5. Diff-in-Diff and Panel Data

- When we have longitudinal data we can try to tackle selection bias/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 least two points 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 Difference-in-Difference and *Fixed Effects Models*

5.1 Difference-in-Difference Estimation

- Consider again the potential outcome framework and introduce a time dimension $t = 1,2$ such that potential outcomes are $Y_{1i}(t)$ and $Y_{0i}(t)$
- Suppose no one is treated at time t $= 1$ and thus $Y_{0i}(1)$ is observed for all
- But some people receive the treatment at date t $= 2$

$$Y_{C_i i}(2) = \begin{cases} Y_{1i}(2) & if \quad C_i = 1 \\ Y_{0i}(2) & if \quad C_i = 0 \end{cases}$$

- We are again interested in estimating the ATT

$$E[Y_{1i}(2) - Y_{0i}(2)|C_i = 1]$$

Note:

- We still cannot observe the counterfactual outcome $Y_{0i}(2)$ for the treated
- But we can observe $Y_{0i}(1)$, that is their outcome before the treatment
- When does this help us the estimate the causal effect?

Now impose the following assumption:

The Parallel Trends Assumption

When

$$E[Y_{0i}(2) - Y_{0i}(1)|C_i = 1] = E[Y_{0i}(2) - Y_{0i}(1)|C_i = 0]$$

then the treated and untreated units have parallel time trends.

Intuition:

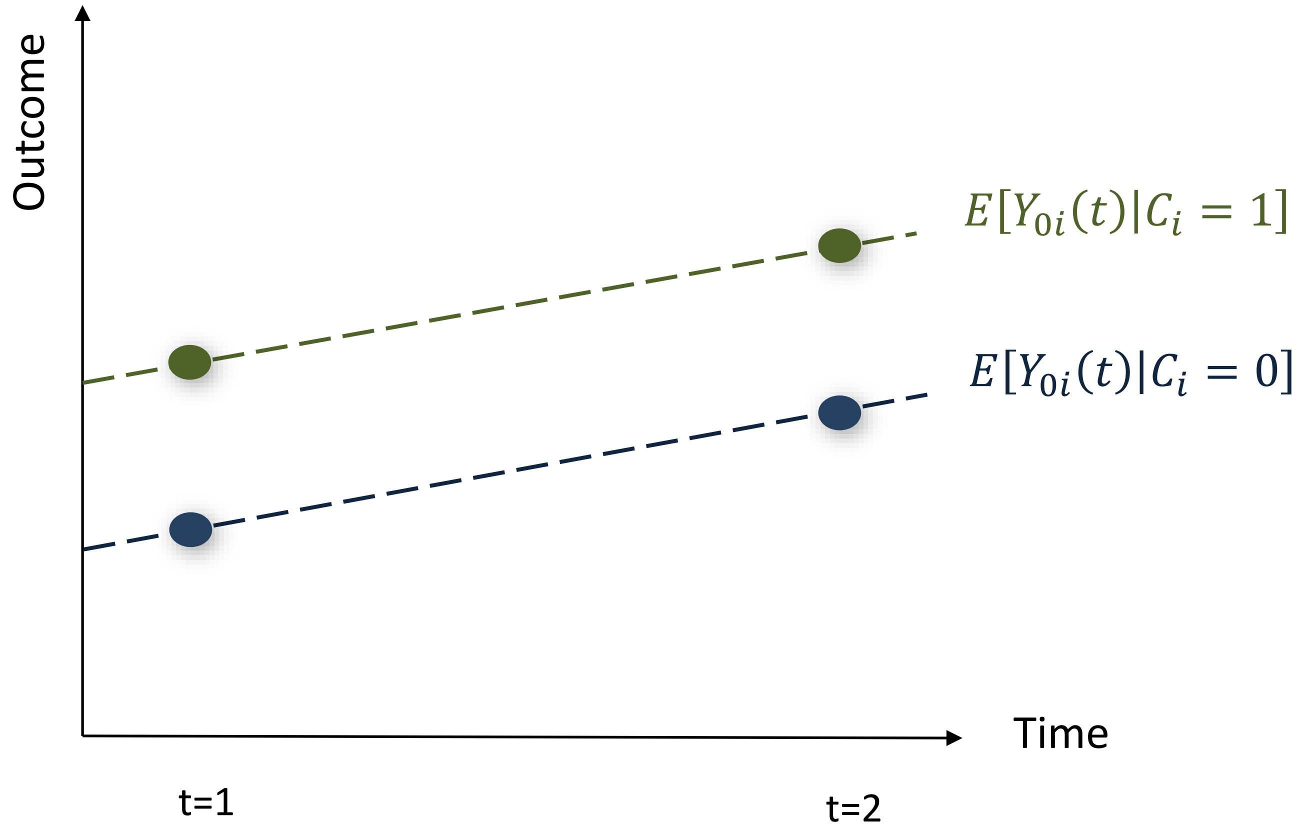
- The treated and untreated units may differ in their levels of performance
- But (without treatment) expected outcomes increase by the same amount from period 1 to period 2

When this assumption holds…

- we can estimate the ATT even with non-random assignment!

The Parallel Trends Assumption:

The same trends without the treatment



Note:

- Now use this common trend assumption

$$E[Y_{0i}(2) - Y_{0i}(1)|C_i = 1] = E[Y_{0i}(2) - Y_{0i}(1)|C_i = 0]$$

- To obtain the expected counterfactual outcome in period 2

$$E[Y_{0i}(2)|C_i = 1] = E[Y_{0i}(1)|C_i = 1] + E[Y_{0i}(2) - Y_{0i}(1)|C_i = 0]$$

- Such that the causal effect:

$$E[Y_{1i}(2) - Y_{0i}(2)|C_i = 1]$$

$$= E[Y_{1i}(2) - Y_{0i}(1)|C_i = 1] - E[Y_{0i}(2) - Y_{0i}(1)|C_i = 0]$$

→ The causal effect is thus just the difference between the

- performance increase in the group that receive $C_i = 1$ and
- performance increase in the group that receive $C_i = 0$

- Key idea: Estimate the counterfactual outcome by assuming that the treated units follow the same time trend

To illustrate the effect it is convenient to rearrange

$$E[Y_{1i}(2) - Y_{0i}(2)|C_i = 1]$$

$$= E[Y_{1i}(2) - Y_{0i}(1)|C_i = 1] - E[Y_{0i}(2) - Y_{0i}(1)|C_i = 0]$$

$$= (E[Y_i(2)|C_i = 1] - E[Y_i(2)|C_i = 0])$$

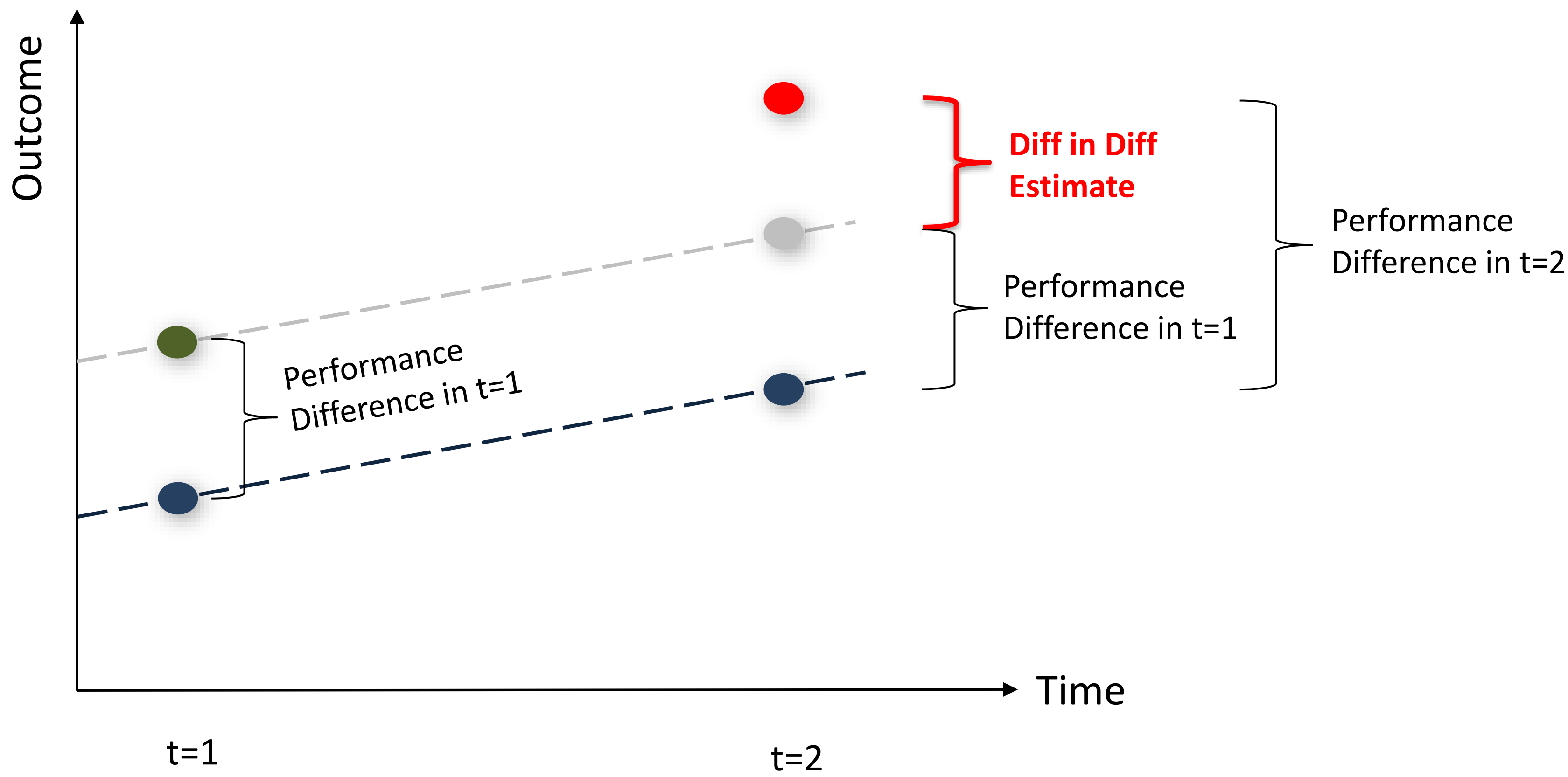
$$-(E[Y_i(1)|C_i = 1] - E[Y_i(1)|C_i = 0])$$

→ The causal effect is just the difference between

- the difference in performance between the groups in $t = 2$
- the difference in performance between the groups in $t = 1$

- Therefore, it is called the **difference-in-difference estimator** (“Diff-in-Diff”, “DiD”)

The Diff-in-Diff Estimator



Your Task — Diff-in-Diff (Simulated Sales Training Evaluation VIII)

- Download the following notebook:
  https://github.com/dsliwka/EEMP2023/blob/main/SalesSimDiD.ipynb

- Go through the simulation code and understand how the data is generated

- Note:
  - tgroup is the group to be trained (it will have value 1 in both periods for those agents who are trained in period 2)
  - training only has value 1 when the agent is indeed trained (in period 2)

- Now plot sales by tgroup and year

  Use: sns.barplot with x='year', y='sales', hue='tgroup'

- Compute average sales by year and group & compute Diff-in-Diff

  df.groupby(['tgroup','year']).sales.mean()

- Save the notebook in your Google Drive

Regression Diff-in-Diff

- Note: we can estimated the causal effect ρ from just working with the differences and replace the expectations with the respective averages
- Typically it is more convenient to simply run a regression
  - Let $TREAT_i$ be a dummy indicating whether an observation comes from the treated group (dummy=1 also before the change!)
  - Let $POST_t$ be a dummy indicating whether an observation comes from a period after the treatment has been implemented
- Then we can regress

$$Y_{it} = \alpha + \beta \cdot TREAT_i + \gamma \cdot POST_t + \rho \cdot (TREAT_i \cdot POST_t) + \epsilon_{it}$$

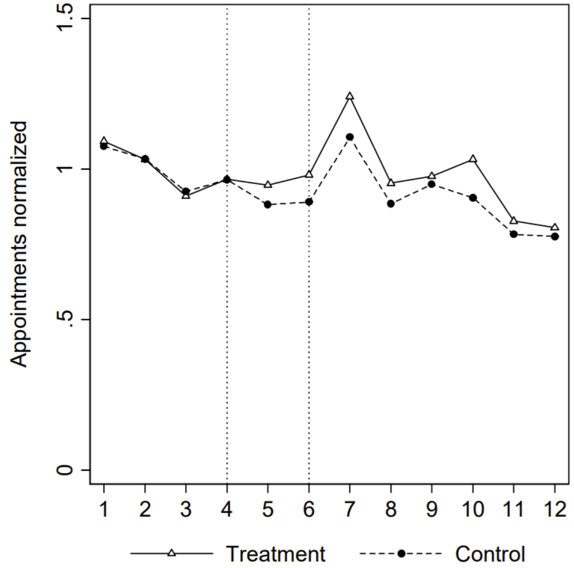
- The coefficient $\tilde{\rho}$ of the interaction term $TREAT_i \cdot POST_t$ yields an estimate of the causal effect
- Note:
  - Regression DiD also provides statistical tests
  - And it can be applied if there are more than two periods

Your Task — Diff-in-Diff (Simulated Sales Training Evaluation VIII)

- Open the notebook that simulated the sales panel data (SalesSimDiD.ipynb)
- Now perform a regression DiD using Statsmodels
- Note: it is convenient to generate a dummy variable post which takes value 1 only in year 2
- Interpret the regression coefficients
- Save the notebook

A note on the common trend assumption

- The common trend assumption can of course not be tested in the treatment period (we don‘t know the counterfactual)
- But: if you have more periods before the intervention it is very useful to check whether it holds in these periods



From Manthei/Sliwka (2019)

Outlook: Synthetic Control Method/Synthetic Difference-in-Differences

See Abadie (2021), Arkhangelsky et al. (2021)

- In Synthetic Control methods you create an artificial control group
- Basic idea: for each “treated“ unit generate a “synthetic” control unit
  - The synthetic control unit’s “outcome” is the weighted average of several other (real) units
  - Weights are derived by minimizing the difference in time trends prior to intervention

Study

Dismissal Protection and Hiring

- Bauernschuster (2013) studies the effect of a firm-size threshold of the German dismissal protection law in 2004 on the hiring behavior of small firms
- From 1999 until the end of 2003 the dismissal protection law applied only to establishments with more than five (full time equivalent) workers
- In 2004, this threshold was shifted up to ten full-time equivalent workers
- Dismissal protection regulation was abandoned for workers hired after December 31, 2003 by establishments with 6 to 10 FTE employees
- Bauernschuster studies the effect on hiring applying a diff-in-diff strategy
  - „Treatment group“: establishments with 6-10 employees
  - „Control group“: establishments with 11-20 employees
- Uses data from the IAB establishment panel (on which the LPP is based)
- Dependent variables are hiring rates and total number of hirings per establishment in the first half of the year

Dismissal Protection and Hiring (continued)

The Dynamics of Hiring

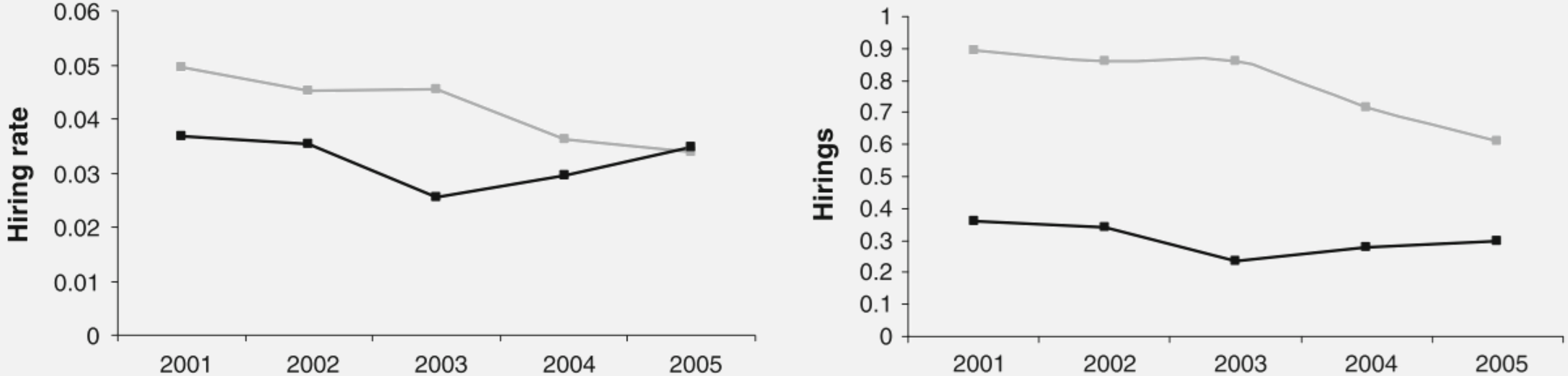


Fig. 2 Dynamics of hirings in treatment and control group. The *left figure* shows average hiring rates for treatment (*black line*) and control groups (*grey line*) over time. The *right figure* shows average absolute hirings for treatment (*black line*) and control groups (*grey line*) over time. The treatment group comprises establishments with more than five and up to ten full-time equivalent workers while the control group consists of establishments with more than ten and up to 20 full-time equivalent employees. Data source: IAB Establishment Panel

Bauernschuster (2013)

Dismissal Protection and Hiring (continued)

Table 1 DiD estimates on hirings

Parameter	Hiring rate (1)		Hiring rate (2)		Hiring rate (3)	
	Coefficient	Standard error	Coefficient	Standard error	Coefficient	Standard error
DiD 2004	0.013**	0.007	0.016**	0.007	0.020**	0.009
DiD 2005	0.021**	0.009	0.021**	0.009	0.020**	0.009
Treatment group	−0.020***	0.006	−0.019***	0.007	−0.024***	0.006
Year 2004	−0.009*	0.004	−0.013***	0.004	−0.014***	0.006
Year 2005	−0.012**	0.005	−0.013***	0.005	−0.013*	0.006
Control set 1	No		Yes		Yes	
Control set 2	No		No		Yes	
N	1,749		1,658		1,285	
R^2	0.0059		0.0725		0.1197	

The table reports the results of OLS difference-in-differences regressions with hiring rates as the dependent variable. The treatment group comprises establishments with more than five and up to ten full-time equivalent workers while the control group consists of establishments with more than ten and up to 20 full-time equivalent employees. The baseline year is 2003. Specification (1) includes no further controls. In specification (2), we additionally control for capital stock, works council, collective labor agreement, age, and industry (control set 1). In specification (3), we add the ratio of female workers, ratio of unqualified workers, ratio of apprentices, wage per worker in the previous year, value added per worker in the previous year as well as net hirings in the previous year as further controls (control set 2). Standard errors are clustered at the establishment level. ***, **, * denote significance at the 1, 5, and 10% levels, respectively. Data source: IAB Establishment Panel

Bauernschuster (2013)

5.2 Fixed Effects

- In the DiD approach we controlled for the group that the agent belonged to (i.e. the treated or untreated)
- The key idea was to at the same time take out
  - “level effects“, i.e. differences between the groups that are there at the outset and stay constant over time
  - and “time effects”, i.e. differences between the periods that are driven by the same underlying time trends in the groups
- We can also apply the same logic, but control for
  - level effects at the level of an individual subject rather than a group of subjects
  - time effects across multiple periods
  - as well as time-varying further control variables
- Then we estimate a so-called ***Fixed Effects Model***

- Again use potential outcome framework (with time index $t = 1, \dots, T$)

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

- Let X_{it} be again a vector of **observed time varying** covariates
- Consider a situation where there is a vector of **unobserved** factors A_i that affect Y_{Cit} but **do not depend on time**
  - This may for instance be a person‘s ability or personality, or
  - a firm‘s location or the quality of its management…
- When these unobservable factors A_i are correlated with C_{it} such that

$$Cov[A_i, C_{it}] \neq 0$$

we have an omitted variable bias when regressing Y_{it} on C_{it}

But: Assume now that $E[Y_{0it}|A_i, X_{it}, t, C_{it}] = E[Y_{0it}|A_i, X_{it}, t]$

We thus make the following key assumption

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

Note:

- The assumption states that C_{it} is as good as randomly assigned conditional on A_i and X_{it}
- That is, if we compare two individuals who share the same values for X_{it} (observable) and A_i (unobservable), we assume that their potential outcomes do not differ
- This is a sensible identifying assumption whenever any unobserved determinants of the treatment (that also may affect the outcomes beyond the treatment) are **constant over time**

- Consider now the following linear model

$$Y_{it} = \alpha + \rho C_{it} + X_{it}'\beta + A_i'\gamma + \lambda_t + \epsilon_{it}$$

- Thus the causal effect of the treatment C is again a constant ρ, i.e.

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

- We can now replace $\alpha_i = \alpha + A_i'\gamma$ and thus rewrite the model as

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

- Then running a regression will estimate the causal effect ρ of C on Y as $E[Y_{0it}|A_i, X_{it}, t, C_{it} = 1] = E[Y_{0it}|A_i, X_{it}, t, C_{it} = 0]$

- This is a **fixed effects model**:
  - The α_i are parameters to be estimated (i.e. estimating a dummy for every person/firm/object)
  - The λ_t are time effects that are also estimated (i.e. estimating a dummy for every period)

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

    dfp=df.set_index([‘entity', 'year’])
  - Note: It is convenient to use a different DataFrame for the panel regressions
- Then fit the model by

  reg = PanelOLS.from_formula('y ~ x + EntityEffects + TimeEffects', data=dfp).fit()
- Then print the output with print(reg)

  (Note the different notation to statsmodels: can directly print the results)

Python

Fixed Effects Regressions in Python

- It is likely that observations of the same entity (firm, person etc.) are not independent
- Hence, when you run a fixed effects panel regression it makes sense to estimate clustered standard errors where the cluster is determined by the respective entity
- In PanelOLS you can do this by adding a cov_type="clustered" option in the fit()-method:

```
reg = PanelOLS.from_formula('y ~ x + EntityEffects + TimeEffects’,
    data=dfp).fit(cov_type="clustered", cluster_entity=True)
```

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 (consider the variables account_id containing a firm identifier and year)
- Bloom et al. (2012) report the following table, where the third colums shows the result of a fixed effects regression
- Replicate the regression using PanelOLS
- You may also replicate it using a simple OLS with dummies for each firm (recall +C(..))
- Note: Further relevant variables are
  - emp: number of employees
  - ppent capital (property, plant & equipment)
  - 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 IX)

- Open again the SalesSimDiD notebook
- Define the time and entity indices
- Run a fixed effects regression
- Save the notebook

Note: Using the estimated fixed effects

Someestimes the α_i estimated for the individuals/firms are useful themselves

Examples:

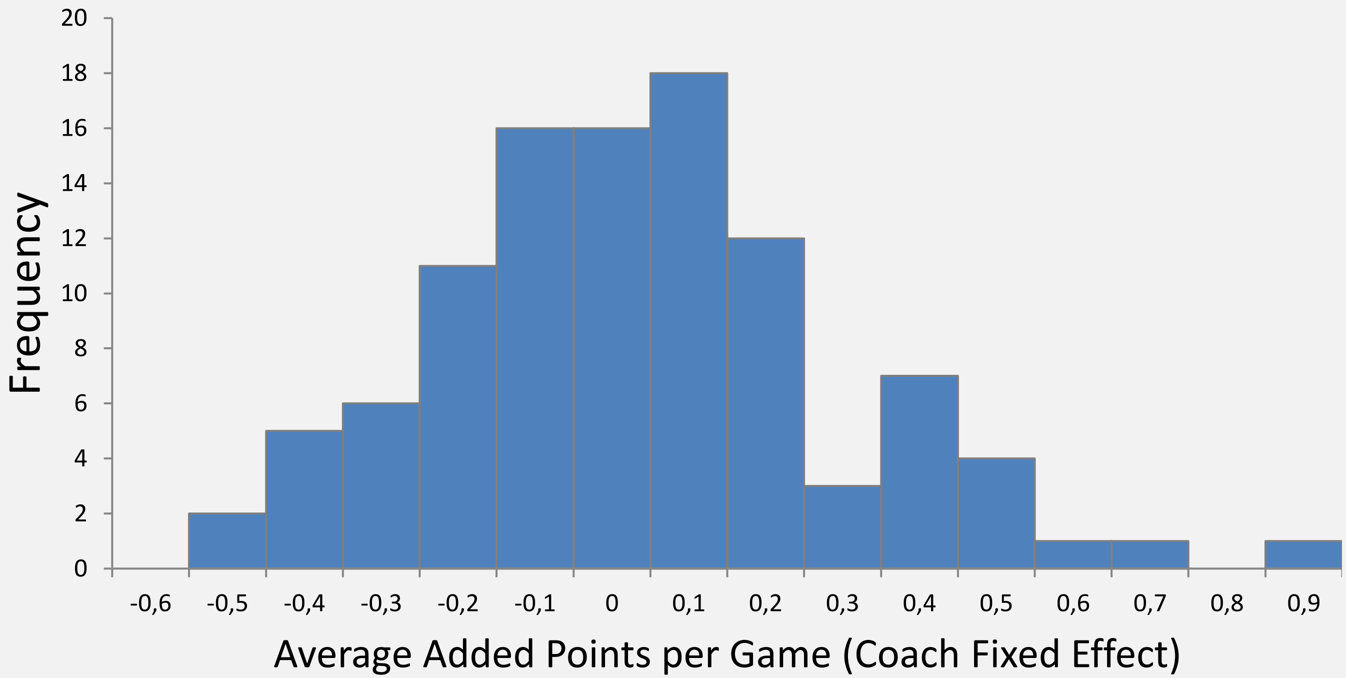
- How much do individual managers matter for firm behavior and performance? Bertrand and Schoar (2003) use
  - Manager-firm matched panel dataset, where they can track individual top managers across different firms over time.
  - Estimate how much of the variation in firm practices can be attributed to manager fixed effects
- Are there Team-Players, i.e. people that makes teams systematically better? Weidmann/Deming (2021) run an experiment where they
  - Repeatedly randomly assign people to group tasks
  - Estimate team-player fixed effects controlling for subjects‘ ability for the tasks
  - Find that there are people who by their social skills systematically make teams better

Study

Estimating the Value Added of Managers

Mühlheusser/Schneemann/Sliwka/Wallmeier (2017) consider soccer coaches

- Coaches frequently move between teams
- Allows to disentangle the effect of the coach from the strength of the club by estimating models with manager and team fixed effects
- Data on 20 seasons of the German Bundesliga



Conclusion

1. When you want to interpret the results of a DiD or Fixed Effects regression causally, a key underlying assumption is again the **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