5. Panel Data and Fixed Effects

- When we have longitudinal data we can potentially tackle OVB when the unobserved omitted factors are *stable over time*
- Setting:
  - We can measure the outcome variable for a set of objects (people, firms, ...) at several point in time
  - The key variable of interest (the „treatment“) changes over time
  - We study the association between the change in the treatment variable and the change in the outcome variable
- Here: Consider *Fixed Effects Models* as one important approach

Fixed Effects

- Consider again the potential outcome framework (time index $t = 1, \dots, T$)

$$Y_{C_{it}it} = \begin{cases} Y_{1it} & if \quad C_{it} = 1 \\ Y_{0it} & if \quad C_{it} = 0 \end{cases}$$

- Assume now that

$$E[Y_{0it}|A_i, X_{it}, t, C_{it}] = E[Y_{0it}|A_i, X_{it}, t]$$

where

  – X_{it} is a vector of observed (time varying) covariates and

  – A_i is is a vector of *unobservable* factors that are fixed over time (no time index t! For instance, a person‘s ability or personality)

- The assumption states that C_{it} is as good as randomly assigned conditional on A_i and X_{it}
- This is a sensible identifying assumption whenever any unobserved determinants of the treatment (that also may affect the outcomes beyond the treatment) are time constant

- Consider now the following linear model

$$E[Y_{0it}|A_i, X_{it}, t] = \alpha + X_{it}'\beta + A_i'\gamma + \lambda_t$$

- And assume that the causal effect is a constant ρ

$$E[Y_{1it}|A_i, X_{it}, t] - E[Y_{0it}|A_i, X_{it}, t] = \rho$$

- Hence, we can write

$$Y_{it} = \alpha_i + \lambda_t + \rho C_{it} + X_{it}'\beta + \epsilon_{it}$$

where $\epsilon_{it} = Y_{0it} - E[Y_{0it}|A_i, X_{it}, t]$ and $\alpha_i = \alpha + A_i'\gamma$

- When we impose these assumptions, running a regression will estimate the causal effect ρ of C on Y

- This is a fixed effects model:
  - The α_i are parameters to be estimated (estimating a dummy for every person)
  - The γ_i are time effects that are also estimated (estimating a dummy for every period)

Study

Lazear's (2000) study on Performance Pay at Safelite

- Safelite is a large auto glass company in the US
- Business: replace broken windshields.
- New compensation scheme in January 1994: Piece rate scheme (PPP) replaced hourly-wage scheme in 1994
- The piece rate scheme was phased in over 19 months, starting from the headquarter town.
- The gradual implementation of piece rate allows for within-worker variation identifying the incentive effect of piece rate on effort.
- But: also high turnover rates; many workers also hired after the introduction of the PPP
- In the following:
  - Unit of observation = Worker in a given month;
  - Productivity measure: Average windshields installed by the worker on a given day.

Safelite: Regression analysis

TABLE 3—REGRESSION RESULTS

Regression number	Dummy for PPP person-month observation	Tenure	Time since PPP	New regime	R^2	Description
1	0.368 (0.013)				0.04	Dummies for month and year included
2	0.197 (0.009)				0.73	Dummies for month and year; worker-specific dummies included (2,755 individual workers)
3	0.313 (0.014)	0.343 (0.017)	0.107 (0.024)		0.05	Dummies for month and year included
4	0.202 (0.009)	0.224 (0.058)	0.273 (0.018)		0.76	Dummies for month and year; worker-specific dummies included (2,755 individual workers)
5	0.309 (0.014)	0.424 (0.019)	0.130 (0.024)	0.243 (0.025)	0.06	Dummies for month and year included

Notes: Standard errors are reported in parentheses below the coefficients.
Dependent variable: ln output-per-worker-per-day.
Number of observations: 29,837.

Safelite (continued): What do the worker fixed effects do here?

- Regression without worker fixed effects (row 1)
  - this gives us an estimate of the causal effect of the treatment on the *average performance* of all workers working at a given point in time
  - (when believe that the treatment is as good as randomly assigned conditional on the time period which seems very plausible here)

- However: if we are interested in the causal effect of the treatment on the performance of an *average given worker* this is a „biased“ estimate
  - This is the case when the ability of workers depends on the treatment
  - For instance, when the PPP allows to hire better workers
  - Then $E[Y_{0it}|t, C_{it} = 1] > E[Y_{0it}|t, C_{it} = 0]$

    i.e. workers hired under the PPP would be better even without the PPP
  - In this respect the conditional independence assumption is violated
  - There is a classical selection bias and the PPP dummy should give a too high estimate for the causal effect of the PPP on a *given* worker

What do the worker fixed effects do here?

- The worker fixed effects model (row 2) takes this problem into account
- It imposes the weaker assumption that

$$E[Y_{0it}|A_i, t, C_{it}] = E[Y_{0it}|A_i, t]$$

- When A_i captures the workers unobserved ability this assumption states that for workers *of the same ability* the counterfactual performance is independent of the treatment
- The fixed effects model in a sense estimates the unoberved abilities of the workers (using that a worker‘s performance is observed over many months)
- It thus estimates the causal effect of the PPP conditional on worker‘s abilities
- Note: The model without fixed effects is here not wrong, it estimates something different
  - Without worker fixed effects it estimates the total effect on performance which includes a *selection* and an *incentive effect*
  - With worker fixed effects it estimates the pure incentive effect

Estimating Fixed Effects Models

- Estimating the coefficients of individual dummy variables seems demanding in large panels (1000 employees = 1000 fixed effects)
- However, if we are not interested in knowing the specific values of the individual fixed effects, we can estimate the model in a simpler manner
- Consider

$$Y_{it} = \alpha_i + \lambda_t + \rho C_{it} + X_{it}'\beta + \epsilon_{it}$$

- Now take the average across all time periods $\overline{Y_i} = \frac{1}{T}\sum_{t=1}^{T} Y_{it}$

$$\overline{Y}_i = \alpha_i + \overline{\lambda} + \rho\overline{C}_i + \overline{X}_i'\beta + \overline{\epsilon}_i$$

and substract this from Y_{it}

$$Y_{it} - \overline{Y}_i = \lambda_t - \overline{\lambda} + \rho\left(C_{it} - \overline{C}_i\right) + (X_{it}' - \overline{X}_i')\beta + \epsilon_{it} - \overline{\epsilon}_i$$

→ The α_i are eliminated!

$$Y_{it} - \overline{Y}_i = \lambda_t - \overline{\lambda} + \rho\left(C_{it} - \overline{C}_i\right) + (X'_{it} - \overline{X}'_i)\beta + \epsilon_{it} - \overline{\epsilon}_i$$

- Hence,
  - replace the outcome variable by its deviation from the mean over time
  - replace the explanatory variables by their deviations from their means over time
  - Regress the „de-meaned“ outcome on the „de-meaned“ explanatory variables

  $\rightarrow$ This gives us an estimate of ρ

  $\rightarrow$ We can estimate ρ and β without having to estimate the α_i

- This model is sometimes also called the *within-estimator:*

  It estimates the effect of ρ on Y from the within person variation in C

Python

Fixed Effects Regressions in Python

- Panel regressions in Python can be done with library linearmodels
- Install by !pip install linearmodels
- Import by from linearmodels import PanelOLS
- In order to run a panel regression use a MultiIndex DataFrame that is a DataFrame that uses two indices
  - one index for the entity variable (the omitted time constant variable)
  - one index for the time variable

```
df=df.set_index([‘entity', 'year’])
```

- Then fit the model by

```
reg = PanelOLS.from_formula('y ~ x + EntityEffects + TimeEffects', data=df).fit()
```

- Then print the output with print(reg)
  (Note the different notation to statsmodels: can directly print the results)

Your Task **Fixed Effects**

- Open the notebook in which you estimated the association between Management Practices and ROCE
- For a part of the observations the data set contains panel data
- The paper by Bloom et al. (2012) contains the following table, where the third colums shows the result of a fixed effects regression
- Please replicate this regression using PanelOLS
- Note:
  - The variable account_id contains an identifier for each firm
  - The variable emp contains the number of employees and ppent the capital (fixed assets)
  - You can generate logs by using np.log(x) directly in the formula

	(1)	(2)	(3)
Sector	**Manufact.**	**Manufact.**	**Manufact.**
Dependent variable	**Log (Sales)**	**Log (Sales)**	**Log (Sales)**
Management	0.523*** (0.030)	0.233*** (0.024)	0.048** (0.022)
Ln(Employees)	0.915*** (0.019)	0.659*** (0.026)	0.364*** (0.109)
Ln(Capital)		0.289*** (0.020)	0.244*** (0.087)
Country controls	No	Yes	NA
Industry controls	No	Yes	NA
General controls	No	Yes	NA
Firm fixed effects	No	No	Yes
Organizations	2,927	2,927	1,453
Observations	7,094	7,094	5,561

Your Task — Fixed Effects (Simulated Sales Training Evaluation VII)

Generate the following notebook

```
n=2000
df1=pd.DataFrame(index=range(n))
df1['ability']=np.random.normal(100,15,n)
df1['year']=1
df1['persnr']=df1.index
df1['training']=0
## Now copy the DataFrame (i.e. generate observations for second year)
df2=df1.copy()
df2['year']=2
## Training only in year 2:
df2['training']=(df2.ability+np.random.normal(0,10,n)>=100)
## Generate DataFrame that spans both years by appending the two data frames
df=pd.concat([df1,df2], sort=False)
df['sales']= 10000 + df.training*5000 + df.ability*100 + df.year*2000
              + np.random.normal(0,4000,2*n)
```

Note:

- The script generated a data frame simulating two years of data in which
  - Sales of each subject are observed in each year
  - training is affected by ability
  - subjects are only trained in year 2

Now analyze the generated data:

- Run an OLS regression of sales on training and year
- Define the time and entity indices
- Run a fixed effects regression

But note important caveats:

1. When you want to interpret the results of a Fixed Effects regression causally, a key underlying assumption is the so-called *common trend assumption*
   – That is „treatment" and „control" units follow the same underlying time trend
   – This is a key identifying assumption
2. When the treatment C_{it} hardly varies over time it is hard to evaluate the causal effect effect ρ
   – In the extreme when C_{it} is completely stable then $C_{it} = \overline{C}_{it}$
   – Not identifying a significant effect in the data then does not necessarily imply that there is no such effect
3. Fixed effects can only eliminate *time constant* omitted variables
   – If the treatment is correlated with time varying unobserved variables omitted variable issues remain