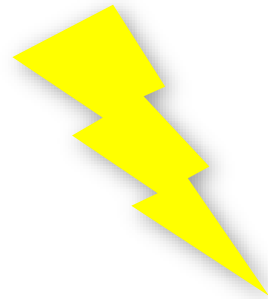


5. Regression and Causality

Reconsider the Bloom et al. (2012) studies that we saw in the introduction:

- Randomly drawn sample of 10,000 organizations.
- Authors define “best” management practices
- Firms adopting theses practices are more productive/profitable, grow faster and survive longer.
- Potential conclusion: Firms must “just” adopt these management practices and they will automatically be more productive/profitable, grow faster and survive longer



What is the problem with such a conclusion?

Recall:

- Regressions give us an approximation to Conditional Expectations
- Conditional Expectations *predict* the outcome of a variable on the basis of other variables
- If we know $E[Y|X]$ we can tell the following:
 - If you tell me a value of X (say x), what is the average value of Y we can expect when $X = x$?
 - *“Which job satisfaction can we expect in firms with performance pay as opposed to firms without?”*
- While this is a powerful property, it does not necessarily tell you:
 - If you change the value of X (say from x_1 to x_2) for objects in the population how is their average value of Y affected by this?
 - *“When we introduce performance pay, how would this change job satisfaction, on average?”*
- Typical reason: there are other variables affecting both X and Y

Counterfactuals and Causality

- The question whether a regression is causal boils down to the question whether the conditional expectation is causal
- If the CEF is causal we can estimate causal effects with a regression analysis
- To answer this question it is very useful to think about *potential outcomes* or *counterfactuals*
“What would have happened, when a different decision had been made?”
- This seems hard to answer!
(But it is often still a useful thought experiment in real life)
- But we sometimes can say something about the counterfactual using data
- When this is the case empirical research becomes very powerful!

5.1 Thinking about Potential Outcomes

- Suppose we want to investigate whether
 - a certain management practice
(performance pay, wage increase, training,...)
 - causally affects some outcome variable Y_i
(job satisfaction, performance,...)
- Let $C_i \in \{0,1\}$ be a dummy variable indicating whether the practice is implemented for person i
- What we would like to know is: what is the value of Y_i
 - if $C_i = 1$ (“person i is treated”)
 - if $C_i = 0$ (“person i is not treated”)
- Let this *potential outcome* be

$$Y_{C_i i} = \begin{cases} Y_{1i} & \text{if } C_i = 1 \\ Y_{0i} & \text{if } C_i = 0 \end{cases}$$

- The *causal effect* of C_i on Y_i is now $Y_{1i} - Y_{0i}$

The problem is:

- when we implement the practice we only observe Y_{1i}
- when we do not implement the practice we only observe Y_{0i}

In real life we do not observe the *counterfactual*

- What would have happened if we had decided differently?
- The *observed outcome* is Y_i where
$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i}) \cdot C_i$$
- Running a simple regression (or comparing means) in a sample yields
 - $E[Y_i|C_i = 1]$ and
 - $E[Y_i|C_i = 0]$
- Here, one may be tempted to interpret

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

as the causal effect of C on Y

But note that

$$\begin{aligned} & E[Y_i|C_i = 1] - E[Y_i|C_i = 0] \\ &= E[Y_{1i}|C_i = 1] - E[Y_{0i}|C_i = 0] \\ &= E[Y_{1i}|C_i = 1] - E[Y_{0i}|C_i = 1] + E[Y_{0i}|C_i = 1] - E[Y_{0i}|C_i = 0] \\ &= E[Y_{1i} - Y_{0i}|C_i = 1] + E[Y_{0i}|C_i = 1] - E[Y_{0i}|C_i = 0] \end{aligned}$$

- The causal effect of C on the group that is treated ($C = 1$) is

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

- It is called the *average treatment effect on the treated (ATT)*
 - Very often this is what we want to know
 - “*Has job satisfaction increased in a group of employees because this group now receives performance pay?*”
- But: the regression coefficient may not estimate the ATT
 - It includes $E[Y_{0i}|C_i = 1] - E[Y_{0i}|C_i = 0]$
 - This is the *selection bias*

We can thus decompose:

$$\underbrace{E[Y_i|C_i = 1] - E[Y_i|C_i = 0]}_{\text{Observed difference in outcome}} \\ = \underbrace{E[Y_{1i} - Y_{0i}|C_i = 1]}_{\substack{\text{Average treatment effect} \\ \text{on the treated}}} + \underbrace{E[Y_{0i}|C_i = 1] - E[Y_{0i}|C_i = 0]}_{\text{Selection bias}}$$

- If $E[Y_{0i}|C_i = 1]$ differs from $E[Y_{0i}|C_i = 0]$
 - Treated and untreated individuals differ
 - $E[Y_{0i}|C_i = 0]$ is not the counterfactual outcome for the treated
- Then the regression estimates are biased estimates of the causal effect!

Example: Does a university education increase earnings?

- $E[Y_{0i}|C_i = 1]$ is the wage somebody who attended a university would earn when not having attended university
- It is very likely that $E[Y_{0i}|C_i = 1] > E[Y_{0i}|C_i = 0]$
- Hence, we would overestimate the true returns to a university education

Your Task

Simulated data set: Evaluation of a sales training

- Write a script that generates a fictitious data set with 10000 observations

```
n=10000  
df=pd.DataFrame(index=range(n))
```
- Generate a normally distributed random variable *ability* with mean 100 and std. deviation 15:

```
df['ability']=np.random.normal(100,15,n)
```
- Generate a dummy variable *training*:

```
df['training']=(df['ability']+np.random.normal(0,10,n))>=100)
```

(Hence, more able people have a higher likelihood to be trained)
- Generate a variable sales:

```
df['sales']= 10000 + df['training']*5000 + df['ability']*100  
+ np.random.normal(0,4000,n)
```
- This is the true causal relationship: the training increases sales by 5000
- But suppose we as researchers cannot observe *ability*
- Run a regression of sales on training & interpret the results (& save the notebook as SalesSim1)

Recall:

- A regression estimates the Conditional Expectation Function
- The CEF gives us $E[Y_i|C_i = 1] - E[Y_i|C_i = 0]$
- It identifies a causal effect only if $E[Y_{0i}|C_i = 1] - E[Y_{0i}|C_i = 0] = 0$

This is satisfied if C_i is *independent* of (Y_{0i}, Y_{1i})

- That is neither Y_{0i} nor Y_{1i} are systematically different for people with different realizations of C_i
- Let the symbol \perp indicate independence
- If the condition

$$(Y_{0i}, Y_{1i}) \perp C_i$$

is satisfied we can use simple regressions (or here mean comparisons) to identify causal effects

5.2 Why are Experiments so Important?

- Suppose we have a randomized controlled experiment
 - That is C_i is randomly assigned to the individuals i
 - In turn, C_i is by construction independent of Y_{i0}
 - Hence, $E[Y_{0i}|C_i = 1] = E[Y_{0i}|C_i = 0]$
 - The selection bias is eliminated!
 - We obtain an unbiased estimator of the causal impact of C in the population
- In that case

$$E[Y_i|C_i = 1] - E[Y_i|C_i = 0] = E[Y_{i1} - Y_{i0}]$$

- A simple comparison between the averages of treatment and control yields an unbiased estimate of the causal effect
- The same holds for a regression on a treatment dummy

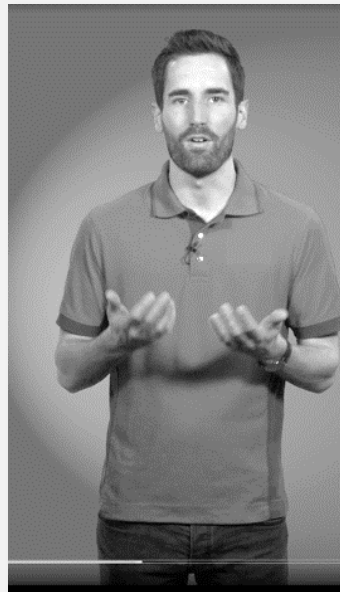
Manthei/Sliwka/Vogelsang (2019) investigate the effect of bonus payments and a training program on profits in a retail chain

Background:

- Store managers previously had pure fixed wages and not much information about profit margins of different products
- Idea: giving more information on margins and training them raises awareness and allows them to make more informed decisions to raise profits
- Bonus payment should raise incentives to do so

360 Stores *randomly* assigned to one of four treatments:

- Control
- Only bonus
- Only learning (E-learning & information on margins)
- Bonus + learning



Deckungsbeitrag 1 (DB1)

Nettoumsatz

- Wareneinsatz
- Inventurdifferenzen
- Verluste
- Personalkosten

= DB1

}

beeinflussbare
Kosten

zusätzliche Informationen zur Spannen- Betrachtung im MDE:

SP1 - sehr hoch (Top 20% der Artikel)

SP2 - hoch

SP3 - durchschnittlich

SP4 - etwas niedrig

SP5 - niedrig (Niedrigste 20% der Artikel)

DB-Position		Apr2017	
		Bud	Ist
9010	Nettoumsatz		
9015	Wareneinsatz		
9020	Spanne		
9050	Inventurdifferenz		
.	davon Inventuren		
.	davon Verluste		
9250	Personalkosten		
Vereinfachter DB1			

Regression:

Table A3 – Regressions only using Treatment Period

	(1) OLS	(2) log OLS	(3) OLS	(4) log OLS
Treatment Effect BONUS	411.41 (342.63)	0.0162 (0.0104)	507.62 (357.36)	0.0206* (0.0106)
Treatment Effect INFORMATION	721.29** (336.68)	0.0149 (0.0111)	706.15** (330.44)	0.0297*** (0.0109)
Treatment Effect BONUS & INFORMATION	907.57** (364.29)	0.0221** (0.0106)	734.83* (391.75)	0.0238** (0.0105)
Time FE	Yes	Yes	Yes	Yes
Store FE	No	No	No	No
District Manager FE	No	No	No	No
Store Manager FE	No	No	No	No
Refurbishments	Yes	Yes	Yes	Yes
Planned Profits	Yes	Yes	Yes	Yes
Further Controls	No	No	Yes	Yes
N Observations	1086	1086	1068	1068
N Stores	363	363	356	356
N Cluster	56	56	56	56
Within R^2				
Overall R^2	0.9260	0.9129	0.9278	0.9082

Note: The table reports results from ordinary least squares estimations with profits at the store level as the dependent variable in columns 1&3 and the log value in columns 2&4. Regressions control for the mean of profits from January 2016 - March 2017 and the randomization pair. All regressions further control for possible refurbishments of a store and the companies planned profits. Columns 3&4 further control for variables with slight imbalance between treatments (gender, FTE, age of the store, age of the store manager, tenure of the store manager). Observations are excluded once a store manager switched the store during the treatment period. Robust standard errors are clustered on the district level at the start of the experiment and displayed in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.001$.

Short recap:

- What does the R^2 in a regression table mean?

N Observations	1086	1086	1068	1068
N Stores	363	363	356	356
N Cluster	56	56	56	56
Within R^2				
Overall R^2	0.9260	0.9129	0.9278	0.9082

- $$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$
- Equal to the share of the variance in the data that can be “explained” by our regression model
 - If $R^2=1$, the regression model would capture all of the variation in Y,
 - if $R^2=0$, the regression model capture none of the variation in Y
- Note:
Here the R^2 is large as profits (DB) are close to plan profits (DP_p)

Your Task

Simulated data set: Evaluation of a sales training II

- Open your SalesSim1 notebook, save it as SalesSim2 to generate a different simulation, and run the whole notebook
- Now suppose that there is new training program which is *randomly assigned*
- Add a cell at the end of the notebook to generate a dummy variable *training2* which takes value 1 for 5% randomly chosen individuals

```
df['training2']=(np.random.uniform(0,100,n)<=5)
```
- **Note:** `np.random.uniform(0,100,n)` generates a vector of `n` random variables that are uniformly distributed between 0 and 100
- Assume that this new program also raises sales by 5000:

```
df['sales']=df['sales']+df['training2']*5000
```
- Run a regression of sales on training and training2
- Interpret the results & save the notebook

Side Remark: The Strategy Method in Experimental Economics

- It is typically stated in econometrics text books that we can never observe the exact counterfactual
- In real life this seems very true: How can we observe a person doing something and at the same time not doing it?
- But: Powerful tool in experimental economics by Selten (1967)
 - Player states her strategy profile in a sequential game in advance
 - Computer then exactly follow the strategy profile
- We observe both Y_{i1} and Y_{i0} and can thus even measure a causal effect for each individual

4.3 Control Variables & Omitted Variable Bias

- But what if we do not have an experiment?
- In multiple regression we “control for” other covariates X_i
- (When) does this help us to identify causal effects?
- We can write $E[Y_i|X_i, C_i = 1] - E[Y_i|X_i, C_i = 0]$

$$= E[Y_{1i} - Y_{0i}|X_i, C_i = 1] + E[Y_{0i}|X_i, C_i = 1] - E[Y_{0i}|X_i, C_i = 0]$$

The Conditional Independence Assumption (CIA)

If the *conditional independence assumption holds*, i.e.

$$Y_{ci} \perp\!\!\!\perp C_i \mid X_i \text{ for all values of } c,$$

(conditional on X the treatment C_i is independent of potential outcomes), then

$$E[Y_i|X_i, C_i = 1] - E[Y_i|X_i, C_i = 0] = E[Y_{1i} - Y_{0i}|X_i, C_i = 1],$$

i.e. the difference in conditional expectations has a causal interpretation.

Note:

- This is a weaker property than the independence assumption
 $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp C_i$ above
- We do not need that C_i is independent from potential values
- But it needs to be independent for people who have the same values for a set of observable co-variates

The ***Conditional Independence Assumption*** is crucial in many applications

- Useful question: is C_i as good as randomly assigned conditional on X_i ?
- Or, in other words: are the variables in X_i the only reason why (Y_{0i}, Y_{1i}) are correlated with C_i ?
- This is also called the “*selection on observables*” assumption: i.e. selection into the treatment only depends on observable variables X_i ; beyond that it is random
- In that case a regression which controls for X_i (in a proper manner) has a causal interpretation

Analogously: Continuous “treatment” variable

- Think in terms of a causal model $Y_{si} \equiv f_i(s)$
 - $f_i(s)$ describes how an object i (person, firm, ...) responds to changes in some variable s
 - or: determines the outcome for all *potential* realizations of s
- Now let $f_i(s) \equiv f(s, X_i)$
- Distinction between CEF $E[Y_i | S_i, X_i]$ (or regression as its approximation) and causal model $f(s, X_i)$
 - The CEF describes the mean of Y when I draw objects with the same values of (S_i, X_i) from the population (and regressions approximate these conditional expectations)
 - The causal model $f(s, X_i)$ describes how Y changes when I change s
- Regressions approximate the causal model when the CIA holds

- Suppose we want to predict wages and that our causal model is

$$f_i(s) = \alpha + \rho s + \eta_i$$

where, for instance, s are the years of education

- That is we assume that the wage is a linear function of s and η_i captures unobserved factors that determine potential earnings
- When s depends on factors in η_i a regression of wages on s will be biased
- But when we can observe these components and we can decompose

$$\eta_i = X_i' \gamma + v_i$$

where γ is a vector of population regression coefficients which satisfies

$$E[\eta_i | X_i] = X_i' \gamma$$

- Then v_i and X_i are uncorrelated and by the CIA we have that

$$E[f_i(s) | X_i, s_i] = E[f_i(s) | X_i] = \alpha + \rho s + E[\eta_i | X]$$

- Then the residual v_i in $Y_i = \alpha + \rho s + X_i' \gamma + v_i$ is uncorrelated with both X_i and s and the regression of Y on s has a causal interpretation when we control for X

A Note on Terminology: *Identifying Assumptions*

- When we use *observational data* (that is data that we observe but which has not been generated by an experiment), we can never be entirely sure that our regression captures the causal effect
- But still for many questions it is hard to design an appropriate field experiment
- We can (and should) still try to say something about causality
- In order to do so, we typically state so called *identifying assumptions*
 - That is: we make clear under what conditions our empirical approach would capture a causal effect
- The conditional independence assumption is an example for such an identifying assumption

Omitted Variable Bias

- Assume that there is a population regression capturing causal effects

$$Y_i = \alpha + \rho \cdot C_i + \gamma \cdot X_i + v_i \quad (4)$$

such that v_i is uncorrelated with all regressors

- When the CIA holds, then ρ is equal to the coefficient in the linear causal model
- But assume that we cannot (or do not) include X_i and estimate

$$Y_i = \tilde{\alpha} + \tilde{\rho} \cdot C_i + \eta_i$$

- The short regression yields (use equation (4))

$$\begin{aligned} \tilde{\rho} &= \frac{\text{Cov}[C_i, Y_i]}{V[C_i]} = \frac{\text{Cov}[C_i, \alpha + \rho \cdot C_i + \gamma \cdot X_i + v_i]}{V[C_i]} \\ &= \rho + \frac{\text{Cov}[C_i, \gamma \cdot X_i + v_i]}{V[C_i]} \\ &= \rho + \gamma \cdot \frac{\text{Cov}[C_i, X_i]}{V[C_i]} \end{aligned}$$

- If $\text{Cov}[C_i, X_i] \neq 0$ the coefficient is biased (“omitted variable bias”)

$$\tilde{\rho} = \rho + \gamma \cdot \frac{Cov[C_i, X_i]}{V[C_i]}$$

- But $\frac{Cov[C_i, X_i]}{V[C_i]}$ is the coefficient in a regression

$$\underbrace{X_i}_{\text{Omitted variable}} = \delta_0 + \delta_c * \underbrace{C_i}_{\substack{\text{Included} \\ \text{"endogenous"} \\ \text{variable}}} + v_i$$

- Then

$$\tilde{\rho} = \frac{Cov[C_i, Y_i]}{V[C_i]} = \rho + \gamma \cdot \delta_c$$

- Hence, Angrist/Pischke's short description of OVB: *"Short equals long plus the effect of omitted times the regression of omitted on included"*
- Note: this also holds when X_i is a vector

- We studied the association between wages and education using the NLSY97
- We found that the CEF of wages is strongly increasing in education
 - But is this a causal effect?
 - It seems quite likely that there is omitted variable bias
- In 1997 and early 1998, the NLSY97 respondents were given the *Armed Services Vocational Aptitude Battery* (ASVAB) which comprises 10 tests that measure knowledge and skills in a number of areas
- First: Regress wages in 2012 on dummy variables for educational degrees
- In a second step:
 - Control for a standardized ASVAB score (Mathematical Knowledge, Arithmetic Reasoning, Word Knowledge, and Paragraph Comprehension)
 - Standardized by subtracting the mean and dividing by the standard deviation

- Open the SalesSim1 notebook
- Again regress
 - *Sales* on *training*
 - *Sales* on *training* and *ability*
- Regress *ability* on the “endogenous” variable *training*
How do you interpret the coefficient of *training* in the last regression?
(Note this is not causal! but think of CEF interpretation of regression)
- Compute the OVB using this coefficient
- Interpret the size of the OVB

Good and Bad Control Variables

Why use control variables? Two purposes:

- Avoiding omitted variable bias:
 - Control variables can help to satisfy the *conditional independence assumption*
 - When we can think of our variable of interest as being as good as randomly assigned conditional on the set of control variables the estimate has a causal interpretation
- Raising statistical power
 - If standard errors are large, our estimates of β are imprecise
 - It is harder to reject the null hypothesis that $\beta = 0$
 - Standard errors are larger when residuals have a higher variance
 - Including control variables can help even if they do not reduce OVB as long as they reduce noise

Your Task

Control variables to reduce noise (Simul. Sales Training IV)

- Open the simulated sales data notebook SalesSim1 and save it as SalesSim3
- Before the line in which you generated the sales variable, generate uniformly distributed variable experience `df['experience']=np.random.uniform(0,30,n)`
- Change the notebook such that sales now also depends on experience:
$$\text{df['sales']} = 10000 + \text{df['training']} * 5000 + \text{df['ability']} * 100 \\ + \text{df['experience']} * 15000 + \text{np.random.normal}(0,4000,n)$$
- At the end of the cell add separate commands to regress sales on
 - training
 - training and experience
 - training and ability
 - training, ability, and experience
- Note: here convenient to use `summary_col` to have the results side by side in a table, i.e. `print(summary_col([reg1,reg2,reg3,reg4], stars=True))`
- Run the do-file several times, comparing the four regression results (inspect the training coefficient and its standard error) & save the notebook

But note: More control is not always better! (See Angrist/Pischke, pp. 64)

- Some variables are bad control variables when we want to estimate causal effects
- Bad control variables are variables that are *themselves affected by our variable of interest*
- The reason is that they introduce a bias in the estimated causal effect of our variable of interest
- Intuitively: a bad control variable may “pick up” a part of the causal effect
- Good control variables are fixed when the variable of interest is determined such that they cannot be affected by this variable

Example:

- We are interested in the earnings difference between university graduates and others
- Think of a regression of earnings on a dummy indicating whether the person has a university degree
- People can work in two occupations, white-and blue-collar jobs
- Suppose you look at a regression model that includes white-collar status
- But: University degree increases likelihood to work in highly paid white-collar job
- Occupation thus highly correlated with
 - education (university education affects occupation)
 - earnings
- Comparison of earnings conditional on white-collar status does *not* have a causal interpretation

- Open again the notebook SalesSim1 (the first version)
- Generate a variable examScore that is the result of an exam the sales agent took part in
$$\text{df['examScore']} = 100 + \text{df['ability']} + \text{df['training']} * 50 + \text{np.random.normal}(0, 5, n)$$
- At the end of the do file add commands to
 - regress sales on training
 - regress sales on training and examScore
- Compare the coefficient of training in the two regression outputs

- To conclude, some variables are bad control variables when we want to estimate causal effects
- Important: Timing matters!
 - Bad control variables are variables that are themselves affected by our variable of interest
 - A bad control variable may “pick up” a part of the causal effect
 - Good control variables are fixed before the variable of interest is determined
- But typical problem in applied research: Timing often not really known
 - What has happened first?
 - We need to have convincing arguments for why control variables are themselves not caused by the variable of interest

5.4 Measurement Error

- The previous considerations suggest that multiple regressions come close to causal effects when there are proper control variables
- What if we can imperfectly measure variables, i.e. there is *measurement error*
- Suppose that we have a causal model $f_i(x) = \alpha + \gamma \cdot x + v_i$
- Suppose that
 - we cannot measure the X_i precisely,
 - but measure $\tilde{X}_i = X_i + \eta_i$ where $\eta_i \sim N(0, \sigma_\eta^2)$
- If we run a regression

$$Y_i = \tilde{\alpha} + \tilde{\gamma} \cdot \tilde{X}_i + \varepsilon_i$$

- We obtain a coefficient (use equation (4))

$$\tilde{\gamma} = \frac{\text{Cov}[\tilde{X}_i, Y_i]}{V[\tilde{X}_i]} = \frac{\text{Cov}[X_i + \eta_i, \alpha + \gamma \cdot X_i + v_i]}{V[X_i + \eta_i]}$$

But

$$\frac{\text{Cov}[X_i + \eta_i, \alpha + \gamma \cdot X_i + v_i]}{V[X_i + \eta_i]} \\ = \gamma \cdot \frac{\sigma_x^2}{\sigma_x^2 + \sigma_\eta^2}$$

- This is strictly smaller than the true causal effect γ as $\frac{\sigma_x^2}{\sigma_x^2 + \sigma_\eta^2} < 1$
- This is called the *attenuation bias*:
If there is measurement error in a variable regressions underestimate its causal effect
- Note: If you still observe a positive and significant effect
 - you are safe to conclude that the variable measured with error has an impact
 - the true effect is even larger

Measurement Error in a Covariate

- However, measurement error can be more problematic, if it affects an important control variable
- Consider a causal population regression (i.e. the CIA holds $Y_{ci} \perp\!\!\!\perp C_i \mid X_i$)

$$Y_i = \alpha + \rho \cdot C_i + \gamma \cdot X_i + v_i$$

- Assume that we are interested in the effect of C which we can measure precisely
- But we measure only a “proxy” $\tilde{X}_i = X_i + \eta_i$

Example: C_i is a training dummy and X_i some test outcome for ability

- The problem is:
 - measurement error leads to a (downward) biased estimate of γ
 - we therefore do not properly condition on X
 - we may get a biased estimate of C as the coefficient of C captures some of the remaining influence of X on Y
- Note: Section in Wooldridge on plug-in solution a bit misleading

- Open again SalesSim1 (the first version)
- Suppose now that we cannot measure ability but have a proxy variable iq which is $ability + \varepsilon$ where $\varepsilon \sim N(0,8)$
- Generate such a variable:

```
df['iq']=df['ability'] +np.random.normal(0,8,n)
```
- Now add three regressions, of sales on
 - training
 - training and ability
 - training and iq
- Compare the coefficient of ability from the second regression with that of iq in the third and interpret the results
- Compare the coefficient of *training* in the three regressions and interpret the results

- Continue with the do file you created on the slide before and save it under a different name (SalesSim4)
- Now replace the line in which you generated the sales variable
$$\text{df['sales']} = 10000 + \text{df['ability']} * 100 + \text{np.random.normal}(0, 4000, n)$$
- And now regress *sales* on *training* and *iq*
- What do you find?
- Interpret your observation

6. Fixed Effects, Difference-in-Difference, and Panel Data

- 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
- Two most important approaches
 - *Fixed Effects* (when we have panel data, that is we observe the same objects repeatedly)
 - *Difference-in-Difference* estimation (when we have repeated cross-sections and the treatment varies at an aggregate level)

6.1 Fixed Effects

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

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

- 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 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: In 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?

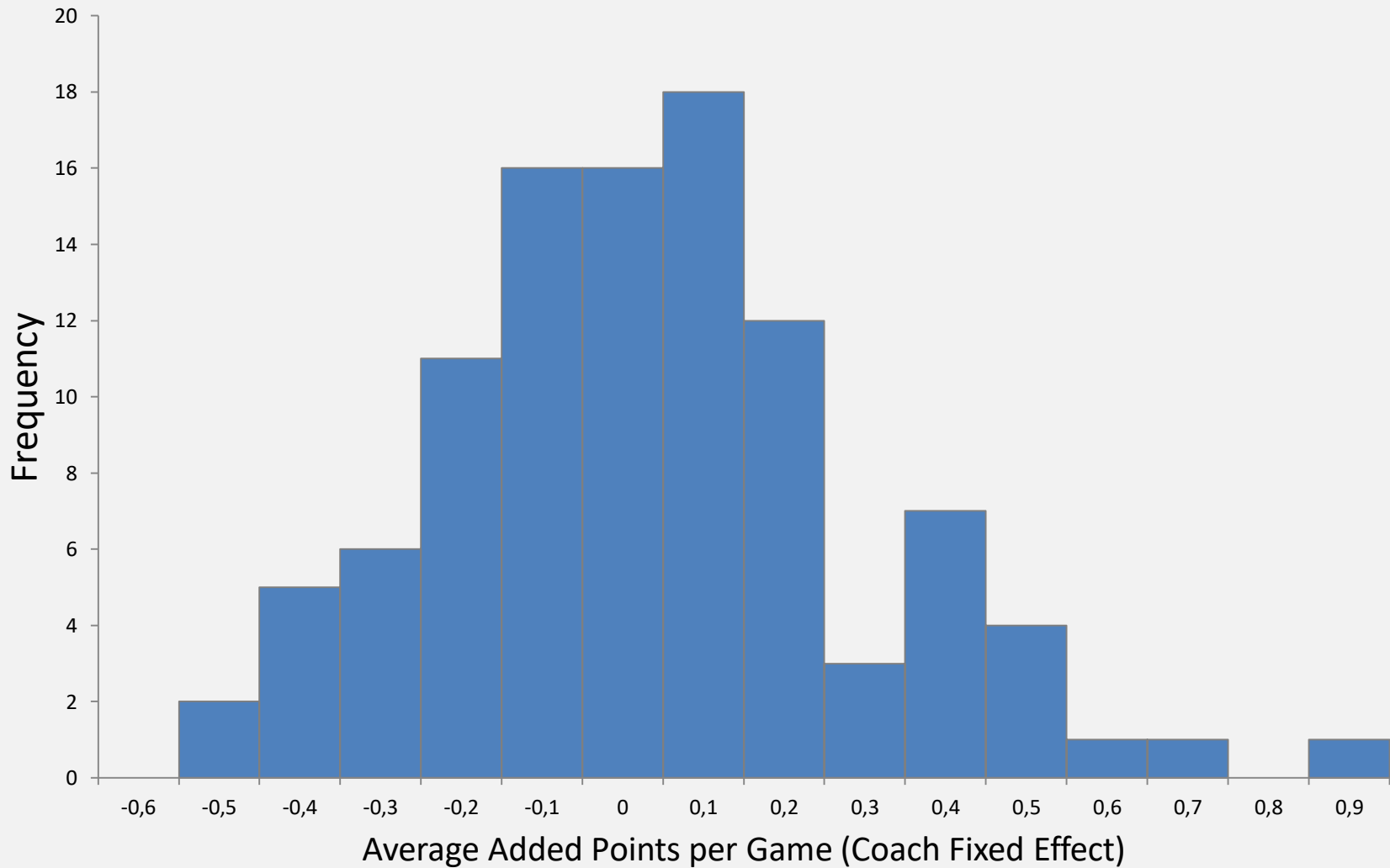
- The worker fixed effects model (row 2) takes this problem into account
- It imposes the weaker assumption that

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

- 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 unobserved 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

- 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 unexplained variation in firm practices can be attributed to manager fixed effects (controlling for firm fixed effects and time-varying firm characteristics)
- Mühlheusser/Schneemann/Sliwka/Wallmeier (2017) consider the case of 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




Estimated Soccer Coach Fixed Effects for the Bundesliga



Study

Mystery shopping in a bakery chain (Friebel et al, mimeo)

- Each month, each shop of a bakery chain is visited by a mystery shopper.
- The mystery shopper buys some bread and evaluates the performance of the employees in a bakery (“mystery shopping score”) using the following sheet:

Bakery:	Albertus-bakery, Magnus street		
Date:	Friday, Dec 05th, 2014		
Name the mystery shopper	Max Mustermann		
<hr/>			
1	Cleanness		9/9
2	Shopping experience		20/24
3	Freshness		11/12
Overall score			40/45
<hr/>			
1 Cleanness			
1.1.1	The shop is clean.	Points:	1/1
...
<hr/>			
2 Shopping experience			
2.1.1	The employee is friendly.	Points:	3/5
2.2.1	The employees says "Thank you".	Points:	1/1
...
<hr/>			
3 Freshness			
3.1.1	The products look fresh	Points:	1/1
...

- Goal of the mystery shopping: Measure the effort of employees. Shop-managers receive a bonus based on the mystery shopping scores
- But Friebel et al find:
 - Regressing shop performance on mystery shopping scores shows that they have no predictive power
 - But the R^2 increases substantially (from 2% to 23%) when including mystery-shopper-fixed effects in the regressions.

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 $\bar{Y}_i = \frac{1}{T} \sum_{t=1}^T Y_{it}$

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

and subtract this from Y_{it}

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

→ The α_i are eliminated!

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

- 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-meanned“ outcome on the „de-meanned“ explanatory variables
 - This gives us an estimate of ρ
 - 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

- 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 `MultilIndex 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 column 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

Sector	(1)	(2)	(3)
	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 three important caveats:

1. When the treatment C_{it} hardly varies over time it is hard to evaluate the causal effect ρ
 - In the extreme when C_{it} is completely stable then $C_{it} = \bar{C}_{it}$
 - Not identifying a significant effect in the data then does not necessarily imply that there is no such effect
2. When on top of that there is measurement error in C_{it} this can generate most of the variation in C_{it} over time
 - this will lead to attenuation bias
 - thus there can be a tendency to underestimate the true 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

6.2 Difference-in-Difference Estimation

- Sometimes the regressor of interest varies only at a more aggregate level (say a state, or a firm) which we index by s
- Moreover, sometimes we do not observe the same individuals repeatedly but have different samples at different points in time t
- We can then use a so-called difference-in-difference estimation strategy
- This is like a fixed effects model with a fixed effect at a more aggregate level
- The underlying idea is again that there is an additive structure in the potential outcomes that is

$$E[Y_{0ist}|s, t] = \gamma_s + \lambda_t$$

- Assume that the causal effect of the treatment is a constant

$$E[Y_{1ist} - Y_{0ist}|s, t] = \delta$$

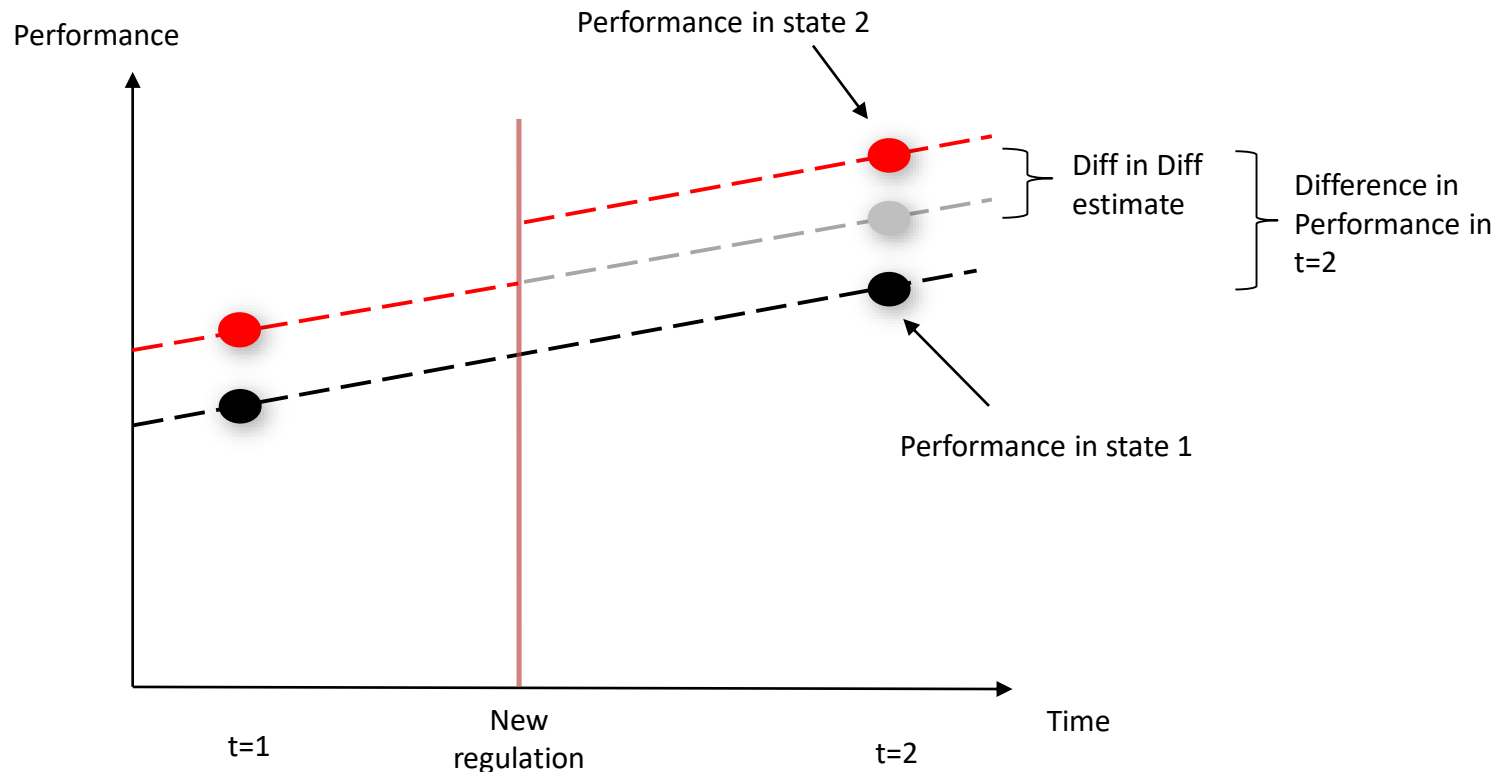
- Then

$$Y_{ist} = \gamma_s + \lambda_t + \delta \cdot C_{st} + \epsilon_{ist}$$

where $E[\epsilon_{ist}|s, t] = 0$

Diff-in-Diff: Graphical illustration

- Suppose that there are two states (regions, firms, departments,...)
- A new regulation is introduced in state 2 but not in state 1
- We want to study the effect on firm performance



Consider

$$Y_{ist} = \gamma_s + \lambda_t + \delta \cdot C_{st} + \epsilon_{ist}$$

- Say we have two periods $t = 1, 2$ and two states $s = 1, 2$
- A new policy is adopted in state 2 in period 2 such that
 - $C_{11} = C_{21} = C_{21} = 0$ and $C_{22} = 1$
- Then
 - $E[Y_{ist}|s = 1, t = 2] - E[Y_{ist}|s = 1, t = 1] = \lambda_2 - \lambda_1$ and
 - $E[Y_{ist}|s = 2, t = 2] - E[Y_{ist}|s = 2, t = 1] = \lambda_2 - \lambda_1 + \delta$
- The difference-in-difference is

$$\begin{aligned} & (E[Y_{ist}|s = 2, t = 2] - E[Y_{ist}|s = 2, t = 1]) \\ & - (E[Y_{ist}|s = 1, t = 2] - E[Y_{ist}|s = 1, t = 1]) = \delta \end{aligned}$$

which is the causal effect of interest

Regression Diff-in-Diff

- Note: we could estimate the causal effect δ 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 region
 - Let $POST_t$ be a dummy indicating whether an observation comes from a period after the treatment has been implemented

- Then we can regress

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

- The coefficient δ 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

- 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 9 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

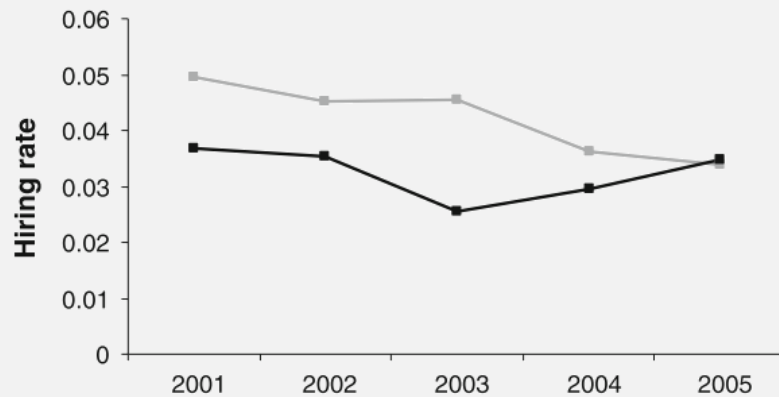
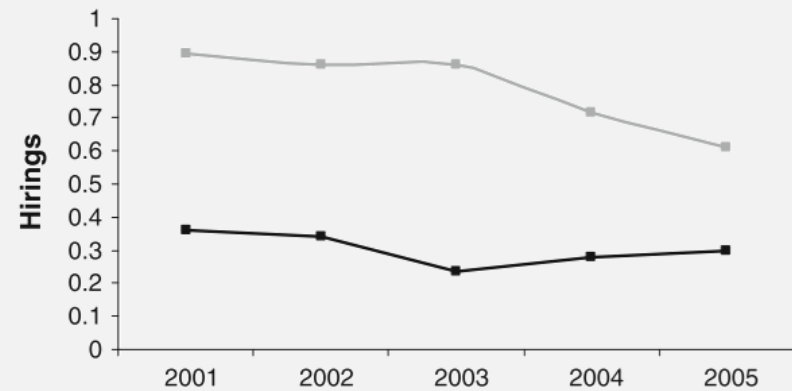


Fig. 2 Dynamics of hirings in treatment and control group. The *left figure* shows average hiring rates for treatment (*black line*) and control groups (*grey line*) over time. The *right figure* shows average absolute hirings for treatment (*black line*) and control groups (*grey line*) over time. The 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	
<i>N</i>	1,749		1,658		1,285	
<i>R</i> 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)

The Common Trend Assumption

- Crucial for both a Diff-in-Diff and a Fixed Effects estimation strategy: Treatment and control group follow the same underlying time trend!
- If this is violated, then both approaches yield biased estimates
- It is called the *common trend assumption*
- Again this is an identifying assumption: We can claim that we identify a causal effect when the common trend assumption holds
- It can be very useful to check the trends in several periods
 - before the change occurs
 - after the change occurs
- If trends differ already before the intervention both strategies are problematic to identify causal effects