Your Task — Evaluation of a Sales Training

- Assume that an HR analytics unit has collected information on training participation and sales for the firm's sales agents from the ERP system
- Start a new notebook
- Read the data set "sales_data.csv" which contains a simulated data set of sales agents
- Analyze the association between training and sales

- To do so:
  - Plot a bar chart with sales for trained and untrained workers
  - Run a regression of sales on training
  - Interpret the results (& save the notebook as SalesSim)

3. 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 data analysis becomes very powerful as you can actually say what works!

3.1 The Potential Outcome Framework

Suppose our aim is to investigate whether

- a certain management practice (bonus, wage increase, training,...)
- causally affects some variable Y_i (job satisfaction, performance,...)

Consider the important framework also know as the Rubin causal model:

- 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 $Y_{1i} - Y_{0i}$

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* directly

- This is sometimes called the *Fundamental Problem of Causal Inference*
- 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 cell that generates a fictitious data set with 500 observations

```
n=500
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']= 40000 + df.training*5000 + df.ability*100
                + np.random.normal(0,4000,n)
```

- Question: What is the true causal effect of the training?

Your Task

Simulated data set: Evaluation of a sales training

- What is the coefficient of training in a regression of sales on training?
- Interpret the regression results
- Increase the sample size to 50000
- Compare the regression results

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

„No Causation without Manipulation“

- This moto was stated in a famous article by Holland (1986)
- Key idea: X can only have a causal effect on Y if X can be changed
- That is:
  - We can make causal claims only on things that can be actively changed
  - For variables that cannot be changed try to avoid ***causal language***
  - In this case: Rather speak of an ***association*** than of an ***effect***
- We must be careful about claiming causality when studying things like
  - the association between gender and wages
  - the association between personality traits and performance,…
- But we may, for instance, be able to say something on the causal effect
  - of a management practice on profits or
  - of wage on job satisfaction, as both can be manipulated
- This is exactly the reason why experiments are so important to identify causal effects: when we can manipulate something we can run an experiment!

3.2 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

A/B Testing and RCTs in Firms

- In the past, people have often claimed that you cannot experiment with employees
- This view has changed
- Firms have for a long time experimented with customers, varying for instance
  - marketing campaigns
  - presentation of goods in stores…
- Has been done much more systematically on online platform in A/B tests to
  - investigate the effects of platform changes on user engagement
  - learn from this on optimal roll out for features to all users
- This has increased the awareness of practitioners of the importance and power of experiments also in other domains

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
  - Often: Negotiation with works councils

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, 2024): 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 SalesSim notebook and copy the cell with the simulation to a new cell at the bottom of the notebook
- Now suppose that there is new training program which the firm implements in an A/B test with a randomly selected 5% of the workforce
- 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

3.3 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?
- Also called the *"selection on observables"* assumption: selection into the treatment only depends on observables 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
- But: It is a strong assumption!

Your Task

Control Variables (Simulated Sales Training Evaluation IIIa)

- Open the SalesSim notebook & again copy the cell where you created the first simulation (in Exercise I)
- Suppose now that ability is in the data set
- Add a regression of sales on training and ability
- Interpret your results

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
  - But we should be aware that it is still a strong 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 (OVB)* is $\gamma \cdot \delta_c$
- Note: The OVB corresponds to the selection bias in the potential outcome framework

Your Task

OVB (Simulated Sales Training Evaluation IIIb)

- Open the SalesSim notebook and in a new cell 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 when you want to estimate causal effects?

1. **Reduce 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

2. **Reduce standard errors**
   - 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 and improve the statistical power

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

- Open SalesSim and copy again the cell where you generated the first simulation in a new cell at the bottom of the notebook & adapt the cell
- 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
- Run the notebook several times, comparing the four regression results (inspect the training coefficient and its standard error)

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)

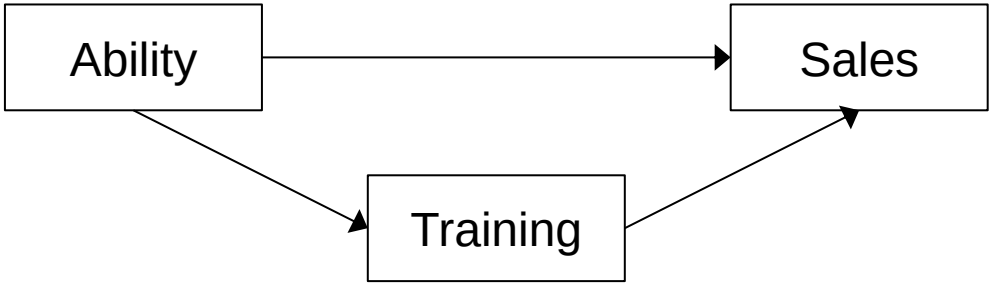
- Open again the notebook SalesSim and copy again the cell where you created the first simulation in a new cell at the bottom of the notebook
- Generate a variable testScore that is the result of a test the sales agent took part in after the training

```
df['testScore'] = 100 + df['ability'] + df['training']*50
                      + np.random.normal(0,5,n)
```

- At the end of the notebook add code to
  - **regress** sales **on** training
  - **regress** sales **on** training **and** testScore
- Compare the coefficient of training in the two regression outputs

3.4 Notes on Mediation Analysis

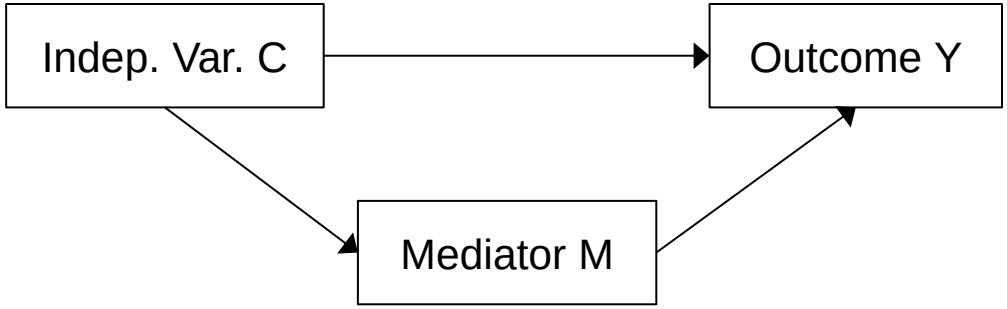
- Sometimes we are interested in the mechanism through which a variable influences the outcome
- In our sales example: Suppose we are not mainly interested in the causal impact of training on sales but in the impact of ability on sales and want to distinguish between the
  - direct impact of ability (independent var.) on sales (dependent var.)
  - and the indirect impact through training (mediator)



- Here you may consider to add an “endogenous” variable on the right-hand side ➔ then we perform a mediation analysis

Mediation Analyses are frequently used in Psychology and Management

Often based on an approach proposed by Baron and Kenny (Journal of Personality and Social Psychology, 1986).



Three steps:

1. *Regress the outcome on the independent variable* (Y ~ C)
   Purpose: Study whether independent variable is a predictor of outcome

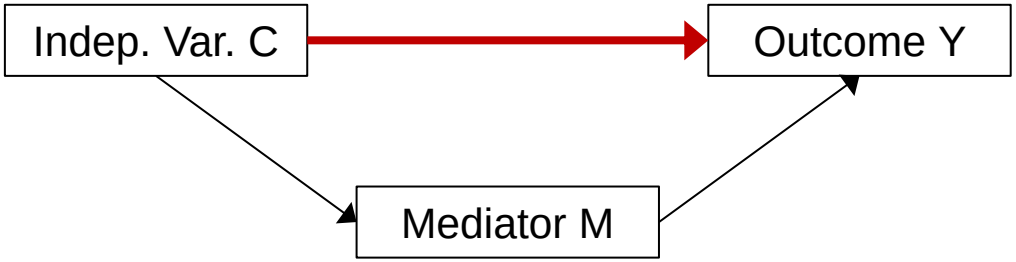
2. *Regress mediator on the independent variable* (M ~ C)
   Purpose: Study whether indep. var is also a predictor of the mediator

3. *Regress outcome on both mediator and independent variable* (Y ~ C + M)
   Then investigate whether
   - (i) mediator is still a significant predictor and
   - (ii) coefficient of the independent variable is smaller than in step 1

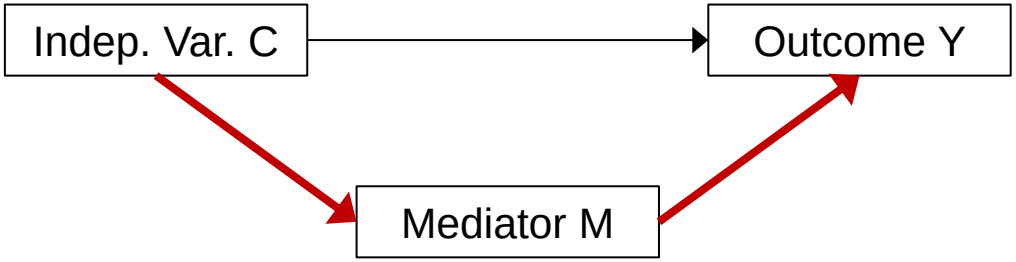
This literature then proposes the following interpretation:

- The **"total effect"**: from Y ~ C

- The **"direct effect"**: from Y ~ C + M



- The **"indirect effect"**: from M ~ C and Y ~ C + M



To obtain the indirect effect multiply the coefficients of C from M ~ C with that of M in Y ~ C + M

Your Task — Mediation Analysis (Simulated Sales Training Evaluation VI)

- Your task is now to study the association between sales and ability and to perform a mediation analysis considering training as a mediator
- What is your estimate of the
  - Total “effect”
  - Direct “effect”
  - Indirect “effect”
- To compute the indirect effect you can either copy the coefficient estimates you need from the table or access the coefficients by using for instance reg.coef().ability

Note: Procedure comes with very strong assumptions!

- Bullock et al. (Journal of Personality and Social Psychology 2010): "*inference about mediators is far more difficult than previous research suggests*"
- Imai et al. (Am. Pol. Science Review 2011): "…*methods rely upon untestable assumptions and are often inappropriate even under those assumptions*."
- Heckman/Pinto (Econometric Reviews, 2014): "*A fundamental problem of mediation analysis is that even though we might observe experimental variation in some inputs and outputs, the relationship between inputs and outputs might be confounded by unobserved variables.*"

Key problem: Even if C is as good as randomly assigned,…

- there will very likely be other unobserved variables that affect both M and Y
- hence, we then have OVB in the regression of Y on C and M
- Or, including M in the regression Y ~ C + M "*messes up the magic of random assignment*" (see https://datacolada.org/103)

Recommendation: If you perform a mediation analysis add a disclaimer that your results do not have a clean causal interpretation!

Your Task — Mediation Analysis (Simulated Sales Training Evaluation VI)

- Copy the cell where you created the simulation
- Add another variable ("motivation") to your simulation: adapt the code

```
df['motivation'] = np.random.normal(0,5,n)
df['training']=(df.ability +df.motivation +np.random.normal(0,10,n)>=100)*1
df['sales']=40000+df.training*5000 + df.motivation*2000+df.ability*100
        +np.random.normal(0,4000,n)
df.drop(columns=['motivation'])
```

- Now perform exactly the same mediation analysis as before (just copy the code and run it)

3.5 Measurement Error

- The previous considerations suggest that multiple regressions come close to causal effects when there are proper control variables
- But what if we can only 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
- 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
- Note: There are methods to tackle measurement error through using *instrumental variables* (not covered here)

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

- Open again the notebook SalesSim and copy again the cell where you created the first simulation in a new cell at the bottom of the notebook
- Suppose now that the firm cannot measure ability but has access to a work sample test that the sales agents had to take when being hired
- This $hiringTest$ outcome is equal to $ability + \varepsilon$ where $\varepsilon \sim N(0,8)$
- Generate such a variable:

  df['hiringTest']=df['ability'] +np.random.normal(0,8,n)
- Now suppose the firm is interested in studying the effect of ability on sales
- As ability cannot be measured directly it uses the hiring test instead
- Run a regression of sales on hiringTest and training and interpret the coefficent of hiringTest → what does it tell you about the effects of ability on sales?

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

- Now suppose the HR department is again interested in evaluating the causal effect of the training
- Being aware of potential selection bias issues the analyst wants to include the hiringTest as a proxy for ability
- Run three regressions, of sales on
  - training
  - training and hiringTest
  - training and ability (Note: the firm cannot do this regression)
- Compare the coefficients of training in the three regressions and interpret the results
- Increase the sample size to n=100000 and repeat the exercise

Measurement Error in a Covariate

- Measurement error can thus be particularly 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 VIII)

- Open again the notebook SalesSim and copy again the cell where you created the first simulation in a new cell at the bottom of the notebook
- 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 hiringTest
- What do you find?
- Interpret your observation

Conclusion

- Data science projects yield most useful insights when identifying causal effects
  - Then directly give actionable insights & leave little room for debate
  - For instance: If the causal effect of a management practice on profits exceeds the costs & the practice also does not reduce employee satisfaction → implement the practice
- But to identify causal effects it does not suffice to have "big data"
- Here it is important to understand the logic of counterfactuals & the potential outcome framework
  - We can decompose the association between an intervention and an outcome variable into the causal effect & the selection bias
  - It is easy to estimate causal effects with random assignment as this eliminates the selection bias (in large enough samples)
  - Without random assignment, important to reduce omitted variable bias
  - Control variables help but do not eliminate the problem entirely

Outlook

There are several other methods to eliminate or at least reduce the selection bias when you do not have *experimental* but merely *observational data*

Matching

- Basic idea: For each treated subject find an untreated subject as similar as possible in the determinants of the potential outcome
- But note: Can only match with observable data

Regression Discontinuity Designs (RDD)

- Basic idea: Sometimes subjects are assigned to a “treatment“ based on a threshold rule
- For instance: All employees are assigned to a training when their outcome in some test is below a threshold value
- When considering only subjects close to the threshold, assignment is often “as good as random”
- Note: Needs large samples to have enough subjects close to the threshold

Instrumental Variables (IV)

- Sometimes when the treatment is not randomly assigned, there is a variable that is as good as random that affects the treatment
- Then use only the variation in the treatment caused by this "instrumental variable" in a clever way to identify the causal effect
- Example:
  - Training selection is voluntary, but firm has informed only a random subset of the employees about the training program
  - When more people in this group register for the training, this can be used to estimate the causal effect

Panel Data

- Often we have data on X and Y over multiple time points
- Then instead of studying the association between X and Y we can for instance study the association between *changes* in Y and *changes* in X
- This can bring us closer to the causal effect
- We will cover this in the next chapter