4. Regression and Causality

- 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 subjects 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!

4.1 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

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

- The *causal effect* of C_i on Y_i is now $Y_{1i} - Y_{i0}$

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

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

- 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:

$$\underbrace{E[Y_i|C_i = 1] - E[Y_i|C_i = 0]}_{\text{Observed difference in outcome}}$$

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

- 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)*1

  (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 ⫫ indicate independence
- If the condition

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

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

4.2 Why are Experiments so Important?

- Suppose we have a *Randomized Controlled Trial* (RCT, A/B Test)
  - That is C_i is randomly (that is *exogenously*) 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

Implementing RCTs in Firms: Typical Project Timeline

1. Preparation (1-3 months)

- Analysis:
  - Detailed analysis of current design of HR practice
  - Collection of outcome data (i.e. KPI, performance evaluations, survey data)
  - Prior qualitative analysis of HR practice
  - Analysis of quantitative data
  - Statistical power analysis
- Design
  - Development for redesign of HR practice
  - Treatment design for A/B test
  - Survey design
  - Communication strategy

2. A/B Testing (3-12 months)

- Duration of A/B test and number of treatments fixed
- A/B Test implemented
  - Random assignment of units to treatment (stratified randomization)
  - Communication strategy implemented
- Posterior employee survey

3. Evaluation (1-3 months)

- Data collection
  - Outcome data for treatment and control units before and during the treatment time collected
- Analysis
  Estimate causal effect on
  - Performance outcomes
  - Employee attidudes (survey outcomes)
- Presentation & Choice
  - Present & discuss results
  - Proposal for roll-out

RCT in Retailing: Talking about Performance

Manthei/Sliwka/Vogelsang (Management Science, 2022): Four treatments

	Bonus	No Bonus
Review	N=63	N=51
No Review	N=50	N=60

- April 2017 – June 2017 (3 Month)
- Performance Incentive:
  €0.05 for ever €1 *profit* above 80% of the planned value
- Monitoring/Performance Review:
  Biweekly reviews meetings with district managers

Protocol:
- What did the store manager do?
- Which problems did occur?
- What does he/she plan to do next?

RCT in Retailing: Talking about Performance

Table 2. Main Treatment Effects on Profits

Model	(1) Profits	(2) Profits	(3) CI 90%	(4) Log (profits)	(5) Log (profits)	(6) CI 90%
Treatment effect *Bonus*	−51.85	156.2	[−1,049.6;	−0.00441	0.0141	[−0.0825;
	(607.3)	(710.5)	1,362.7]	(0.0417)	(0.0569)	0.1108]
Treatment effect *Review*	1,370.2**	1,492.3**	[361.6;	0.0732***	0.0858**	[0.0161;
	(559.0)	(666.2)	2,622.9]	(0.0238)	(0.0411)	0.1554]
Treatment effect *Bonus&Review*	−376.3	−397.7	[−1,355.5;	−0.00485	−0.00390	[−0.0889;
	(605.1)	(564.3)	560.0]	(0.0351)	(0.0501)	0.0811]
Wald test *Review* = *Bonus&Review*	$p = 0.0162$	$p = 0.0090$		$p = 0.0218$	$p = 0.0330$	
Time fixed effects	Yes	Yes		Yes	Yes	
Store fixed effects	Yes	Yes		Yes	Yes	
District manager fixed effects	No	Yes		No	Yes	
Store manager fixed effects	No	Yes		No	Yes	
Refurbishments	Yes	Yes		Yes	Yes	
Planned profits	Yes	Yes		Yes	Yes	
No. of observations	3,975	3,777		3,966	3,768	
No. of stores	224	224		224	224	
Cluster	31	31		31	31	
Within R^2	0.2370	0.2722		0.1621	0.1875	
Overall R^2	0.7577	0.5955		0.6158	0.4316	

Notes. The table reports results from a fixed effects regression with the profits on the store level as the dependent variable. The regressions compare pretreatment observations (January 2016–March 2017) with the observations during the experiment (April 2017–June 2017). Treatment effect thus refers to the difference-in-difference estimator. All regressions control for possible refurbishments of a store and the companies' planned value of profits. Observations are excluded when a store manager switched the store during the treatment period. Robust standard errors are clustered on the district level of the treatment start and displayed in parentheses. Columns (3) and (6) display 90% confidence intervals (CIs) of the specification in columns (2) and (5), respectively.

**$p < 0.05$; ***$p < 0.01$.

RCT in Retailing: Talking about Performance

Table 5. Survey Results Perceptions on Activities

Model	(1) Satisfaction job	(2) Satisfaction compensation	(3) Satisfaction workload	(4) Profit aim	(5) Feedback	(6) Motivate
Treatment effect *Bonus*	−0.282	0.310	−0.301	−0.111	0.156	0.328
	(0.322)	(0.273)	(0.522)	(0.251)	(0.274)	(0.339)
Treatment effect *Review*	0.169	−0.0767	−0.383	0.0498	0.959***	0.896*
	(0.267)	(0.324)	(0.546)	(0.265)	(0.307)	(0.458)
Treatment effect *Bonus&Review*	−0.114	−0.0263	−0.605	0.509**	0.409	0.0463
	(0.308)	(0.242)	(0.474)	(0.242)	(0.243)	(0.338)
Wald test *Review* = *Bonus&Review*	$p = 0.3438$	$p = 0.8680$	$p = 0.5934$	$p = 0.1016$	$p = 0.0432$	$p = 0.0412$
No. of observations	95	95	95	93	94	94
Cluster	28	28	28	28	28	28
Overall R^2	0.141	0.280	0.108	0.181	0.199	0.187

Notes. The table reports results from OLS regressions with the respective survey response as the dependent variable (scale from 1–6). In addition to general job satisfaction, satisfaction with the compensation, and satisfaction with the workload, controls are store size, number of employees, store manager's age, and prior performance evaluation, as well as randomization group. Standard errors are clustered on the district level of the treatment start and displayed in parentheses.

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

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.binomial(1, 0.05, n)
```

- Note: np.random. binomial(1,0.05,n) generates a vector of n binomial random variables with 1 trial each (taking value 1 with 5% probability)
- 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

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 $f_i(s) \equiv f(s, X_i)$ where X_i is a vector of i's characteristics
- 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 (exogenously) changing s
- Regressions powerful to estimate the causal model when the CIA holds

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 such an identifying assumption

Omitted Variable Bias

- Assume that the causal relationship between Y_i and C_i is determined by

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

  where v_i is uncorrelated with all regressors

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

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

- The short regression yields (use the true causal relationship)

$$\widetilde{\rho} = \frac{Cov[C_i, Y_i]}{V[C_i]} = \frac{Cov[C_i, \alpha + \rho \cdot C_i + \gamma \cdot X_i + v_i]}{V[C_i]}$$

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

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

- If $Cov[C_i, X_i] \neq 0$ 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: If C_i is *endogenously* determined by X_i and we cannot observe X_i

- then the regression will yield a biased estimate of the causal effect
- the size of this *omitted variable bias* is $\gamma \cdot \delta_c$

Your Task

OVB (Simulated Sales Training Evaluation III)

- 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 the 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
- Reducing standard errors and thus increasing 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:

```
df['sales']= 10000 + df['training']*5000+df['ability']*100
                   + df['experience']*15000+ 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 it is 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))
- Moreover, to order the explanatory variables add the option `regressor_order=['training','ability','experience']`

Your Task

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

- 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 2008, 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

Your Task — Bad control (Simulated Sales Training Evaluation V)

- Open again the notebook SalesSim1 (the first version)
- Generate a variable testScore that is the result of a test the sales agent took part in after the training

```
df['testScore'] = 100 + df['ability'] + df['training']*50
                      + 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 testScore
- Compare the coefficient of training in the two regression outputs

4.4 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 $\widetilde{X}_i = X_i + \eta_i$ where $\eta_i \sim N(0, \sigma_\eta^2)$
- If we run a regression

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

we obtain a coefficient

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

But

$$\frac{Cov[X_i + \eta_i, \alpha + \gamma \cdot X_i + v_i]}{V[X_i + \eta_i]}$$

$$= \gamma \cdot \frac{\sigma_x^2}{\sigma_x^2 + \sigma_\eta^2}$$

- 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

Your Task — Measurement Error (Simulated Sales Training Evaluation VI)

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

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

Measurement Error in a Covariate

- 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
- But we measure only a "proxy" $\widetilde{X}_i = X_i + \eta_i$ for the omitted variable X_i

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
- we thus reduce omitted variable bias but we do not eliminate it

Your Task — Measurement Error (Simulated Sales Training Evaluation VII)

- 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

  df['sales']= 10000 + df['ability']*100 + np.random.normal(0,4000,n)

- And now regress $sales$ on $training$ and iq
- What do you find?
- Interpret your observation