Seminar 9 Solutions

Giulio Rossetti*

giuliorossetti94.github.io

March 13, 2025

* email: giulio.rossetti.1@wbs.ac.uk

Disclaimer

Full solutions are available on my.wbs. All exercises are examinable material, not just the ones we covered in the seminars.

Roadmap

Exercise 1

Exercise 4

CS

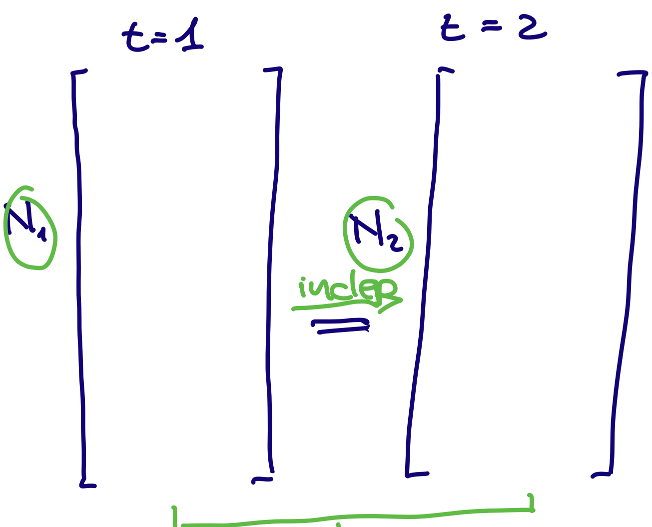
n individuals t = fixed

TS

T periods n = fixed = 1

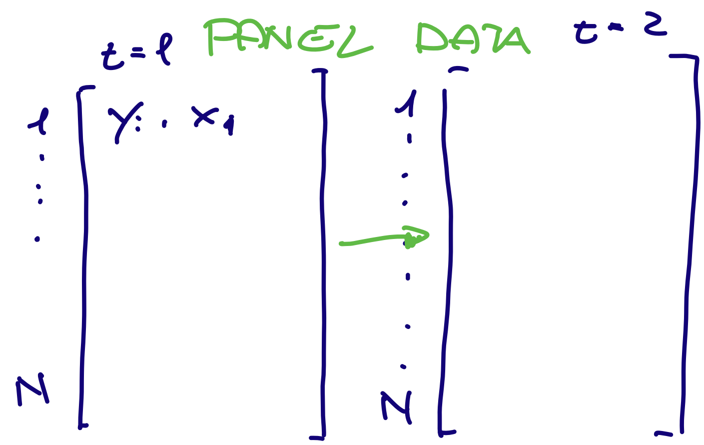
PANEL DATA

POOLED DATA



pooled OLS

PANEL DATA



FIXED EFFECT MODELS

RANDOM EFFECT MODELS

i = London, Coventry .... N cities

$t = 1 \dots T$

$$crime_{it} = \beta_0 + \beta_1 unemp_{it} + \alpha_i + \overbrace{u_{it}}^{\text{idiosyncr.}}$$

α_i → unobserved heterogeneity → city specific characteristic

pooled OLS => $cov(\alpha_i, unemp) \neq 0$

$\Downarrow$

$\hat{\beta}$ not consistent

FEM

① First difference estimator → remove α_i by first diff

② Fixed effect estimator → remove α_i by demeaning

① $crime_{it} = \beta_0 + \beta_1 unemp_{it} + \alpha_i + u_{it}$

$$crime_{i2} - crime_{i1} = \beta_1 (unemp_{i2} - unemp_{i1}) + \cancel{\alpha_i} - \cancel{\alpha_i} + (u_{i2} - u_{i1})$$

$$\Delta crime_i = \beta_1 \Delta unemp_i + \Delta u_i$$

↳ POOLED OLS

② $\overline{crim_i} = \frac{1}{T} \sum_{t=1}^{T} crime_{it}$ $\overline{unemp_i} = \frac{1}{T} \sum_{t=1}^{T} unemp_{it}$

$\overline{u_i} = \frac{1}{T} \sum_{t=1}^{T} u_{it}$

$$crime_{it} - \overline{crime_i} = \beta_1 (unemp_{it} - \overline{unemp_i}) + \cancel{\alpha_i} - \cancel{\alpha_i} + (u_{it} - \overline{u_i})$$

↳ POOLED OLS

$cov(\alpha_i, u_{it}) = 0$ α_i RV

RANDOM EFFECT MODEL

↳ POOLED OLS

↳ adjust SEs

Part 1: Theory

$$\hat{\beta}_{FD} = \hat{\beta}_{FE} \text{ iff } T=2$$

Exercise 1

For $T = 2$, consider the standard panel data model:

$$y_{it} = x_{it}'\beta + \alpha_i + u_{it}, \quad t = 1, 2, \quad i = 1, \dots, n$$

where i denotes the cross-sectional unit and t denotes the time dimension. For simplicity, assume that in this model there is no intercept.

First-Difference Estimator

Show that the fixed-effects(**FE**) and first-difference (**FD**) estimators are identical (they deliver the same beta estimates.)

- **FD**: Remove unobs heterogeneity by differencing over time:

$$y_{i2} - y_{i1} = (x_{i2} - x_{i1})'\beta + (u_{i2} - u_{i1})$$

$$\Delta y_i = \Delta x_i'\beta + \Delta u_i + \alpha_i - \alpha_i$$

- Assuming independence of the error terms, β_{FD}:

$$\hat{\beta}_{FD} = \left(\sum_{i=1}^{n} \Delta x_i \Delta x_i' \right)^{-1} \sum_{i=1}^{n} \Delta x_i \Delta y_i.$$

Fixed-Effects Estimator

- **FE** : Remove unobs heterogeneity by demeaning:

$$\bar{y}_i = \frac{1}{2}(y_{i1} + y_{i2}), \quad \bar{x}_i = \frac{1}{2}(x_{i1} + x_{i2}), \quad \bar{u}_i = \frac{1}{2}(u_{i1} + u_{i2}).$$

- Then, we have:

$$y_{it} - \bar{y}_i = (x_{it} - \bar{x}_i)'\beta + u_{it} - \bar{u}_i, \quad t = 1, 2.$$

- β_{FE}:

$$\hat{\beta}_{FE} = \left(\underbrace{\sum_{i=1}^{n} \sum_{t=1}^{2} (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)'}_{1/2\, \Delta x_i \Delta x_i'} \right)^{-1} \underbrace{\sum_{i=1}^{n} \sum_{t=1}^{2} (x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i)}_{1/2\, \Delta x_i \Delta y_i}.$$

Equivalence of FE and FD

Note that:

$$\sum_{t=1}^{2}(x_{it}-\bar{x}_i)(x_{it}-\bar{x}_i)' = \sum_{t=1}^{2}\left(x_{it}-\frac{x_{i1}+x_{i2}}{2}\right)\left(x_{it}-\frac{x_{i1}+x_{i2}}{2}\right)'$$

$$= \left(\frac{x_{i1}-x_{i2}}{2}\right)\left(\frac{x_{i1}-x_{i2}}{2}\right)' + \left(\frac{x_{i2}-x_{i1}}{2}\right)\left(\frac{x_{i2}-x_{i1}}{2}\right)'$$

$$= \boxed{\frac{1}{2}\Delta x_i \Delta x_i'.}$$

Equivalence of FE and FD

Similarly:

$$\sum_{t=1}^{2}(x_{it}-\bar{x}_i)(y_{it}-\bar{y}_i)=\boxed{\frac{1}{2}\Delta x_i \Delta y_i.}$$

Substituting into the FE estimator, we obtain:

$$\hat{\beta}_{FE}=\left(\cancel{\frac{1}{2}}\sum_{i=1}^{n}\Delta x_i \Delta x_i'\right)^{-1}\left(\cancel{\frac{1}{2}}\sum_{i=1}^{n}\Delta x_i \Delta y_i\right).$$

$$=\left(\sum_{i=1}^{n}\Delta x_i \Delta x_i'\right)^{-1}\sum_{i=1}^{n}\Delta x_i \Delta y_i=\hat{\beta}_{FD}.$$

Conclusion: The fixed-effects and first-difference estimators are identical when $T=2$.

Including *age* as a Regressor

t = 1 j = 10 yrs i = 12 yrs k = 24 yrs

t = 2 j = 11 i = 13 k = 25

Suppose that we include the variable *age* as an additional regressor and use first differencing to estimate a fixed effects model.

- Requirements behind the FD estimator: Δx_{it} must have some variation across i.
- This fails if an explanatory variable such as *age* is included.
  - *age* changes by the same amount for each of the individuals over time

$$y_{i1} = \beta_1 x_{i1} + \beta_2 x_{i2} + \alpha_i + u_{i1}, \quad t = 1, \quad i = 1, \ldots, n$$

$$y_{i2} = \beta_1 x_{i2} + \beta_2 x_{i2} + \alpha_i + u_{i2}, \quad t = 2, \quad i = 1, \ldots, n.$$

Differencing the Model

By subtracting the first equation from the second, we obtain:

$$\Delta y_i = \beta_1 \Delta x_{i1} + \beta_2 \Delta \underline{\underline{x_{i2}}} + \Delta u_i, \quad i = 1, \ldots, n.$$

Since x_{i2} increases by the same amount c across individuals:

$$\begin{aligned} \Delta y_i &= \beta_1 \Delta x_{i1} + \boxed{\beta_2 c} + \Delta u_i. \\ &= \beta_1 \Delta x_{i1} + \boxed{\delta} + \Delta u_i. \end{aligned}$$

where $\delta = \beta_2 c$ is a constant term.

Interpretation

Key issue: The constant term δ makes it problematic to identify β_2.

- δ does not represent the intercept (since there was no intercept in the original model).
- It also does not represent any change in the intercept by definition:
  - Since we allow α_i to be correlated with x_{i2}, we cannot separate the effect of α_i on y_i from the effect of any other variable that does not change over time.

Implications of Cov$(x_{it}, \alpha_i) = 0$

RE model

Suppose that Cov$(x_{it}, \alpha_i) = 0$. What does this imply for the FE and FD estimators?

- When we assume that Cov$(x_{it}, \alpha_i) = 0$, the original model becomes a *random effects model*.
- The random effects assumptions include all of the fixed effects assumptions plus the additional requirement that α_i is independent of all explanatory variables in all time periods.
- WNote that given Cov$(x_{it}, \alpha_i) = 0$, β can be consistently estimated by Pooled OLS.

Composite Error Term

However, this ignores a key feature of the model. If we define the composite error term as:

$$v_{it} = \alpha_i + u_{it},$$

corr of composite e.t. over time

we can show that:

$\neq 0$

$$\text{corr}(v_{it}, v_{is}) = \boxed{\frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_u^2}}, \quad t \neq s,$$

where:

$$\sigma_\alpha^2 = \text{Var}(\alpha_i), \quad \sigma_u^2 = \text{Var}(u_{it}).$$

Implications for Estimation

REM FEM

Hausman test

- positive serial correlation in the error term makes pooled OLS standard errors incorrect.
- We must:
  - Either correct the OLS SE, or
  - Use the GLS random effects estimator

Roadmap

Exercise 1

Exercise 4

Exercise 4: Rental Prices and Student Presence

The data for the years 1980 and 1990 include rental prices and other variables for college towns. The goal is to determine whether a stronger presence of students affects rental rates. The model is:

$$\log(\text{rent}_{it}) = \beta_0 + \delta_0 y90_t + \beta_1 \log(\text{pop}_{it}) + \beta_2 \log(\text{avginc}_{it}) + \beta_3 \text{pctstu}_{it} + e_{it},$$

where:

- pop is city population,
- avginc is average income,
- pctstu is student population as a percentage of city population (during the school year).

Pooled OLS Estimation Results

You estimate the model with pooled OLS and obtain the following results:

Source	SS	df	MS
Model	12.1080112	4	3.02700281
Residual	1.9501234	123	.015854662
Total	14.0581346	127	.110693974

Number of obs	=	128
F(4, 123)	=	190.92
Prob > F	=	0.0000
R-squared	=	0.8613
Adj R-squared	=	0.8568
Root MSE	=	.12592

lrent	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
y90	.2622267	.0347632	7.54	0.000	.1934151	.3310384
lpop	.0406863	.0225154	1.81	0.073	-.0038815	.0852541
lavginc	.5714461	.0530981	10.76	0.000	.4663417	.6765504
pctstu	.0050436	.0010192	4.95	0.000	.0030262	.007061
_cons	-.5688069	.5348808	-1.06	0.290	-1.627571	.4899568

Figure: Pooled OLS Estimation Results for Rental Prices and Student Presence

Interpreting the Regression Results

- Almost all regressors are statistically significant.
- City population is borderline significant.
- However, population per se is not a strong driving factor:
  - The number of inhabitants affects rents only if land size is limited.
  - This constraint is not explicitly considered in the model.
- There is a clear omitted variable bias:
  - City size is not constant and may depend on the city itself.
  - Example: London and Coventry do not have the same size.
- This leads to the so-called heterogeneous bias.
- To address this issue:
  - A fixed effects model can be used if regressors are correlated with city-specific effects.
  - A random effects model can be used if regressors are uncorrelated with city-specific effects.

Pooled OLS Estimation Results

Now you estimate the model with fixed effect and obtain the following results:

```
                                                F(4,60)           =    624.15
corr(u_i, Xb)  = -0.1297                        Prob > F          =    0.0000
```

lrent	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
y90	.3855214	.0368245	10.47	0.000	.3118615	.4591813
lpop	.0722456	.0883426	0.82	0.417	-.104466	.2489571
lavginc	.3099605	.0664771	4.66	0.000	.1769865	.4429346
pctstu	.0112033	.0041319	2.71	0.009	.0029382	.0194684
_cons	1.409384	1.167238	1.21	0.232	-.9254394	3.744208
sigma_u	.15905877					
sigma_e	.06372873					
rho	.8616755	(fraction of variance due to u_i)				

```
F test that all u_i=0:     F(63, 60) =    10.20            Prob > F = 0.0000
```

Figure: FE Estimation Results for Rental Prices and Student Presence

Fixed Effects and Model Selection

- By fully acknowledging unobservable fixed effects, the impact of *lpop* disappears.
- From the output, we see that:

cov(α_i, x_{it}) = 0
$\hookrightarrow$ REM

$$\text{corr}(\alpha_i, x_{it}) = -0.129,$$

which is relatively small.

- Given this small correlation, it might be sensible to use a *random effects model* instead.
- However, determining the appropriate model is difficult without first implementing a **Hausman test**.