Panel Data Models

Ani Katchova

Outline

- Panel data set up and variations
- Pooled OLS estimator
- First differences estimator
- Fixed effects (within) estimator
- Dummy variables regression with fixed effects
- Random effects estimator
- Hausman test for fixed effects versus random effects

Panel data and variations

- Panel data variables:
  - Varying regressors x_{it} (annual income for a person, monthly food expenditures)
  - Time-invariant regressors $x_{it} = x_i$ for all t (gender, race, education)
  - Individual-invariant regressors $x_{it} = x_t$ for all i (time trend, US unemployment rate)
- Variations:
  - x_{it} is the individual value, $\bar{x}_i$ is the individual mean, and $\bar{x}$ is the overall mean
  - Overall variation (over time and individuals) $x_{it} - \bar{x}$
  - Between variation (variation between individuals) $\bar{x}_i - \bar{x}$
  - Within variation (variation within individuals over time) $x_{it} - \bar{x}_i$

Panel data variations

Id	Time	Variable	Individual mean	Overall mean	Overall deviation	Between deviation	Within deviation	Within deviation (modified)	First differences
i	t	x_{it}	$\bar{x}_i$	$\bar{x}$	$x_{it} - \bar{x}$	$\bar{x}_i - \bar{x}$	$x_{it} - \bar{x}$	$x_{it} - \bar{x}_i + \bar{x}$	$x_{it} - x_{i(t-1)}$
1	2019	9	10	20	-11	-10	-1	19	.
1	2020	10	10	20	-10	-10	0	20	1
1	2021	11	10	20	-9	-10	1	21	1
2	2019	20	20	20	0	0	0	20	.
2	2020	20	20	20	0	0	0	20	0
2	2021	20	20	20	0	0	0	20	0
3	2019	25	30	20	5	10	-5	15	.
3	2020	30	30	20	10	10	0	20	5
3	2021	35	30	20	15	10	5	25	5

Pooled OLS estimator

- The pooled OLS estimator uses both the between and within variation to estimate the parameters.
- Stack the data over i and t into one long regression
- Pooled OLS estimator:
- $y_{it} = \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + u_{it}, i = 1 \dots N, t = 1 \dots T$
- NT observations

Between estimator

- The between estimator uses the between variation (between individuals)
- Uses time averages of all variables $\bar{y}_i$, $\bar{x}_{1i}$, $\bar{x}_{2i}$
- Between estimator:

$$\bar{y}_i = \beta_0 + \beta_1 \bar{x}_{1i} + \beta_2 \bar{x}_{2i} + \bar{u}_i, i = 1 \dots N$$

- N observations
- Rarely used because panel data gets collapsed over time.

First differences estimator

- The first differences estimator uses the one period changes for each individual.
- Panel data model: $y_{it} = \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + a_i + u_{it}, i = 1 \dots N, t = 1 \dots T$
- Individual specific effects a_i are unobserved factors attributed to each individual.
- First differences estimator:

$$y_{it} - y_{i(t-1)} = \beta_0 + \beta_1 (x_{1it} - x_{1i(t-1)}) + \beta_2 (x_{2it} - x_{2i(t-1)}) + (u_{it} - u_{i(t-1)})$$

- OLS estimation of the one period changes of the dependent variable on the one period changes in the independent variables.
- Number of observations: $N(T-1)$ (first period is missing because of differencing)
- The individual specific effects a_i cancel out.
- Time-invariant variables are dropped from the model and their coefficients are not estimated.

Fixed effects within estimator

- The fixed effects estimator uses within variation (within same individual, over time) by using time demeaned variables.
- Panel data model: $y_{it} = \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + a_i + u_{it}, i = 1 \dots N, t = 1 \dots T$
- The individual specific effect a_i is potentially correlated with the independent variables.
- Time-averages (taking averages over time): $\bar{y}_i = \beta_0 + \beta_1 \bar{x}_{1i} + \beta_2 \bar{x}_{2i} + \bar{a}_i + \bar{u}_i$
- Subtract the second equation from the first equation, individual specific effect cancels out because it does not vary over time.
- Fixed effects within estimator (uses time demeaned variables):

$$y_{it} - \bar{y}_i = \beta_1 (x_{1it} - \bar{x}_{1i}) + \beta_2 (x_{2it} - \bar{x}_{2i}) + (u_{it} - \bar{u}_i)$$

Fixed effects within estimator

- The time-demeaned model does not include the individual specific effect (a_i) and can be estimated by OLS.
- Number of observations is NT.
- The within estimator cannot include time-invariant variables because they will drop out.
- After the fixed effects estimation, the individual specific effects a_i can be estimated as:

$$\hat{a}_i = \bar{y}_i - \hat{\beta}_0 - \hat{\beta}_1 \bar{x}_{1i} - \hat{\beta}_2 \bar{x}_{2i}$$

- The individual specific effects a_i sum up to zero across all individuals i, $\sum_{i=1}^{N} a_i = 0$

Dummy variables regression with fixed effects

- Regression with fixed effects as dummy variables:

$$y_{it} = \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + a_1 d_1 + a_2 d_2 + \cdots + a_N d_N + u_{it}$$

$$i = 1 \ldots N, t = 1 \ldots T$$

- Dummy variable $d_1 = 1$ for the first individual and zero for everyone else, $d_2 = 1$ for the second individual and zero for everyone else, …, $d_N = 1$ for the Nth individual and zero for everyone else.
- In the dummy variables regression, the coefficient $a_1 = 0$ is normalized to zero, so the dummy variable d_1 is the reference category not included in the model.
- The individual specific effects a_i are the coefficients to be estimated in the dummy variables regression.
- The fixed effects (within) estimator is equivalent to a regression with dummy variables for each individual.

Panel data example

fcode	year	lscrap	tothrs	d88	d89	grant	grant_1
410523	1987	-2.81	40	0	0	0	0
410523	1988	-3.00	40	1	0	0	0
410523	1989	-3.00	60	0	1	0	0
410538	1987	0.97	0	0	0	0	0
410538	1988	1.01	0	1	0	0	0
410538	1989	0.93	30	0	1	0	0
410563	1987	1.79	0	0	0	0	0
410563	1988	1.95	0	1	0	0	0
410563	1989	1.61	50	0	1	0	0

Panel data (3 years, 54 firms). Firm's log scrap rates (lscrap) are explained by whether the firm received a grant (grant) and total hours of training (tothrs). d88 and d89 are time dummies for each year.

Do firms with more hours of training or grants have lower scrap rates?

Means and variations for panel variables

Variable		Mean	Std. Dev.	Min	Max
fcode	overall	416314	3759	410523	419483
	between		3782	410523	419483
	within		0	416314	416314
year	overall	1988	0.82	1987	1989
	between		0	1988	1988
	within		0.82	1987	1989
lscrap	overall	0.39	1.49	-4.61	3.40
	between		1.43	-3.01	3.21
	within		0.45	-2.03	2.12
tothrs	overall	23.71	28.01	0	154
	between		21.06	0	93
	within		18.58	-19.62	100.38

Panel data set up:
fcode (the firm code) has only between (3782) and not within (0) variation.
year has only within (0.82) and no between (0) variation.

Standard deviations:
There is more variation for the log scrap rate between firms (1.43) than within firms (0.45).
Total hours of training has about the same variation between firms (21.06) and within firms (18.58).

Pooled OLS, between estimator, and first differences estimator

- Pooled OLS

$$lscrap_{it} = \beta_0 + \beta_1 tothrs_{it} + \beta_2 d88_t + \beta_3 d89_t + \beta_4 grant_{it} + \beta_5 grant\_1_{it} + u_{it}$$

- Between estimator

$$\overline{lscrap_i} = \beta_0 + \beta_1 \overline{tothrs_i} + \beta_2 \overline{d88} + \beta_3 \overline{d89} + \beta_4 \overline{grant_i} + \beta_5 \overline{grant\_1_i} + \bar{u}_i$$

- First differences estimator

$$(lscrap_{it} - lscrap_{i(t-1)}) = \beta_1 (tothrs_{it} - tothrs_{i(t-1)}) + \beta_2 (grant_{it} - grant_{i(t-1)}) + (u_{it} - u_{i(t-1)})$$

- In the first differences estimator, time dummies are differenced out and cannot be included.

Pooled OLS and between estimator

	Pooled OLS	Between estimator
VARIABLES	lscrap	lscrap
tothrs	-0.005	-0.01
	(0.005)	(0.01)
d88	-0.27	-3.14
	(0.33)	(5.11)
d89	-0.51	
	(0.37)	
grant	0.38	2.49
	(0.38)	(1.81)
grant_1	0.08	-1.19
	(0.45)	(1.77)
Constant	0.66***	1.27
	(0.22)	(1.71)

Pooled OLS model stacks data for individuals and years and estimates the model by OLS. Results show that the effect of one additional hour of training across firms and over time on the log of scrap rate. This effect is not significant.

The between estimator uses the averages of variables over time for each firm. Results show the effect of one additional hour of training for a firm in comparison to another firm on the log of scrap rate for this firm in comparison to another firm. This effect is not significant.

First differences estimator

	First differences
VARIABLES	dlscrap
dtothrs	-0.003
	(0.003)
dgrant	0.05
	(0.12)
Constant	-0.21***
	(0.06)

First differences estimator uses first differences of all variables. The results show the effect of increase in total hours of training from one year to the next for the same firm on the change in log scrap rate from one year to the next for the same firm. This effect is not significant.

Fixed effects within estimator and dummy variables regression

- Fixed effects within estimator:

$$(lscrap_{it} - \overline{lscrap_i}) = \beta_1(tothrs_{it} - \overline{tothrs_i}) + \beta_2(d88_t - \overline{d88}) + \beta_3(d89_t - \overline{d89}) + \beta_4(grant_{it} - \overline{grant_i}) + \beta_5(grant\_1_{it} - \overline{grant\_1_i}) + (u_{it} - \bar{u}_i)$$

- The individual specific effects can be recovered as:

$$\hat{a}_i = \overline{lscrap_i} - \hat{\beta}_0 - \hat{\beta}_1\overline{tothrs_i} - \hat{\beta}_2\overline{d88} - \hat{\beta}_3\overline{d89} - \hat{\beta}_4\overline{grant_i} - \hat{\beta}_5\overline{grant\_1_i}$$

- The individual specific effects a_i sum up to zero.

- Dummy variable regression:
- $lscrap_{it} = \beta_0 + \beta_1 tothrs_{it} + \beta_2 d88_t + \beta_3 d89_t + \beta_4 grant_{it} + \beta_5 grant\_1_{it} + a_1 d_1 + a_2 d_2 + \cdots + a_N d_N + u_{it}$
- There are N=54 dummy variables, one for each firm. $a_1 = 0$ is normalized to zero for the first firm. The coefficients a_i are estimated and included in the regression output.

Fixed effects within estimator

	Fixed effects within estimator
VARIABLES	lscrap
tothrs	-0.005
	(0.003)
d88	-0.07
	(0.12)
d89	-0.22
	(0.16)
grant	-0.12
	(0.18)
grant_1	-0.41*
	(0.23)
Constant	0.66***
	(0.09)
R-squared	0.23

The within transformation has each variable minus its average over time.
Results show an increase of total hours of training for the same firm from its mean on the log of scrap rate from its mean for the same firm. This effect is not significant.

Receiving a grant in the previous period is associated with 41% lower scrap rates.

	Fixed effects within estimator	Dummy variables regression
VARIABLES	lscrap	lscrap
tothrs	-0.005	-0.005
	(0.003)	(0.003)
d88	-0.07	-0.07
	(0.12)	(0.12)
d89	-0.22	-0.22
	(0.16)	(0.16)
grant	-0.12	-0.12
	(0.18)	(0.18)
grant_1	-0.41*	-0.41*
	(0.23)	(0.23)
410538.fcode		3.73***
		(0.43)
410563.fcode		4.58***
		(0.43)
….		…
419483.fcode		5.95***
		(0.44)
Constant	0.66***	-2.62***
	(0.09)	(0.33)
R-squared	0.23	0.92

Comparison of fixed effects using within estimator and dummy variable regression.
The coefficients on the variables are the same in both models.
The dummy variable regression includes the coefficients on the dummy variables (410538.fcode, 410563.fcode, etc.) but because there are (N-1)=53 coefficients, they are typically not included when presenting the results.

Individual specific effects

	Fixed effects (within estimator)	Fixed effects with dummy variables regression	Individual specific effect plus intercept
fcode 410523	$\hat{a}_1 = -3.28$	$\hat{a}_1 = 0$	$-3.28 + 0.66 = 0 + (-2.62)$
fcode 410538	$\hat{a}_2 = 0.45$	$\hat{a}_2 = 3.73$	$0.45 + 0.66 = 3.73 + (-2.62)$
fcode 410563	$\hat{a}_3 = 1.30$	$\hat{a}_3 = 4.58$	$1.30 + 0.66 = 4.58 + (-2.62)$
Intercept	0.66	-2.62	

For the within estimator, the individual specific effects sum up to zero. Log scrap rates for the second firm are 0.45 higher than the average log scrap rates across firms.
For the dummy variables regression, the coefficient on the fixed effects dummy variables is normalized to zero for the first firm. Log scrap rates for the second firm are 3.73 higher compared to the first firm.
The individual specific effect plus the regression intercept is the individual specific intercept for this firm.

R-squared for within estimator vs dummy variables regression

- The R-squared is higher for the dummy variables regression than FE within estimator that only explains within variation.
- For FE within estimator: Model SS = 7, Residual SS = 24,

  R-squared = Model SS / (Model SS + Residual SS) = 7/(7+24)= 0.22
- For dummy variables regression: Model/Explained SS = 287, Residual SS = 24, R-squared = Model SS / (Model SS + Residual SS) = 287/(287+24)= 0.92
- Same residual variation.
- Variables are demeaned for FE within estimator $(lscrap_{it} - \overline{lscrap_i})$, so there is less total variation to be explained in the FE estimator and it is mostly residual variation leading to lower R-squared.
- There is more total variation for the dummy variables regression because the variables are not demeaned. The dummy variables (d_1, d_2,… d_N) help explain model variation, leading to higher R-squared.

Discussion on fixed effects estimator

- The original model needs to have exogeneity, where variables are not correlated with the error term.
- The R-squared for the demeaned model (within estimator) is not appropriate because the variables are demeaned so they only show the within variation, not the overall variation.
- Time-invariant variables cannot be included because they drop out. Interaction terms with time-invariant variables can be included in the model.
- Variables that change deterministically over time (such as experience which increases by one year every year) cannot be included in the model.

Fixed effects versus first difference estimator

- The fixed effects and first difference estimators are identical with two time periods (T=2).
  - First difference, $\Delta y_{it} = y_{it} - y_{i(t-1)}$
  - Within estimator, time demeaned variables:
  - $y_{it} - \bar{y}_i = y_{it} - \frac{y_{it}+y_{i(t-1)}}{2} = \frac{y_{it}-y_{i(t-1)}}{2}$
- With more time periods (T>2)
  - The fixed effects is more efficient if the classical assumptions hold.
  - First differencing may be better if there are many time periods with strong serial correlation in the errors.

Random effects estimator

- Panel data model: $y_{it} = \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + a_i + u_{it}, i = 1 \dots N, t = 1 \dots T$
- Random effects assumption: the individual specific effect a_i is assumed to be "random" and not correlated with the independent variables, $cov(x_{jit}, a_i) = 0$.
- The error term $a_i + u_{it}$ is serially correlated within the individual:
- $cov(a_i + u_{it}, a_i + u_{i(t-1)}) = cov(a_i, a_i) = \sigma_a^2$
- The individual specific effects a_i are correlated within the individual over time.

Random effects estimator

- Under the random effects assumption, the independent variables are uncorrelated with the error term, so the pooled OLS will provide consistent but inefficient estimates.
- With the OLS, the standard errors will need to be adjusted because they are correlated over time.
- Transform the model so the new errors are not correlated:

$$y_{it} - \theta\bar{y}_i = \beta_0 + \beta_1(x_{1it} - \theta\bar{x}_{1i}) + \beta_2(x_{2it} - \theta\bar{x}_{2i}) + (a_i - \theta\bar{a}_i) + (u_{it} - \theta\bar{u}_i)$$

Random effects estimator

- Random effects estimator:
$$y_{it} - \theta\bar{y}_i = \beta_0 + \beta_1(x_{1it} - \theta\bar{x}_{1i}) + \beta_2(x_{2it} - \theta\bar{x}_{2i}) + (a_i - \theta\bar{a}_i) + (u_{it} - \theta\bar{u}_i)$$
- The random effects parameter θ can be estimated: $\theta = 1 - \sqrt{\frac{\sigma_u^2}{\sigma_u^2 + T\sigma_a^2}}$
- $\hat{\theta} = 0$ corresponds to pooled OLS (a_i is not important)
- $\hat{\theta} = 1$ corresponds to the FE within estimator (a_i is important)
- The random effects estimator is a weighted average of the pooled OLS and the FE within estimator.
- The random effects estimator can include time-invariant variables.

Random effects estimator

- Random effects estimator
- $(lscrap_{it} - \theta\overline{lscrap_i}) = \beta_1(tothrs_{it} - \theta\overline{tothrs_i}) + \beta_2(d88_t - \theta\overline{d88}) + \beta_3(d89_t - \theta\overline{d89}) + \beta_4(grant_{it} - \theta\overline{grant}_i) + \beta_5(grant\_1_{it} - \theta\overline{grant\_1}_i) + (u_{it} - \theta\bar{u}_i)$

- The RE estimation output reports $\sigma_a = 1.39$ and $\sigma_u = 0.51$. The individual specific effects a_i are more important than the idiosyncratic error term u_{it}.
- The median for the θ parameter is:
- $\theta = 1 - \sqrt{\frac{\sigma_u^2}{\sigma_u^2 + T\sigma_a^2}} = 1 - \sqrt{\frac{0.51^2}{0.51^2 + 3*1.39^2}} = 0.79$, which is closer to the FE model than the pooled OLS model. This means that the individual specific effects are more important.
- (Stata denotes a_i as u_i and denotes u_{it} as e_{it}.)

Random effects estimator

	Random effects
VARIABLES	lscrap
tothrs	-0.005*
	(0.005)
d88	-0.09
	(0.12)
d89	-0.25
	(0.15)
grant	-0.07
	(0.18)
grant_1	-0.35
	(0.22)
Constant	0.66***
	(0.22)

For the random effects estimator, for each additional hour in total training, the scrap rate is lower by 0.5%. The coefficients on time dummies and grants are not significant.

Hausman test

- The Hausman test is used to decide whether to use the fixed effects (FE) or random effects (RE) estimator.
- H_0: no corrrelation between individial specific effects and independent variables, FE and RE coefficients are not signficantly different from each other
- H_a: correlation between individual specific effects and independent variables, FE and RE coefficients are significantly different from each other
- Calculate the difference in RE and FE coefficients ($\beta_{RE} - \beta_{FE}$) and their covariances.
- The Hausman test statistic:

$$W = (\beta_{RE} - \beta_{FE})'\left(var(\beta_{RE}) - var(\beta_{FE})\right)^{-1}(\beta_{RE} - \beta_{FE}) \sim \chi^2$$

Hausman test

Estimator	H_0 is true	H_a is true
RE estimator	Consistent and efficient	Inconsistent
FE estimator	Consistent but inefficient	Consistent

- If the Hausman test statistic W is not significantly different from zero, then both the FE and RE estimators are consistent. RE estimator should be used because it is more efficient.
- If the Hausman test statistic W is significantly different from zero, then only the FE estimator is consistent and should be used.
- The Hausman test evaluates the consistency of an RE estimator against a less efficient FE estimator that is known to be consistent.
- The individual specific effects are typically correlated with the independent variables, making the FE estimator more appropriate.

Hausman test example

	Fixed effects	Random effects
VARIABLES	lscrap	lscrap
tothrs	-0.005	-0.005*
	(0.003)	(0.005)
d88	-0.07	-0.09
	(0.12)	(0.12)
d89	-0.22	-0.25
	(0.16)	(0.15)
grant	-0.12	-0.07
	(0.18)	(0.18)
grant_1	-0.41*	-0.35
	(0.23)	(0.22)
Constant	0.66***	0.66***
	(0.09)	(0.22)

The coefficients for the FE and RE estimators (β_{FE} and β_{RE}) are similar.
Hausman test statistic W= 1.51, and p-value=0.91.
No significant difference in FE and RE coefficients, so both are estimators are consistent. RE estimator should be used because it is more efficient.

Review questions

- Describe overall, between, and within variation.
- Describe individual specific effects. How are the fixed effects recovered after estimation?
- Describe the different estimation methods for panel data models
  - Pooled model/ Between estimator/ First difference estimator
  - Fixed effects (within) estimator/ Dummy variables regression
  - Random effects estimator
- Compare the fixed effects within estimator and dummy variables regression.
- Describe the Hausman test for fixed effects versus random effects. What assumption do the fixed effects and random effects estimators make about the individual specific effects?