

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 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!

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

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

- 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

4.2 Why are Experiments so Important?

- Suppose we have a randomized controlled experiment
 - 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

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 $f_i(s) \equiv f(s, X_i)$
- 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

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 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

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

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

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

- 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: 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 $\gamma \cdot \delta_c$

- Consider association between wages and education in the NLSY97
- We find that 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)

- 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