Seminar 9 Solutions

Panel Data: Fixed Effects and First Differencing

Giulio Rossetti*

giuliorossetti94.github.io

March 20, 2026

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

Roadmap

Disclaimer

Disclaimer

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

CS

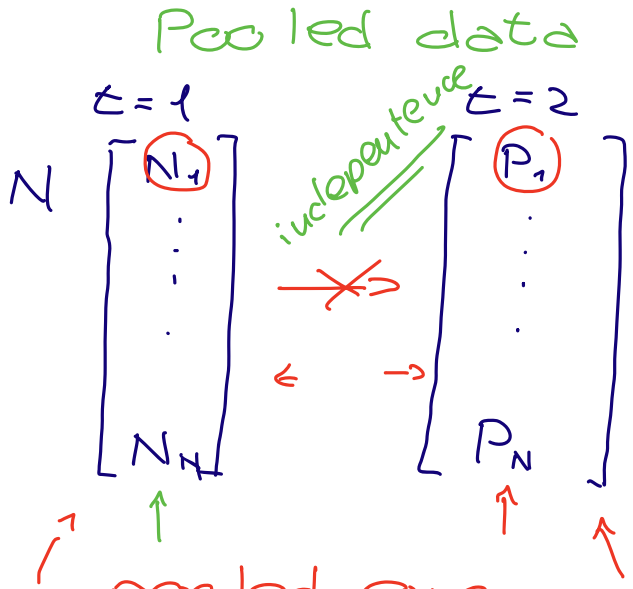
t [1 ... N] individuals

TS

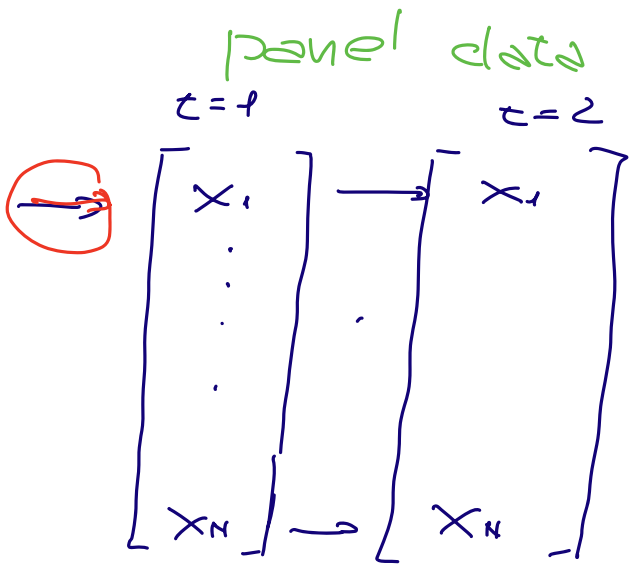
1 ind $\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_T \end{bmatrix}$

PANEL DATA

Pooled data



panel data



FIXED EFFECT MODELS

$$y_{it} = \beta_0 + x_{it}'\beta + u_{it}$$

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

α_i: unobserved heterogeneity

FIXED EFFECT MODELS $cov(\alpha_i, x_{it}) \neq 0$

① first difference: remove α_i by differencing

$$y_{i2} - y_{i1} = \beta_1 (x_{i2} - x_{i1}) + \cancel{\alpha_i} - \cancel{\alpha_i} + (u_{i2} - u_{i1})$$

$$\rightarrow \Delta y = \beta_1 \Delta x_i + \Delta u_i$$

↳ estimate β_1 by pooled OLS

② fixed effect: remove α_i by demeaning

$$\bar{y}_i = \frac{1}{T} \sum_{t=1}^{T} y_{it} \qquad \bar{x}_i = \frac{1}{T} \sum_{t=1}^{T} x_{it} \qquad \bar{u}_i = \frac{1}{T} \sum_{t=1}^{T} u_{it}$$

$$y_{it} - \bar{y}_i = (x_{it} - \bar{x}_i)\beta_1 + \cancel{\alpha_i} - \cancel{\alpha_i} + (u_{it} - \bar{u}_i)$$

$$\ddot{y}_{it} = \ddot{x}_{it}\beta + \ddot{u}_{it}$$

↳ estimate by pooled OLS

RANDOM EFFECT MODELS $cov(\alpha_i, x_{it}) = 0$

Roadmap

Exercise 1: FE and FD Equivalence

- Setup
- First-Difference Estimator
- Fixed-Effects Estimator
- Equivalence Proof
- Age as a Regressor
- Random Effects

Exercise 3: Difference-in-Differences (Incinerator)

Exercise 4: Rental Prices and Student Presence

- Setup
- Pooled OLS Estimation
- Interpretation
- Fixed Effects Estimation

Exercise 1

Panel Data Model

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

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

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

$t = 1 \quad y_{i1} = x_{i1}'\beta + \alpha_i + u_{i1}$

$t = 2 \quad y_{i2} = x_{i2}'\beta + \alpha_i + u_{i2}$

Exercise 1

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:

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

$$\Delta y_i = \Delta x_i'\beta + \Delta u_i.$$

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

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

Exercise 1

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}:

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

Exercise 1

Equivalence of FE and FD – Part 1

Note that:

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

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

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

Exercise 1

Equivalence of FE and FD – Part 2

Similarly:

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

Substituting into the FE estimator, we obtain:

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

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

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

Exercise 1

Including age as a Regressor

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

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

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

Exercise 1

Differencing the Model

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

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

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

$$\Delta y_i = \beta_1 \Delta x_{i1} + \beta_2 c + \Delta u_i.$$

$$= \beta_1 \Delta x_{i1} + \delta + \Delta u_i.$$

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

Exercise 1

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.

Exercise 1

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

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.
- Note that given Cov$(x_{it}, \alpha_i) = 0$, β can be consistently estimated by Pooled OLS.

Exercise 1

Composite Error Term

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

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

we can show that:

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

where:

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

Exercise 1

Implications for Estimation

- 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: FE and FD Equivalence

- Setup
- First-Difference Estimator
- Fixed-Effects Estimator
- Equivalence Proof
- Age as a Regressor
- Random Effects

Exercise 3: Difference-in-Differences (Incinerator)

Exercise 4: Rental Prices and Student Presence

- Setup
- Pooled OLS Estimation
- Interpretation
- Fixed Effects Estimation

Why Does This Matter?

Motivation

Difference-in-Differences (DiD) is a workhorse method for causal inference. This exercise uses the construction of an incinerator to study its effect on nearby house prices.

Exercise 3: The Model

$$\log(price) = \beta_0 + \delta_0 y81 + \beta_1 \log(dist) + \delta_1 y81 \cdot \log(dist) + u$$

- $dist$: distance from each home to the incinerator site
- $y81$: dummy $= 1$ for 1981 (when incinerator is built)
- δ_1: the DiD coefficient — how the distance effect changes after construction

Expected signs:

- $\beta_1 > 0$: further from incinerator $\Rightarrow$ higher price (incinerators are in depressed areas)
- $\delta_1 > 0$: if the incinerator reduces nearby home values, being further away becomes even more valuable after construction

Results: Basic Model

Source	SS	df	MS
Model	24.3172548	3	8.10575159
Residual	37.1217306	317	.117103251
Total	61.4389853	320	.191996829

Number of obs	= 321
F(3, 317)	= 69.22
Prob > F	= 0.0000
R-squared	= 0.3958
Adj R-squared	= 0.3901
Root MSE	= .3422

lprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
y81	-.0113101	.8050622	-0.01	0.989	-1.59525	1.57263
ldist	.316689	.0515323	6.15	0.000	.2153005	.4180775
y81ldist	.0481862	.0817929	0.59	0.556	-.1127394	.2091117
_cons	8.058468	.5084358	15.85	0.000	7.058133	9.058803

Conclusion: Neither $y81$ nor the interaction $y81 \cdot \log(dist)$ is statistically significant. The incinerator does not appear to affect house prices in this simple specification.

Results: With Controls

Adding age, $rooms$, $baths$, $\log(land)$, $\log(area)$:

Source	SS	df	MS
Model	47.6052846	8	5.95066057
Residual	13.8337008	312	.044338785
Total	61.4389853	320	.191996829

Number of obs	= 321
F(8, 312)	= 134.21
Prob > F	= 0.0000
R-squared	= 0.7748
Adj R-squared	= 0.7691
Root MSE	= .21057

lprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
y81	.0416659	.5028206	0.08	0.934	-.9476821	1.031014
ldist	.0001984	.0400148	0.00	0.996	-.0785346	.0789314
y81ldist	.0354178	.0511078	0.69	0.489	-.0651417	.1359773
age	-.0023866	.000428	-5.58	0.000	-.0032288	-.0015445
rooms	.0572316	.0175194	3.27	0.001	.0227604	.0917027
baths	.1430244	.0263956	5.42	0.000	.0910886	.1949603
lland	.0807137	.0208169	3.88	0.000	.0397544	.121673
larea	.3403781	.0530962	6.41	0.000	.2359061	.44485
_cons	7.117956	.4891775	14.55	0.000	6.155452	8.08046

Results: With Controls

Key finding: $\log(dist)$ is no longer significant. Controls like $\log(land)$ and $\log(area)$ capture the same information — houses in better areas are both further from the incinerator *and* more expensive.

Roadmap

Exercise 1: FE and FD Equivalence

- Setup
- First-Difference Estimator
- Fixed-Effects Estimator
- Equivalence Proof
- Age as a Regressor
- Random Effects

Exercise 3: Difference-in-Differences (Incinerator)

Exercise 4: Rental Prices and Student Presence

- Setup
- Pooled OLS Estimation
- Interpretation
- Fixed Effects Estimation

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).

Exercise 4

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

Exercise 4

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.
  - Omitted variable bias: city-specific factors (geography, amenities) affect rents
  - 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

Exercise 4

Fixed Effects Estimation Results

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

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

lrcnb	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
y90	.3853219	.0368245	10.47	0.000	.3118615	.4587823
lpop	.0722456	.0882426	0.82	0.417	-.104466	.2489571
lavginc	.2099605	.0664771	4.66	0.000	.1769865	.4429246
pctstu	.0112083	.0041319	2.71	0.009	.0029382	.0194684
_cons	1.409984	1.167238	1.21	0.232	-.9254904	3.744208
sigma_u	.15906877					
sigma_e	.06372673					
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
```

Exercise 4

Fixed Effects and Model Selection

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

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

  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**.