + \delta_0 d2 + \beta_1 dT + \delta_1(d2 \times dT) + \boldsymbol{x}_{it}\boldsymbol{\gamma} + u_{it},$$

$$i = 1, \ldots, N; \quad t = 1, 2.$$

where:

- $d2$ is a dummy variable, $d2 = 1$ for the second period (post treatment),
- dT is a dummy variable, equals 1 for the individuals in the treatment group,
- $\boldsymbol{x}_{it}$ is a $1 \times k$ (row) vector of additional regressors and $\boldsymbol{\gamma}$ is a $k \times 1$ vector of coefficients.

Example: policy analysis with pooled cross sections

$$y_{it} = \beta_0 + \delta_0 d2 + \beta_1 dT + \delta_1(d2 \times dT) + u_{it},$$
$$i = 1, \ldots, N; \quad t = 1, 2.$$

In this simplified model (we drop $\boldsymbol{x}_{it}\boldsymbol{\gamma}$), the estimated δ_1 has a convenient DiD interpretation:

$$\hat{\delta}_1 = (\overline{y}_{Tr,\, t=2} - \overline{y}_{Co,\, t=2}) - (\overline{y}_{Tr,\, t=1} - \overline{y}_{Co,\, t=1}),$$

which may be rearranged as:

$$= (\overline{y}_{Tr,\, t=2} - \overline{y}_{Tr,\, t=1}) - (\overline{y}_{Co,\, t=2} - \overline{y}_{Co,\, t=1})$$

Example: policy analysis with pooled cross sections

Table 1: Illustration of the DiD estimator

$E(y_{it}|d2, dT)$	Before $(t=1)$	After $(t=2)$	After - Before
Control	β_0	$\beta_0 + \delta_0$	δ_0
Treatment	$\beta_0 + \beta_1$	$\beta_0 + \delta_0 + \beta_1 + \delta_1$	$\delta_0 + \delta_1$
Treatment - Control	β_1	$\beta_1 + \delta_1$	δ_1

Even if $\boldsymbol{x}_{it}\boldsymbol{\gamma}$ is added back to the equation, interpretation of δ_1 remains essentially unchanged.

Example: policy analysis with pooled cross sections

What is the effect of building garbage incinerator on housing prices?

Dependent Variable: RPRICE
Included observations: 321

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	82517.23	2726.910	30.26034	0.0000
Y81	18790.29	4050.065	4.639502	0.0000
NEARINC	-18824.37	4875.322	-3.861154	0.0001
Y81*NEARINC	-11863.90	7456.646	-1.591051	0.1126

R-squared	0.173948	Mean dependent var	83721.36
Adjusted R-squared	0.166131	S.D. dependent var	33118.79
S.E. of regression	30242.90	Akaike info criterion	23.48429
Sum squared resid	2.90E+11	Schwarz criterion	23.53129
Log likelihood	-3765.229	Hannan-Quinn criter.	23.50306
F-statistic			
Prob(F-statistic)			

PRICE - house price in real terms (USD)
$Y81$ – dummy variable for 1981, $(t = 1978, 1981)$
1978 – before "rumors" ; 1981 – incinerator operational
NEARINC – dummy for the treatment group

Example: policy analysis with pooled cross sections

Incinerator effect on prices example, contd:

The model may be easily expanded by explanatory variables such as: *HOUSE.AGE, ROOMS, AREA, LOT.AREA*, etc. and the DiD interpretation remains basically unchanged ...

Selection bias (treatment effect vs. selection bias) example:

Assumption: "Unbiased DiD estimates require that the treatment is not systematically related to factors affecting the outcome that are not explicitly accounted for."

Say, we have a "poor neighborhood" with relatively old and small houses and low house-prices. For complex reasons, it suffers from a representation deficit within the local city council (as compared to other "rich neighborhoods") and is therefore more likely to get the incinerator.

We do not have variables to control for this factor $\rightarrow$ the DiD estimator may be severely biased

Policy analysis with pooled cross sections

Treatment effects

Topic not fully covered here, for detailed information see:

1. Wooldridge: Econometric analysis of C-S and panel data, chapter 21 Estimating Average Treatment Effects

2. Greene: Econometric analysis, chapter 19.6

3. Angrist, Pischke: Mostly Harmless Econometrics

Panel data

- N individual CS units are followed over T time periods
- Short panels: $N \gg T$
  Working with short panels is similar to CS data analysis. If CS units are randomly drawn from a population and T is small and fixed, then asymptotic analysis (properties) hold for arbitrary time dependence and distributional heterogeneity across time.
- Long panels: $T \gg N$
  Working with long panels is similar to time-series analysis. In TS analysis, stationarity & weak dependency conditions apply. SURE (Seemingly Unrelated Regression Equation) approach can be used: for the regression equations (with identical regressor structure), we estimate contemporaneous error covariances and use this information to improve efficiency of the estimate (see Greene, chapter 10.2)
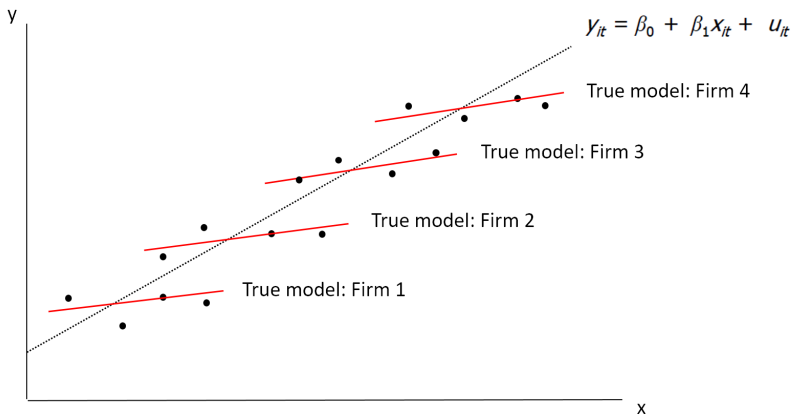
Panel data

- Large panel datasets: T and N large
  Both CS and TS analysis assumptions apply, specialized estimators exist for large (heterogeneous) panels. Cointegrated series in panels: estimation and tests by Pesaran.

- **Balanced panels:** obs. available for all T on all CS units. Often assumed for simplicity of interpretation.

- **Unbalanced panels:** mechanics of coefficient estimation do not differ. Model interpretation may require formal description of why the panel may be unbalanced. Problems may be caused by:
  - **Sample selection bias:** with e.g. self-selection, coefficients can be be biased and inconsistent.
  - **Attrition bias:** even if participants are randomly selected at the beginning of observation, they often leave (medical study, school, etc.) on a non-random basis.

Panel data

Pooled regression with panel data:
Heterogeneity bias
(Similar principle as the Simpson's paradox)

Panel data

Variation for the dependent variable and regressors

- Overall variation: variation over time and individuals.
- Between variation: variation between individuals.
- Within variation: variation within individuals (over time).

Id	Time	Variable	Individual mean	Overall mean	Overall deviation	Between deviation	Within deviation	Within deviation (modified)
i	t	x_{it}	$\overline{x}_i$	$\overline{x}$	$x_{it} - \overline{x}$	$\overline{x}_i - \overline{x}$	$x_{it} - \overline{x}_i$	$x_{it} - \overline{x}_i + \overline{x}$
1	1	9	10	20	-11	-10	-1	19
1	2	10	10	20	-10	-10	0	20
1	3	11	10	20	-9	-10	1	21
2	1	20	20	20	0	0	0	20
2	2	20	20	20	0	0	0	20
2	3	20	20	20	0	0	0	20
3	1	25	30	20	5	10	-5	15
3	2	30	30	20	10	10	0	20
3	3	35	30	20	15	10	5	25

Panel data

Panel data model - example

Unobserved individual effect, constant over time

Random element

$$\log(wage_{it}) = \beta_0 + \beta_1 trend_t + \beta_2 educ_{it} + a_i + u_{it}$$

Panel data model - a general notation

$$y_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta} + a_i + u_{it}$$

where $t = 1, 2, \ldots, T; \quad i = 1, 2, \ldots, N,$

$\boldsymbol{x}_{it}$ is a $1 \times k$ (row) vector

$\boldsymbol{\beta}$ is a $k \times 1$ vector

a_i unobserved effect, unobserved heterogeneity, individual effect, firm effect, etc,

u_{it} the usual random element.

Panel data

"Population version" of the panel data model in conditional expectation form:

(individual subscripts omitted)

$E(y_t|\boldsymbol{x_t}, a) = \boldsymbol{x_t}\boldsymbol{\beta} + a$

Therefore $\beta_j = \frac{\partial E(y_t|x_t, a)}{\partial x_{tj}}$ is the partial effect of the j-th explanatory variable on y_t (while holding a fixed).

This model may appear restrictive because $\boldsymbol{\beta}$ is time-invariant (the same in each time period).

However, by appropriately choosing x_{it}, we can allow for regression parameters to change over time.

Panel data

Panel data model - a structured notation

$y_{it} = \boldsymbol{g}_t\boldsymbol{\theta} + \boldsymbol{z}_i\boldsymbol{\delta} + \boldsymbol{w}_{it}\boldsymbol{\gamma} + a_i + u_{it}$

where $t = 1, 2, \ldots, T; \quad i = 1, 2, \ldots, N,$

$\boldsymbol{g}_t$ is a vector of aggregate time effects (often time dummies),

$\boldsymbol{z}_i$ is a set of time-constant observed variables,

$\boldsymbol{w}_{it}$ changes across i and t (for at least some units i and time periods t), can include interactions among time-constant and time varying variables,

$\boldsymbol{\theta}, \boldsymbol{\delta}$ and $\boldsymbol{\gamma}$ - regression coefficients

Panel data

Panel data model - a structured notation example

$$\log(wage_{it}) = \theta_0 + \theta_1 d91_t + \theta_2 d92_t + \delta_1 female_i + \delta_2 educ_i +$$
$$+ \gamma_1 exper_{it} + \gamma_2 (female \times exper)_{it} + a_i + u_{it}$$

Where $t = 1990, 1991, 1992; \quad i = 1, 2, \ldots, 100.$
For a balanced panel, $T \times N = 300$

We follow 100 individuals across three years.

$d91_t$ and $d92_t$ are time dummies,
$female_i$ and $educ_i$ do not change over time
(individuals in our dataset are not active students ...),
$exper_{it}$ changes between individuals and across time periods,
$(female \times exper)_{it}$ is an interaction element, changes between
individuals and across time.

LSDV regression

In the model $y_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta} + a_i + u_{it}$,

a_i are usually regarded as unobservable variables.
This approach gives appropriate interpretation of $\boldsymbol{\beta}$.
Traditional (old) approaches to fixed effects estimation view the a_i as parameters to be estimated along with $\boldsymbol{\beta}$.

How to estimate a_i values along with $\boldsymbol{\beta}$?

- Define N dummy variables - one for each cross-section.
- Convenient LSDV model expansion: use interactions to control for individual slopes for chosen regressors.
  Sample interaction element: $\delta_1(ind1_i \times x_{1,it})$

δ_1 measures the difference in x_1 of slope for $ind1$	Dummy, $ind1_i = 1$ for individual 1($i = 1$) and zero otherwise	Regressor x_1 for individual i at time t

LSDV regression - example

$$y_{it} = \alpha_1 ind1_i + \alpha_2 ind2_i + \cdots + \alpha_N indN_i + \beta_1 x_{it1} + \cdots + \beta_k x_{itk} + u_{it}$$

Dummy equals 1 only if observations (time-invariant) relate to i-th C-S unit.

- $\hat{\boldsymbol{\beta}}_{LSDV}$ is identical to $\hat{\boldsymbol{\beta}}_{FE}$ (explained next) - it is a consistent estimator of $\boldsymbol{\beta}$ if we hold T fixed and $N \to \infty$ (atually, consistency applies more generally).
- For $\hat{\boldsymbol{\alpha}}$ (vector of individual $\hat{\alpha}_i$ values), such consistency does not hold: as $N \to \infty$, information does not accumulate for a_i.

FD estimator

We can simply eliminate unobserved heterogeneity from the regression: $y_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta} + a_i + u_{it}$
by first differences (FD) transformation:
$\Delta y_{it} = y_{it} - y_{i,t-1} = \Delta\boldsymbol{x}_{it}\boldsymbol{\beta} + \Delta a_i + \Delta u_{it} = \Delta\boldsymbol{x}_{it}\boldsymbol{\beta} + \Delta u_{it}$

  ✓ Removes any unobserved heterogeneity.

  × We remove all time-invariant factors in $\boldsymbol{x}$.
    If the time-invariant regressors are of no interest, this is
    a robust estimator.

Estimation can be done with FGLS (autocorrelation of transformed residuals), or OLS with HAC robust errors.

FD is most suitable when we have $t = 1; 2$ – two period panel (FD may be used with more time periods, we have $N(T-1)$ observations after differencing)

FD estimator – more than two time periods

$$y_{it} = \delta_1 + \delta_2 d2_t + \delta_3 d3_t + \boldsymbol{x}_{it}\boldsymbol{\beta} + a_i + u_{it};$$

$$\text{3 periods } (t = 1, 2, 3), \text{ 3N total observations}$$

First differences (FD) transformation:

$$\Delta y_{it} = \delta_2 \Delta d2_t + \delta_3 \Delta d3_t + \Delta \boldsymbol{x}_{it}\boldsymbol{\beta} + \Delta u_{it} \qquad (1)$$

- We only have data for $t = 2, 3$. i.e. we have $N(T-1)$ observations after differencing.
- If G-M assumptions are satisfied, we use pooled OLS for estimation.
- For $t = 2$, $\Delta d2_t = 1$ and $\Delta d3_t = 0$.
  For $t = 3$, $\Delta d2_t = -1$ and $\Delta d3_t = 1$.
- FD equation does not contain intercept
  ($\ldots R^2$ calculation & other complications)

FD estimator – more than two time periods

Unless time intercepts are of direct interest, we transform the FD equation so that it contains intercept and -undifferenced- time dummies (usually, we leave out $d2$)

$$\Delta y_{it} = \alpha_0 + \alpha_3 d3_t + \Delta \boldsymbol{x}_{it}\boldsymbol{\beta} + \Delta u_{it} \qquad (2)$$

this may be generalized for $T > 3$:

$$\Delta y_{it} = \alpha_0 + \alpha_3 d3_t + \alpha_4 d4_t + \cdots + \alpha_T dT_t + \Delta \boldsymbol{x}_{it}\boldsymbol{\beta} + \Delta u_{it}$$

Both (1) & (2) model specifications lead to identical $\hat{\boldsymbol{\beta}}$.

FD estimator – assumptions

FD.1 Functional form: $y_{it} = \beta_1 x_{it1} + \cdots + \beta_k x_{itk} + a_i + u_{it}$,
$i = 1, \ldots, N, \ \ t = 1, \ldots, T$

FD.2 We have random sample from cross-sectional units.

FD.3 Each regressor changes in time at least for some i and no perfect linear combination exists among regressors.

FD.4 For each i and t, $E(u_{it} \mid \boldsymbol{X}_i, a_i) = 0$. [Alt.: regressors are strictly exogenous conditional on unobserved effects: $corr(x_{itj}, u_{is} \mid a_i) = 0, \quad \forall \ t, s$]

FD.5 Variance of differenced errors conditional on all regressors is constant: $var(\Delta u_{it} \mid \boldsymbol{X}_i) = \sigma^2, \quad t = 2, 3, \ldots, T$. [homoskedasticity]

FD.6 No serial correlation exists among differenced errors. $cov(\Delta u_{it}, \Delta u_{is} \mid \boldsymbol{X}_i) = 0, \quad t \neq s$

FD.7 Differenced errors are normally distributed conditional on all regressors $\boldsymbol{X}_i$.

FD estimator – assumptions

Under **FD.1 - FD.4**
FD estimator is unbiased.
FD estimator is consistent for fixed T as $N \to \infty$.
For unbiasedness, $E(\Delta u_{it} \mid \boldsymbol{X}_i) = 0$ (for $t = 2, 3, \dots$) is
sufficient (instead of FD.4)

Under **FD.1 - FD.6**
FD estimator is BLUE (conditional on explanatory variables).
Asymptotic inference for FD estimator holds (t and F statistics
asymptotically follow corresponding distributions).

Under **FD.1 - FD.7**
FD estimator is BLUE (conditional on explanatory variables).
FD estimators - i.e. pooled OLS on first differences - are
normally distributed (t and F statistics have exact t and F
distributions).

FD estimator example

$$crmrte_{it} = \beta_0 + \delta_0 \, d87_{it} + \beta_1 \, unem_{it} + a_i + u_{it},$$

$$t = 1982, 1987$$

Dummy for the second time period

Model expanded: written separately for each time period

$$crmrte_{i1987} = \beta_0 + \delta_0 \cdot 1 + \beta_1 \, unem_{i1987} + a_i + u_{i1987}$$

$$crmrte_{i1982} = \beta_0 + \delta_0 \cdot 0 + \beta_1 \, unem_{i1982} + a_i + u_{i1982}$$

FD applied

$$\Rightarrow \Delta crmrte_i = \delta_0 + \beta_1 \, \Delta unem_i + \Delta u_i$$

δ_0 has a time effect interpretation

Individual unobserved effect disappears

$$\widehat{\Delta crmrte} = 15.40 + 2.22 \, \Delta unem$$

$$(4.70) \quad (.88)$$

With OLS estimation, HAC errors should be used

FD estimator

Problems related to the FD estimator:

- First-differenced estimates will be imprecise if explanatory variables vary only to a small extent over time (no estimate possible if regressors are time-invariant).
- Potentially, there is insufficient (lower) variability in differenced variables.
- Without strict exogeneity of regressors (e.g. in the case of a lagged dependent variable /say, $y_{i,t-1}$/ among regressors or with measurement errors), adding further periods does not reduce inconsistency.
- FD estimator may be worse than pooled OLS if explanatory variables are subject to measurement errors (errors in variables - EIV).

FE estimator

$y_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta} + a_i + u_{it}$

If a_i is unobserved, but correlated with $\boldsymbol{x}_{it}$,
then OLS on the observed variables

$$y_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta} + u_{it}$$

is biased and inconsistent (omitted variable bias).

Fixed effect (FE): yet another method for eliminating a_i from
the panel data model.

"Fixed" means correlation of a_i and $\boldsymbol{x}_{it}$, not that a_i is
non-stochastic.

FE estimator

$N \times T$ observations

We can rewrite $y_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta} + a_i + u_{it}$ as follows:

$y_{it} = \beta_1 x_{it1} + \cdots + \beta_k x_{itk} + a_i + u_{it}, \ i = 1, \ldots, N, \ t = 1, \ldots, T$

Now, <u>for each i</u>, we <u>average</u> the above equation <u>over time</u>:

$\overline{y}_i = \beta_1 \overline{x}_{i1} + \cdots + \beta_k \overline{x}_{ik} + \overline{a}_i + \overline{u}_i$ ← N equations with individual averages

By subtracting individual averages from the original observations (time-demeaning), we get:

$\Rightarrow [y_{it} - \overline{y}_i] = \beta_1 [x_{it1} - \overline{x}_{i1}] + \cdots + \beta_k [x_{itk} - \overline{x}_{ik}] + [u_{it} - \overline{u}_i]$

Alternative notation: $\ddot{y}_{it} = \ddot{\boldsymbol{x}}_{it}\boldsymbol{\beta} + \ddot{u}_{it}$; where $\ddot{y}_{it} = y_{it} - \overline{y}_i$, etc.

FE estimator, denoted $\hat{\boldsymbol{\beta}}_{FE}$, is the pooled OLS estimator applied to time-demeaned data.

FE estimator

FE estimator: by time demeaning, we get rid of the a_i element - as it does not vary over time

- $a_i = \overline{a}_i \ \rightarrow \ a_i - \overline{a}_i = 0$
- Intercept and all time-invariant regressors are also eliminated using the FE (within) transformation.

After FE estimation, a_i elements may be estimated as follows:
$\hat{a}_i = \overline{y}_i - \hat{\beta}_1 \overline{x}_{i1} - \cdots - \hat{\beta}_k \overline{x}_{ik}, \ \ i = 1, \ldots, N$

However, in most practical applications, a_i values bear limited useful information.

FE estimator

Degrees of freedom for the FE estimator:

- We have $N \times T$ observations for the equation $\ddot{y}_{it} = \ddot{\boldsymbol{x}}_{it}\boldsymbol{\beta} + \ddot{u}_{it}$ but the df of the FE estimator are NOT equal to $(N \times T) - k$

- For each C-S observation i, we loose one df.
  ...for each i, the demeaned errors $\ddot{u}_{it}$ add up to zero when summed over time.
- Hence $df = N(T-1) - k$

FE estimator – assumptions

FE.1 Functional form: $y_{it} = \beta_1 x_{it1} + \cdots + \beta_k x_{itk} + a_i + u_{it}$,
$i = 1, \ldots, N, \ t = 1, \ldots, T$

FE.2 We have random sample from cross-sectional units.

FE.3 Each regressor changes in time at least for some i and no perfect linear combination exists among regressors.

FE.4 For each i and t, $E(u_{it} \mid \boldsymbol{X}_i, a_i) = 0$. [Alt.: regressors are strictly exogenous conditional on unobserved effects: $corr(x_{itj}, u_{is} \mid a_i) = 0, \quad \forall \ t, s$]

FE.5 Variance of errors conditional on all regressors is constant: $var(u_{it} \mid \boldsymbol{X}_i, a_i) = var(u_{it}) = \sigma_u^2, \quad t = 1, 2, \ldots, T$. [homoskedasticity]

FE.6 No serial correlation exists among idiosyncratic errors. $cov(u_{it}, u_{is} \mid \boldsymbol{X}_i, a_i) = 0, \quad t \neq s$

FE.7 Errors are normally distributed conditional on all regressors $(\boldsymbol{X}_i, a_i)$.

FE estimator – assumptions

Under **FE.1 - FE.4** (identical to **FD.1 - FD.4**)
FE estimator is unbiased.
FE estimator is consistent for fixed T as $N \rightarrow \infty$.

Under **FE.1 - FE.6**
FE estimator is BLUE.
FD is unbiased
... **FE.6** makes FE better (less variance) than FD.
Asymptotically valid inference for FE estimator holds (t and F).

Under **FE.1 - FE.7**
FE estimator is BLUE and t and F statistics have exact t and F distributions.
FE estimators - i.e. pooled OLS on time demeaned data - are normally distributed.

FE estimator – example

Example: Effect of training grants on firm scrap rate

$$scrap_{it} = \beta_1 d88_{it} + \beta_2 d89_{it} + \beta_3 grant_{it} + \beta_4 grant_{it-1} + \boxed{a_i} + u_{it}$$

> Time-invariant reasons why one firm is more productive than another are controlled for. The important point is that these may be correlated with other explanatory variables.

> Stars denote time-demeaning

Fixed-effects estimation using the years 1987, 1988, 1989:

$$\widehat{scrap}_{it}^* = -.080\ d88_{it}^* - .247\ d89_{it}^* - .252\ grant_{it}^* - .422\ grant_{it-1}^*$$
$$\quad\quad\quad\quad (.109) \quad\quad\quad (.133) \quad\quad\quad (.151) \quad\quad\quad (.210)$$

$$n = 162,\ R^2 = .201$$

> Training grants significantly improve productivity (with a time lag)

FE estimator

Fixed effects testing - test for poolability

H_0: $a_i = 0$; $\forall i$

H_1: $\neg H_0$

This test evaluates the joint significance of the cross-section effects using sums-of-squares (F-test) and/or the likelihood function (χ^2 test).

(restricted model lacks individual effects, but includes intercept)

Alternative FE-redundancy tests can be based on
H_0: $var(a_i) = 0$;   see Wooldridge (2010)

FE estimator

Within estimator vs. **Between estimator**

- **Within estimator**
  For equation $\ddot{y}_{it} = \ddot{\boldsymbol{x}}_{it}\boldsymbol{\beta} + \ddot{u}_{it}$, the FE estimator (pooled OLS on time-demeaned data) is often called within estimator, as it uses variation within each cross-section.

- **Between estimator**
  Is obtained as the OLS estimation of
  $\overline{y}_i = \beta_1 \overline{x}_{i1} + \cdots + \beta_k \overline{x}_{ik} + \overline{a}_i + \overline{u}_i$ (after adding intercept)

  The between estimator uses only variation between the cross-section observations (ignores information on how the variables change over time).

- $\hat{\boldsymbol{\beta}}_{Between}$ is not consistent if a_i is correlated with $\boldsymbol{X}_i$,
  If we can reasonably assume no correlation between $\boldsymbol{X}_i$ and a_i, we use the RE estimator (explained next).

FE vs FD estimator

- For $T = 2$, FE and FD estimators produce identical estimates and inference. (FE must include a time dummy for the second period to be actually identical to the FD estimation output)
- For $T > 2$, FE and FD are both unbiased under FE.1 - FE.4. Both FE and FD are consistent for fixed T as $N \rightarrow \infty$
- If u_{it} is not serially correlated, FE is more efficient than FD
- If u_{it} follows a random walk (hence Δu_{it} is serially uncorrelated) FD is better than FE.
- If u_{it} shows some level of positive serial correlation (not a random walk), FD and FE may not be easily compared. For negative correlation of u_{it}, we prefer FE.

FE vs FD estimator

- For $T \gg N$, especially if non-stationary series are involved, FE may lead to spurious regression problems, while the FD help us transforming integrated series into weakly dependent series.

- If strict exogeneity is violated, both FE and FD are biased. However, FE is likely to have less bias than FD (unless $T = 2$). The bias of FD does not depend on T, while the bias in FE tends to zero at rate $1/T$.

- ... it may be a good idea to use both FD and FE. If the results are not method-sensitive, so much the better. If the results from FE and FD differ significantly, we sometimes report both.

RE estimator

If the individual unobserved effects a_i are strictly uncorrelated with $\boldsymbol{x}_{it}$, then it may be appropriate to model the individual constant terms as randomly distributed across cross-sectional units (appropriate if C-S units are from a large sample).

- Random effects (RE) model will greatly reduce the number of parameters to be estimated.

- But the RE estimator may yield potentially inconsistent estimates, if the above assumption is not correct.

RE estimator

$y_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta} + a_i + u_{it}$

If we can assume that a_i is uncorrelated with each explanatory variable: $cov(\boldsymbol{x}_{it}, a_i) = 0; \;\; t = 1, 2, \ldots, T$
then we may drop a_i from the equation and β_j estimates will remain unbiased.

By dropping a_i from the regression, we effectively create a new error term: $v_{it} = a_i + u_{it}$

As a_i is time-invariant, the random element v_{it} contains a lot of "inertia", i.e. autocorrelation (unless $a_i = 0$).

RE estimator

$$y_{it} = \beta_0 + \beta_1 x_{1,it} + \cdots + \beta_k x_{k,it} + v_{it};$$

where $v_{it} = a_i + u_{it}$ and $cov(\boldsymbol{x}_{it}, a_i) = 0; \ t = 1, 2, \ldots, T$

- The above equation may be estimated by **OLS, LSDV, within** and **between** estimators. All estimators mentioned are consistent but inefficient (due to autocorrelation in v_{it}).

- Under special conditions (RE assumptions), a consistent and asymptotically efficient FGLS-based estimation method exists: the **RE estimator**.
  (More asymptotically efficient than pooled OLS or FE estimator, for T fixed as $N \to \infty$.)

RE estimator - FGLS

$y_{it} = \beta_0 + \beta_1 x_{it1} + \cdots + \beta_k x_{itk} + v_{it};$

The quasi-demeaning (quasi-differencing) parameter λ is used for the FGLS estimation:

$\lambda = 1 - \left[\sigma_u^2/(\sigma_u^2 + T\sigma_a^2)\right]^{1/2}, \quad 0 \le \lambda \le 1$

where $var(a_i) = \sigma_a^2; \quad var(u_i) = \sigma_u^2$

- For each dataset, consistent estimators of σ_a^2 and σ_u^2 are available.
- Their estimation is based on pooled OLS or FE
  also, we use the fact that $\sigma_v^2 = \sigma_a^2 + \sigma_u^2$

RE estimator is a pooled OLS used on the quasi-demeaned data:

$$[y_{it} - \lambda \overline{y}_i] = \beta_1 [x_{it1} - \lambda \overline{x}_{i1}] + \cdots + \beta_k [x_{itk} - \lambda \overline{x}_{ik}] + [a_i - \lambda \overline{a}_i + u_{it} - \lambda \overline{u}_i]$$

Data are transformed

The transformed error follows G-M assumptions (not autocorrelated)

RE estimator - FGLS

$$[y_{it} - \lambda\overline{y}_i] = \beta_1 [x_{it1} - \lambda\overline{x}_{i1}] + \cdots + \beta_k [x_{itk} - \lambda\overline{x}_{ik}] + [a_i - \lambda\overline{a}_i + u_{it} - \lambda\overline{u}_i]$$

Interestingly, the FGLS equation is a general form that encompasses both FE and pooled OLS:

$$\hat{\lambda} \to 1 \quad \to \quad \text{RE} \to \text{FE}$$
$$\hat{\lambda} \to 0 \quad \to \quad \text{RE} \to \text{Pooled}$$

RE estimator – Assumptions

FE.1 Functional form: $y_{it} = \beta_1 x_{it1} + \cdots + \beta_k x_{itk} + a_i + u_{it}$, $i = 1, \ldots, N$, $t = 1, \ldots, T$

FE.2 We have random sample from cross-sectional units.

FE.4 $\forall \, i, t$: $E(u_{it} \mid \boldsymbol{X}_i, a_i) = 0$. [Alt.: $corr(x_{itj}, u_{is} \mid a_i) = 0$, $\forall \, t, s$]

FE.5 Variance of idiosyncratic errors conditional on all regressors is constant: $var(u_{it} \mid \boldsymbol{X}_i, a_i) = var(u_{it}) = \sigma_u^2$, $t = 1, 2, \ldots, T$. [homoskedasticity]

FE.6 No serial correlation exists among idiosyncratic errors. $cov(u_{it}, u_{is} \mid \boldsymbol{X}_i, a_i) = 0$, $t \neq s$

FE.7 [normality of u_{it} has little actual importance for the RE estimator

---

RE.1 There are no perfect linear relationships among explanatory variables. [replaces **FE.3**]

RE.2 In addition to **FE.4**, the expected value of a_i given all regressors is constant: $E(a_i \mid \boldsymbol{X}_i) = \beta_0$. [Rules out correlation between a_i and $\boldsymbol{X}_i$]

RE.3 In addition to **FE.5**, variance of a_i given all regressors is constant: $var(a_i \mid \boldsymbol{X}_i) = \sigma_a^2$ [Homoskedasticity imposed on a_i]

RE estimator – Assumptions

Under **FE.1+FE.2+RE.1+(FE.4+RE.2)**
RE estimator is consistent and asymptotically normal
(for fixed T as $N \to \infty$).
RE standard errors and statistics are not valid unless
(FE.5+RE.3) and **FE.6** conditions are met.

Under
FE.1-FE.2+RE.1+(FE.4+RE.2)+(FE.5+RE.3)+FE.6
RE estimator is consistent and asymptotically normal
(for fixed T as $N \to \infty$).
RE standard errors and statistics are valid.
RE is asymptotically efficient

- lower st.errs. than pooled OLS
- for time-varying variables, RE estimator is more efficient
  than FE (FE cannot be used on time-invariant variables).

RE estimator – Example

Example:
Wage equation using using panel data

> Random effects is used because many of the variables are time-invariant.
> <u>But is the random effects assumption realistic?</u>

$$\widehat{\log}(wage_{it}) = .92 \; educ_{it} - black_{it} + hisp_{it}$$
$$\phantom{\widehat{\log}(wage_{it}) = } (.011) \qquad (.048) \qquad (.043)$$
$$+ .106 \; exper_{it} - .0047 \; exper_{it}^2 + .064 \; married_{it}$$
$$(.015) \qquad (.0007) \qquad (.017)$$
$$+ .106 \; union_{it} + time \; dummies$$
$$(.018)$$

Random effects or fixed effects?
In economics, unobserved individual effects are rarely
uncorrelated with explanatory variables.
Hence, FE model/estimation is more convincing.

RE vs FE estimator

Hausman test / Hausman statistics may be used to choose between RE and FE:

$$H = (\hat{\boldsymbol{\beta}}_{FE} - \hat{\boldsymbol{\beta}}_{RE})^T [\widehat{Avar}(\hat{\boldsymbol{\beta}}_{FE}) - \widehat{Avar}(\hat{\boldsymbol{\beta}}_{RE})]^{-1} (\hat{\boldsymbol{\beta}}_{FE} - \hat{\boldsymbol{\beta}}_{RE}) \underset{H_0}{\sim} \chi^2(M)$$

where M is the number of regressors varying across i and t.

H_0: $cov(\boldsymbol{x}_{it}, a_i) = 0$ ... i.e. the crucial RE assumption holds

H_1: RE assumptions violated.

RE vs FE estimator

$$H = (\hat{\boldsymbol{\beta}}_{FE} - \hat{\boldsymbol{\beta}}_{RE})^T [\widehat{Avar}(\hat{\boldsymbol{\beta}}_{FE}) - \widehat{Avar}(\hat{\boldsymbol{\beta}}_{RE})]^{-1} (\hat{\boldsymbol{\beta}}_{FE} - \hat{\boldsymbol{\beta}}_{RE}) \underset{H_0}{\sim} \chi^2(M)$$

If $\hat{\boldsymbol{\beta}}_{FE}$ and $\hat{\boldsymbol{\beta}}_{RE}$ do not differ too much [or when the asymptotic variances are relatively large] we do not reject H_0 ... if we may assume RE assumptions hold, both RE and FE are consistent, and RE is efficient. For asymptotic variance estimators ($\widehat{Avar}$), see Wooldridge (2010).

If we reject H_0, we need to assume that RE assumptions are violated → RE is not consistent [we use FE]
CRE may be used to test FE vs. RE (explained next).

CRE estimator

Correlated Random Effects (CRE) estimator - a synthesis of the RE and FE approaches:

- a_i viewed as random, yet they can be correlated with $\boldsymbol{x}_{it}$.

  Specifically, as a_i do not vary over time, it makes sense to allow for their correlation with the time average of $x_{it} : \overline{x}_i = T^{-1} \sum_{t=1}^{T} x_{it}$

- CRE allows for incorporation of time-invariant regressors (compare to FE).

- CRE allows for convenient testing of FE vs. RE.

CRE estimator

CRE: The individual-specific effect a_i is split up into a part that is related to the time-averages of the explanatory variables and a part r_i (a time-constant unobservable) that is unrelated to the explanatory variables:

For $y_{it} = \beta_1 x_{it} + a_i + u_{it}$, we assume (a single-regressor illustration):

$$a_i = \alpha + \gamma \overline{x}_i + r_i, \text{ now: } corr(r_i, \overline{x}_i) = 0 \Rightarrow corr(r_i, x_{it}) = 0$$

Because $\overline{x}_i$ is a linear function of x_{it}

By substituting for a_i into the first equation, we obtain:

$$y_{it} = \alpha + \beta_1 x_{it} + \gamma \overline{x}_i + r_i + u_{it}$$

This equation can be estimated using RE
As $\gamma \overline{x}_i$ controls for the correlation between a_i and x_{it}, r_i is uncorrelated with regressors.

CRE estimator

CRE: $y_{it} = \alpha + \beta_1 x_{it} + \gamma \overline{x}_i + r_i + u_{it}$

CRE is a modified RE of the original equation $y_{it} = \beta_1 x_{it} + a_i + u_{it}$:

with uncorrelated random effect r_i but with the time averages as additional regressors.

The resulting CRE estimate for β is identical to the FE estimator.
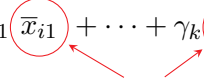
- CRE allows for incorporation of time-invariant regressors:
  Besides $\hat{\beta}_{CRE} = \hat{\beta}_{FE}$, we can include arbitrary time invariant regressors and estimate γ_{CRE} values.

- CRE allows for convenient testing of FE vs. RE:

  H_0: $\gamma = 0$ can be evaluated using $\hat{\gamma}_{CRE}$ and appropriate (HCE) standard errors against

  H_1: $\gamma \neq 0$

[RE assumes $\gamma = 0$: if we reject H_0, we also reject RE in favor of FE]

CRE estimator

The application of CRE to a model with multiple regressors is simple:

$$y_{it} = \beta_1 x_{it1} + \cdots + \beta_k x_{itk} + a_i + u_{it}, \ i = 1, \ldots, N, \ t = 1, \ldots, T$$

$$a_i = \alpha + \gamma_1 \overline{x}_{i1} + \cdots + \gamma_k \overline{x}_{ik} + r_i$$

Within means for all time-variant regressors

$$\Rightarrow y_{it} = \beta_0 + \alpha + \beta_1 x_{it1} + \cdots + \beta_k x_{itk} + \gamma_1 \overline{x}_{i1} + \cdots + \gamma_k \overline{x}_{ik} + [r_i + u_{it}]$$

Applying panel data methods to other data structures

- Panels are designed for two dimensional data.
  (Data are grouped by both cross section and time period.)

- Grouping data is sometimes useful even when there is only
  one dimension to group along
  (clusters, matched pairs, countries, etc.).

  In other words, sometimes it's useful to pretend that data
  come in a panel even when they don't.

- This can be a useful tool for estimating separate intercepts
  for each group/cluster.

Applying panel data methods to other data structures

EViews Illustrated example, based on data from the USA:
(Current Population Survey for March 2004; 100,000 individuals)

$$\log(wage_i) = \beta_0 + \beta_1 educ_i + \beta_2 age_i + \beta_3 asian_i + u_i$$

In pooled OLS, $\hat{\boldsymbol{\beta}}$ may be biased due to different wage levels (average wages) in different federal states. This may be controlled directly – by including dummies for the 51 states …

Often, it is more convenient to pretend that each state identifies a cross section in a panel. Then, we may use the FE/RE/CRE estimation (a_s are the state-specific wage effects):

$$\log(wage_{si}) = \beta_0 + \beta_1 educ_{si} + \beta_2 age_{si} + \beta_3 asian_{si} + a_s + u_{si}$$

Autocorrelation and heteroskedasticity in panel data models

Panel data extensions: $\qquad y_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta} + a_i + u_{it}$

<u>Heteroskedasticity</u>

RE model: $\ var(v_{it} \mid \boldsymbol{X}_i) = \sigma_{a_i}^2 + var(u_{it} \mid \boldsymbol{X}_i) = \begin{cases} \sigma_{a_i}^2 + \sigma_{u_i}^2 \\ \sigma_{a_i}^2 + \sigma_{u_t}^2 \end{cases}$

<u>Correlation between cross-sectional units (contemporaneous correlation)</u>

The general H_0 of no C-S dependence may be written as follows:

$$\rho_{ij} = corr(u_{it}, u_{jt}) = 0 \ \ for \ \ i \neq j$$

Autocorrelation & heteroskedasticity in panel data models

Panel data extensions: $y_{it} = \boldsymbol{x}_{it}\boldsymbol{\beta} + a_i + u_{it}$

Serial correlation (between-period correlation)

$$u_{it} = \begin{cases} \rho\, u_{i,t-1} + \varepsilon_{it} \\ \rho_i\, u_{i,t-1} + \varepsilon_{it} \end{cases}$$

Non-stationarity and panel cointegration

For the above "extensions", tests and GLS methods are usually estimator-specific (FD/FE/RE/CRE). Different types of assumption violations may occur simultaneously. Topic not covered in this course, see e.g. Wooldridge, 2010 for details.

Arellano-Bond estimator (dynamic panels)

Dynamic panel

$y_{it} = \delta_1 y_{i,t-1} + \boldsymbol{x}_{it}'\boldsymbol{\beta} + a_i + u_{it}$

. . . May be expanded using additional lags of the dependent variable or using lagged exogenous regressors.

Nickel Bias

- Related mostly to the lagged exogenous regressors $\boldsymbol{x}$

- FEs take up some part of the dynamic effect and therefore dynamic panel data models lead to overestimated FEs and underestimated dynamic interactions.

- Whether the Nickel bias is significant in a particular model/dataset situation is an empirical question. Nevertheless, in theory this bias persists unless the number of time observations goes to infinity.

- The inclusion of additional cross-sections to the dataset would worsen the bias in most cases.

Arellano-Bond estimator (dynamic panels)

Arellano-Bond (AB) **estimator**

- The model is transformed into first differences to eliminate the individual effects:
  $\Delta y_{it} = \delta_1 \Delta y_{i,t-1} + \Delta \boldsymbol{x}_{it}^T \boldsymbol{\beta} + \Delta u_{it}$,

- then a generalized method of moments (GMM) approach is used to produce asymptotically efficient estimates for the dynamic coefficients.

- AB approach is based on IV (we need instruments for the lagged dependent variable – this is an endogenous regressor, correlated with the errors in the FD model).

- Warning: AR(2) / not AR(1) / autocorrelation in residuals of the AB-estimated model renders the AB estimator inconsistent. After using the AB estimator, always test for AR(2) autocorrelation in the residuals!

Panel data - Extensions

- Advanced course on panel data

  http://people.stern.nyu.edu/wgreene/Econometrics/PanelDataNotes.htm

- Mixed effects model
  Extension to the RE model
  (intercept and -some- coefficients have a random term):

  $$y_{it} = \boldsymbol{x}'_{it}\boldsymbol{\beta} + \boldsymbol{z}'_{it}(\boldsymbol{\gamma} + \boldsymbol{h}_i) + (\alpha + u_i) + \varepsilon_{it}$$

  where $\boldsymbol{h}_i$ describes random variation of the paremeter(s)
  across individuals.

  http://www.bodowinter.com/tutorial/bw_LME_tutorial1.pdf