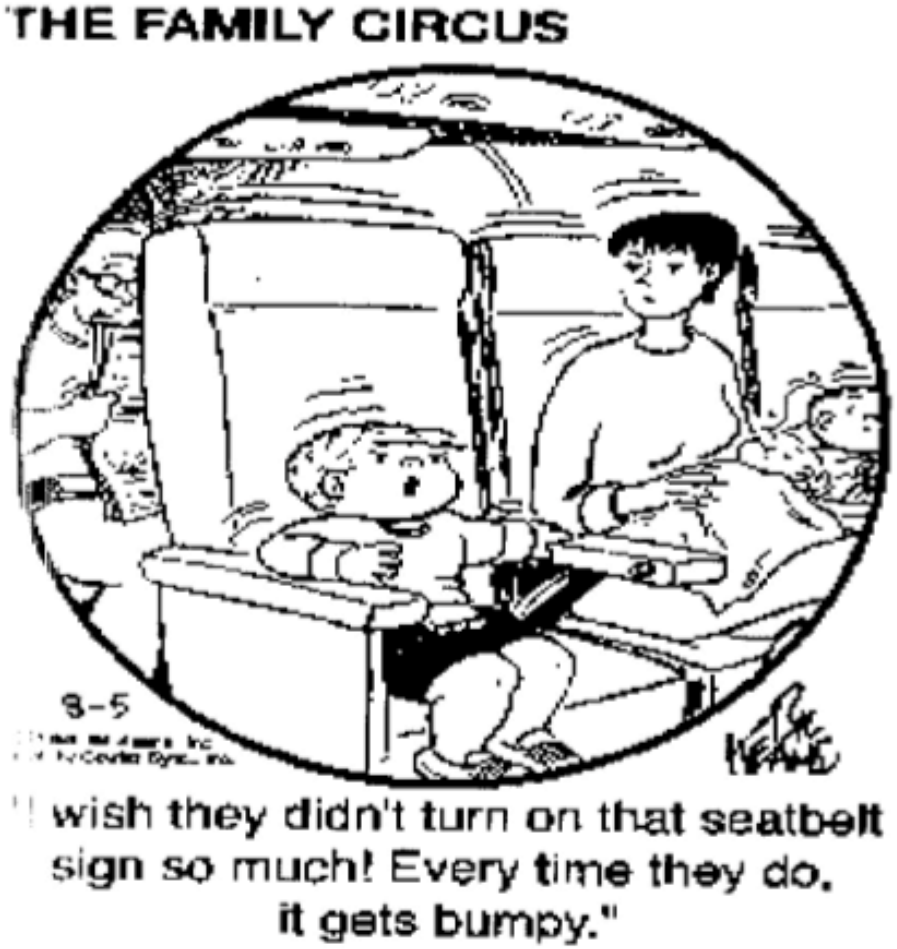
MSBA7003 Decision Analytics

08 Causal Inference I

Agenda

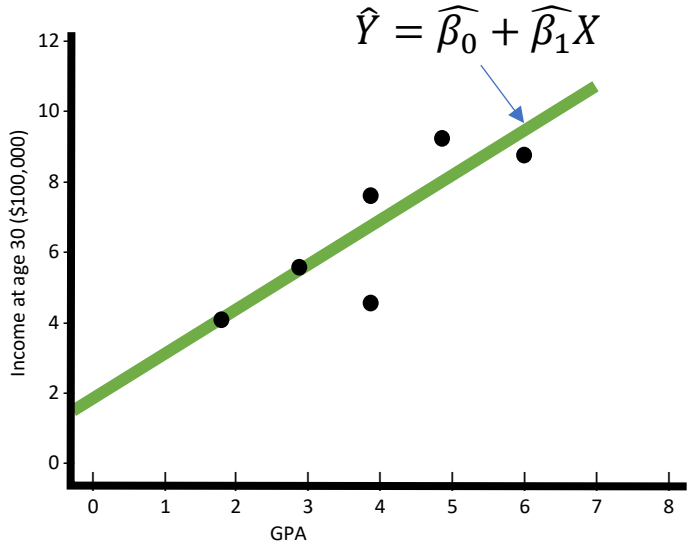
- Decision and Causal Inference
- Counterfactual Model
  - The Naïve estimator
- Causal Identification by Conditioning
  - Matching
- Case: Kjell and Company



Decision and Causal Inference: A Brain Exercise

- The relationship between GPA and income for 6 HKU graduates

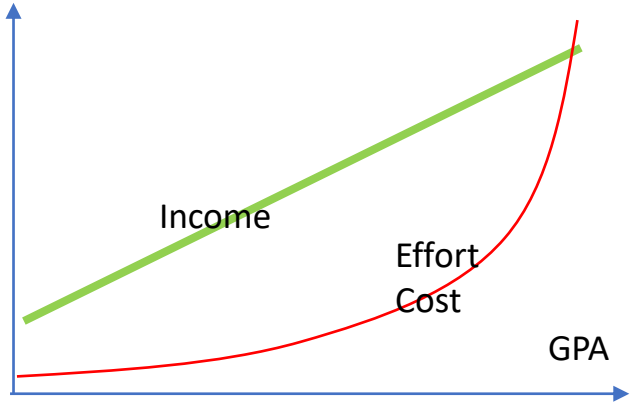
Y: Income at age 30 ($100,000)	X : GPA (1~6 scale)
6	3
8	4
9	6
5	4
4.5	2
9.5	5

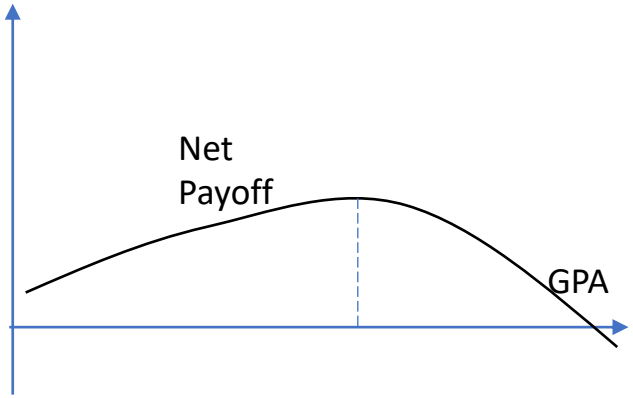


- $\hat{\beta} = \begin{bmatrix} \widehat{\beta_0} \\ \widehat{\beta_1} \end{bmatrix} = \begin{bmatrix} 2 \\ 1.25 \end{bmatrix}$

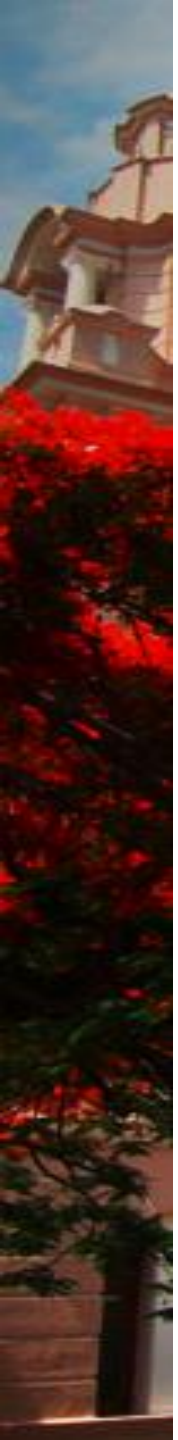
Decision and Causal Inference: A Brain Exercise

- Now you are going to decide how hard to work in order to improve your GPA.
- According to the HKU data, you estimated that *expected income = 2 + 1.25GPA*.
- You also estimated that the expected effort cost to achieve the target GPA is *0.12*GPA^2*, as illustrated by the cost curve.





- Hence, your optimal GPA is the one that peaks the net payoff curve.

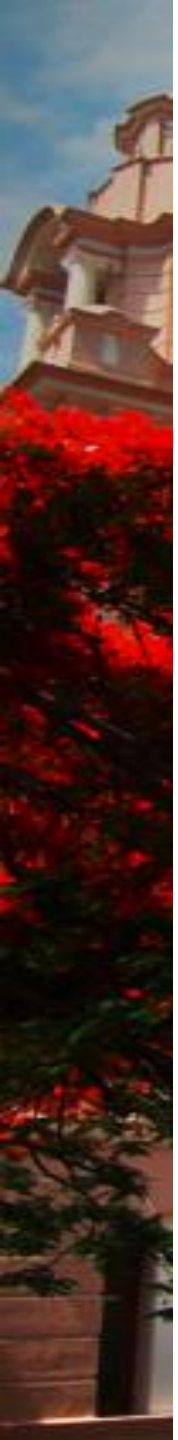


Decision and Causal Inference: A Brain Exercise

- Discussion: What is the real logic behind the relationsip

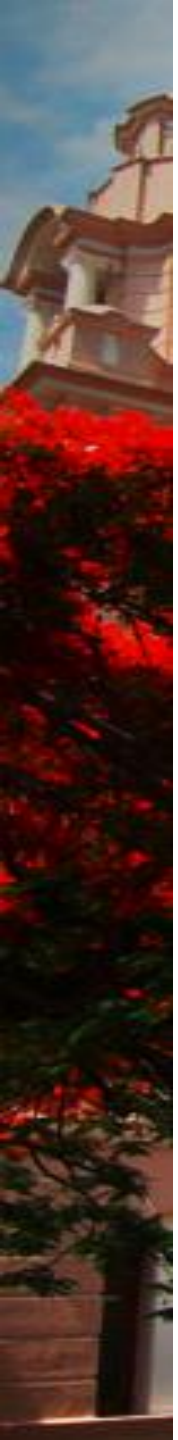
$$\widehat{\text{Income}} = 2 + 1.25 \times \text{GPA}?$$

- Income is associated with GPA in many ways:
  - …
- Hence, if GPA is increased by one unit while other factors are held constant, the changes in income may not be 1.25.
- How much does GPA influence income?



Introduction

- In practice, we often want to evaluate the impact of our decisions.
  - Door-to-door loan collection, special discount offers, etc.
- If we look at the historical data and simply check the correlations, we may get wrong impressions. (Assume randomized experiments are not allowed.)
- How to correctly estimate the causal impact from historical data, so that firms can better allocate their valuable resources?
- In many cases, our decision (the cause) is binary:
  - Factory workers: Enrolment in a training program
  - Job market candidates: Master degree in BA
- Our analysis will focus on binary causes.
- The analysis is analogous for a many-valued cause.



The Counterfactual Model

- For a binary cause (D), two potential states exist for each member of the population: treatment state (1) and control state (0).
- For ***each individual*** (the same person), there are two potential outcome random variables: Y^1 and Y^0.
  - E.g., academic performance of a student after and without joining a training program
- For the same person, everything else is identical under the two possible states. Any difference between Y^1 and Y^0 should be caused by the treatment.
- The individual-level causal effect is $\delta = Y^1 - Y^0$.

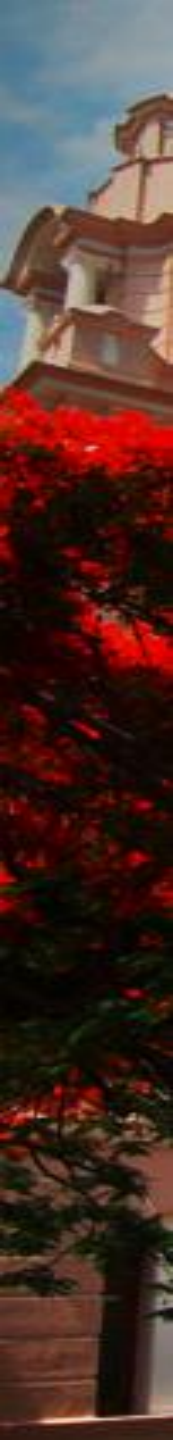
The Counterfactual Model

D = 0

$Y^0 = 85$

D = 1

$Y^1 = 100$



The Counterfactual Model

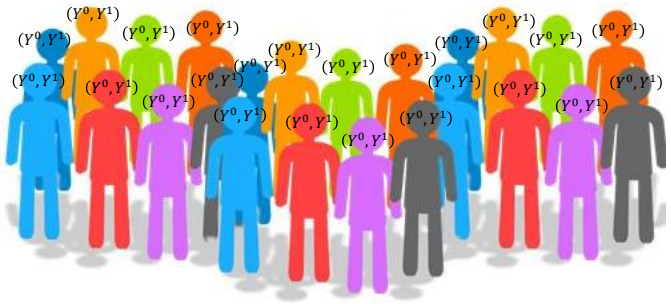
- The fundamental problem of causal inference: Y^1 and Y^0 of the same person exist in parallel universes!

Group	Y^1	Y^0
Treatment group (D = 1)	Observed	Counterfactual
Control group (D = 0)	Counterfactual	Observed

- It is impossible to calculate the individual-level causal effect for each individual i: $\delta_i = Y_i^1 - Y_i^0$. It is impossible to estimate the average causal effect by averaging the individual-level causal effect: $\bar{\delta} = \frac{1}{n}\sum_i \delta_i$.
- How about comparing Y^1 and Y^0 of different people?
- In general, for different individuals, their Y^1 and Y^0 may depend on many different factors. Comparing John's Y^1 and Max's Y^0 is meaningless.

The Counterfactual Model

- However, the answer is yes, if we have the following condition:
- Random assignment and independence condition: $(Y^0, Y^1) \perp D$
  - It means that the treatment state for each individual is random such that it does not depend on any factor that is related to individual performance.
  - It does NOT mean: $Y \perp D$
  - (Y is the observed outcome in the data.)



  - If the treatment does not affect performance at all, Y^1 and Y^0 should follow the same distribution, and thus, the same mean. Any difference in the means should be caused by the treatment.

The Naïve estimator

This equation is due to that subjects in the treatment group are randomly selected so that their average performance in each possible state is representative of the population.

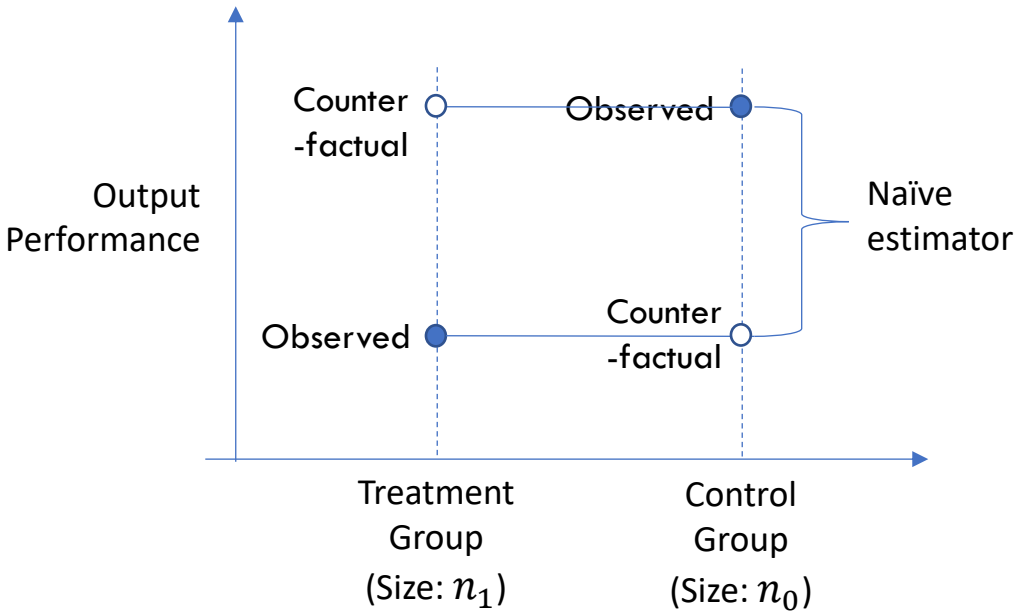
- Given the independence condition, the Naïve estimator of the average treatment effect:

$$\hat{\delta}_{\text{naive}} = \frac{1}{n_1}\sum_i Y_i^1 - \frac{1}{n_0}\sum_i Y_i^0$$

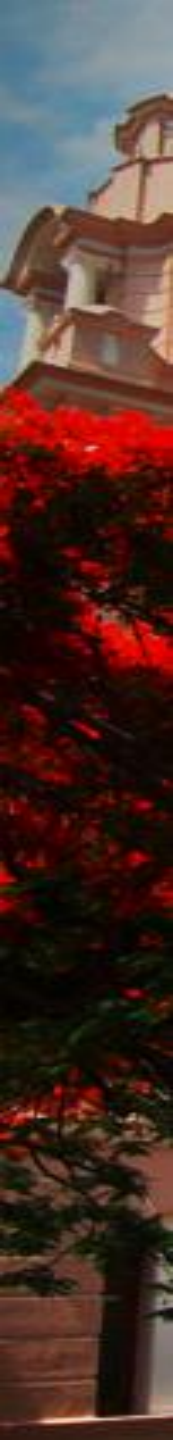
$$\xrightarrow{p} E[Y^1|D=1] - E[Y^0|D=0]$$

$$= E[Y^1 - Y^0] = E[\delta]$$

  - Note that, given the independence condition,
  - $E[Y^1|D=1] = E[Y^1]$ and $E[Y^0|D=0] = E[Y^0]$.



- Without the independence condition, $\hat{\delta}_{\text{naive}}$ is biased.
  - E.g., D depends on factor S, which is also related to the outcome.

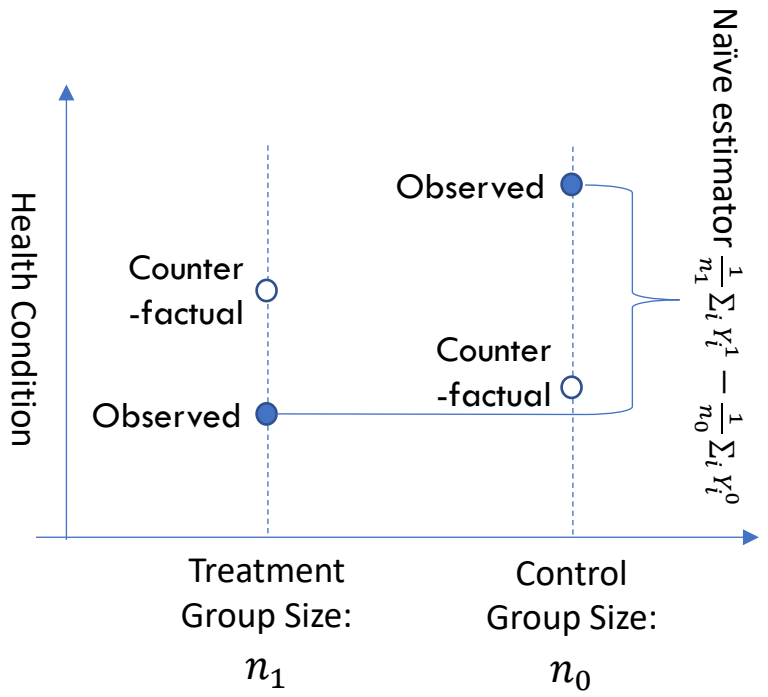


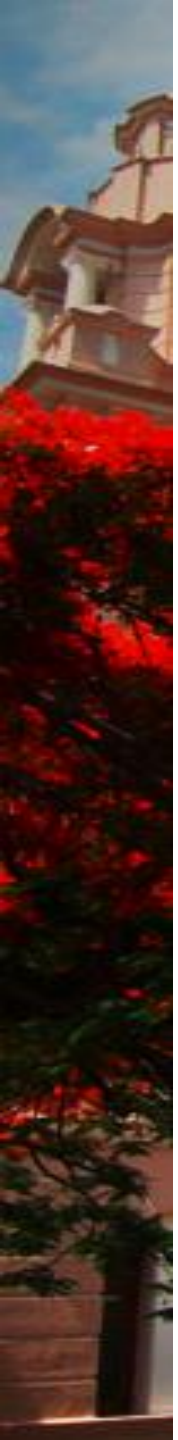
Example 1: Independence Condition Holds

- A group of scientists developed a new medicine for disease X. They tried to evaluate the effect of this medicine by conducting a controlled trial.
- They invited some disease carriers and randomly divided them into two groups of the same size n. One group is the treatment group and the other is the control group. Volunteers in the treatment group will get the medicine, while volunteers in the control group will get no treatment. Their health conditions after 3 days are recorded.
- Let Y_i^0 denote the health condition of volunteer i in the control group
- Let Y_i^1 denote the health condition of volunteer i in the treatment group
- The effect of the vaccine is estimated as $\frac{1}{n}\sum Y_i^1 - \frac{1}{n}\sum Y_i^0$.

Example 2: Independence Condition Is Violated

- A group of students study the effect of smoking on health condition. They invited two groups of people: one has a history of smoking and the other has never smoked. Each group has n participants.
- Let Y_i^0 denote the health condition of participant i in the non-smoking group.
- Let Y_i^1 denote the health condition of participant i in the smoking group.
- The effect of the smoking is estimated as $\frac{1}{n}\sum Y_i^1 - \frac{1}{n}\sum Y_i^0$.
- The estimate is biased.

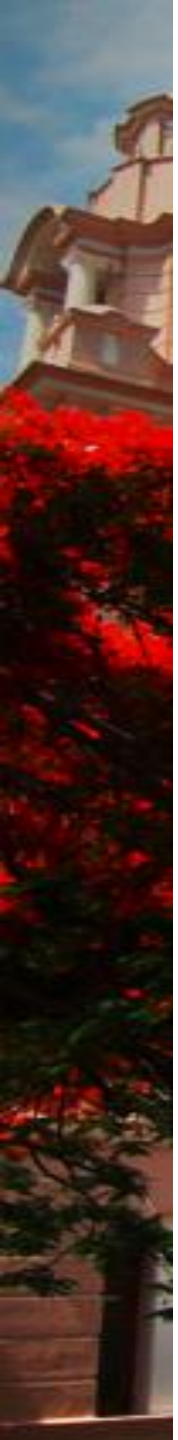




Example 3: Independence Condition Is Violated

- Suppose we studied a sample of job market candidate, and we have (with the god's help) correctly estimated their job market performance of some kind (e.g., monthly income at age 30 in thousand CNY) as follows. The treatment is whether or not an individual has a college degree. We know that 30 percent of the population obtains college degrees.

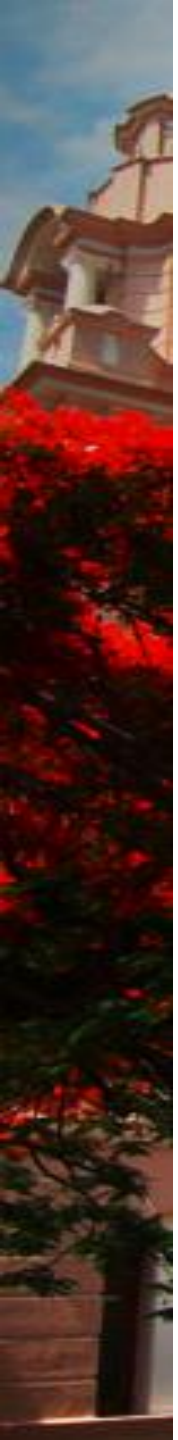
Group	$E[Y^1 \mid D]$	$E[Y^0 \mid D]$
Treatment group (D = 1)	10 (observed)	6 (predicted)
Control group (D = 0)	8 (predicted)	5 (observed)



Example 3: Independence Condition Is Violated

- What is the Naïve estimator?
  - $\hat{\delta}_{\text{naive}} = 10 - 5 = 5$
- What is the average treatment effect for the population?
  - $E[\delta] = 0.3 \times (10 - 6) + 0.7 \times (8 - 5) = 3.3$

- Since $E[\delta] = \pi E[Y^1 - Y^0 | D = 1] + (1 - \pi) E[Y^1 - Y^0 | D = 0]$
- We have $\hat{\delta}_{\text{naive}} = E[Y^1 | D = 1] - E[Y^0 | D = 0] =$
  $E[\delta] + \{E[Y^0 | D = 1] - E[Y^0 | D = 0]\} + (1 - \pi)\{E[\delta | D = 1] - E[\delta | D = 0]\}$
- People with college degree perform better in the job market than those who have not attended college. There are three possible reasons that we have this observation. First, attending college might make people smarter on average (the average treatment effect). Second, individuals who attended college might have been smarter in the first place. Third, individuals who attended college might be better at learning.



The Counterfactual Model

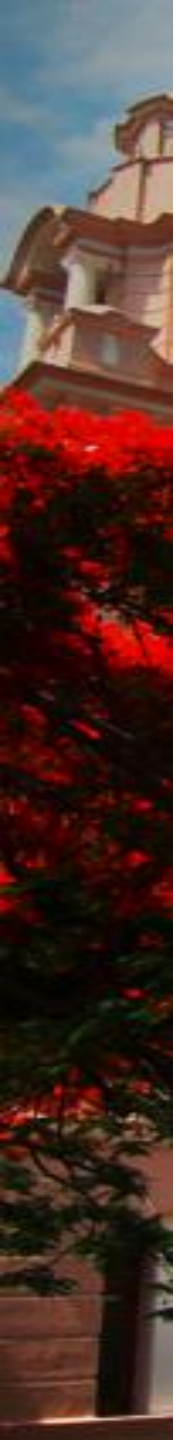
- Conditional Average Treatment Effects
- Two conditional average treatment effects are of particular interest.
- The average treatment effect for those who take the treatment:

$$E[\delta|D=1] = E[Y^1 - Y^0|D=1] = E[Y^1|D=1] - E[Y^0|D=1]$$

  - E.g., how the MSBA program can help its graduates get better jobs?
- The average treatment effect for those did not take the treatment:

$$E[\delta|D=0] = E[Y^1 - Y^0|D=0] = E[Y^1|D=0] - E[Y^0|D=0]$$

  - E.g., how much the firm can benefit from extending the training program to the rest of workers?
- Again, the naïve estimator works only when the independence condition holds.



Conditioning

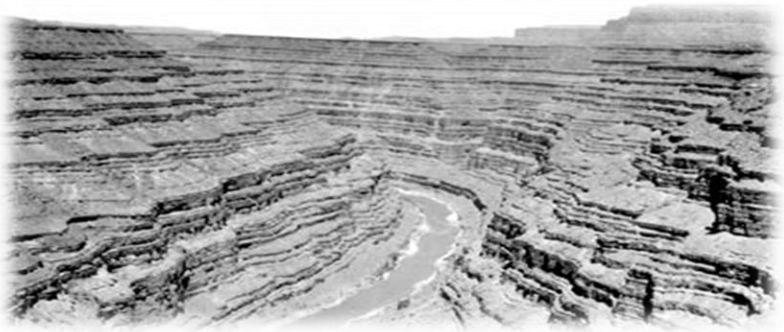
- Suppose assignment only depends on gender (S).
  - Suppose gender is observable and we know the probability of assignment: $\Pr(D = 1|S = M) = 0.7$; $\Pr(D = 1|S = F) = 0.4$.

- Assignment is completely random within each gender class (stratum).
  - Then we have **conditional independence**: $(Y^0, Y^1) \perp D|S$.
  - $E[Y^1|D = 1, S] = E[Y^1|S]$; $E[Y^0|D = 0, S] = E[Y^0|S]$

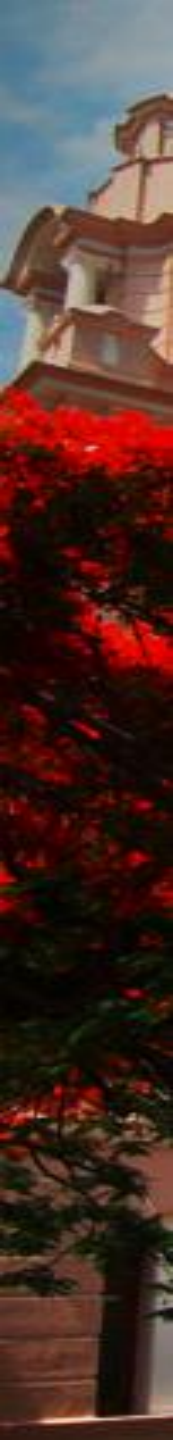
- We can estimate the average treatment effect by first estimating the (locally) average treatment effect for each gender class and then taking the weighted average over two gender classes.

Matching: Conditioning via Stratification

- The basic idea of matching is to put cases of the same characteristics observed as S together and then calculate the average treatment effect for a given value of S.
  - The motivation is to avoid selection bias.

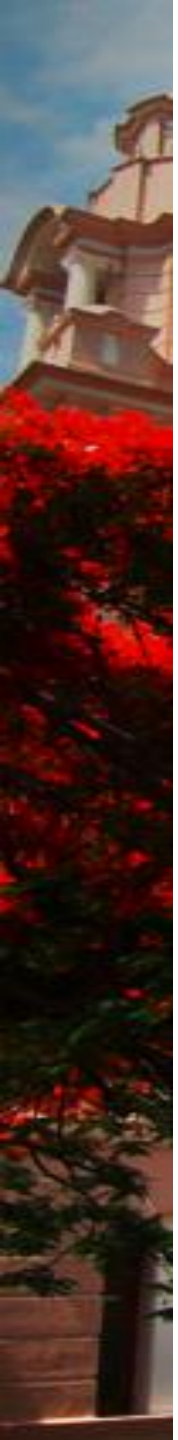


- For matching, we assume:
- Characteristic variables in S are perfectly observable such that the data can be perfectly stratified according to S.
- Treatment assignment in a given stratum is completely random.



Matching

- $Y =$ a measure of an individual's economic success at age 40
- $D =$ an indicator of receipt of a college degree
- $S = \{S_1, S_2\}$ is a set of characteristics that satisfies $(Y^0, Y^1) \perp D|S$.
- $S_1 =$ family background which can be high or low
- $S_2 =$ preparedness for college (e.g., IQ) which can be high or low
- Suppose that S_1 is irrelevant when $S_2 =$ low
- Let $S = \begin{cases} 1 & \text{if } S_2 = \text{low} \\ 2 & \text{if } S_2 = \text{high}, \text{if } S_1 = \text{low} \\ 3 & \text{if } S_2 = \text{high}, \text{if } S_1 = \text{high} \end{cases}$
- We have a large enough sample and the following information.

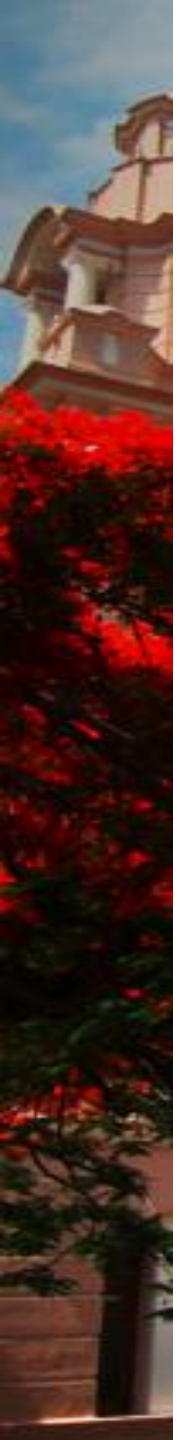


Matching

Joint probability	D = 0	D = 1	Marginal
S = 1	0.36	0.08	0.44
S = 2	0.12	0.12	0.24
S = 3	0.12	0.20	0.32
Marginal	0.6	0.4	

Potential outcome	D = 0	D = 1	
S = 1	$E[Y^0\|S]$ = 2	$E[Y^1\|S]$ = 4	$E[Y^1 - Y^0\|S]$ = 2
S = 2	$E[Y^0\|S]$ = 6	$E[Y^1\|S]$ = 8	$E[Y^1 - Y^0\|S]$ = 2
S = 3	$E[Y^0\|S]$ = 10	$E[Y^1\|S]$ = 14	$E[Y^1 - Y^0\|S]$ = 4
	$E[Y^0\|D = 0]$ = 4.4	$E[Y^1\|D = 1]$ = 10.2	

Index	S	D	Y
1	3	1	14.29901
2	2	0	5.869285
3	1	0	1.795601
4	1	0	2.260264
5	1	0	1.930357
6	3	1	14.13401
7	2	1	7.940315
8	2	0	5.581928
9	1	0	1.878888
10	1	0	2.166757
11	1	0	2.098094
12	1	0	1.822028
13	3	0	10.41953
14	1	0	1.812157
15	2	1	7.506354
16	2	0	5.629711
17	3	1	13.97574
18	1	0	1.65853
19	1	0	2.283716
20	2	0	6.482382
21	1	0	2.352635
22	1	0	1.720699
23	3	0	10.27105
24	3	1	13.86354
25	2	1	8.039263
26	1	0	1.900532
27	1	0	2.312755
28	3	0	10.04247
29	1	0	1.641661
30	1	0	2.148323
31	1	0	1.857471
32	2	0	6.317677
33	3	1	14.13349
34	2	0	6.135376
35	1	0	2.220093
36	1	0	2.204963
37	3	0	9.667611
38	1	0	1.670939
39	1	0	1.563837
40	2	1	8.494299
41	1	0	1.909293
42	1	0	2.175926
43	1	0	1.745386
44	3	1	14.31643
45	2	0	6.476645
46	3	0	10.12682
47	2	1	8.11968
48	1	1	3.962736
49	3	1	13.78759
50	1	0	1.762999



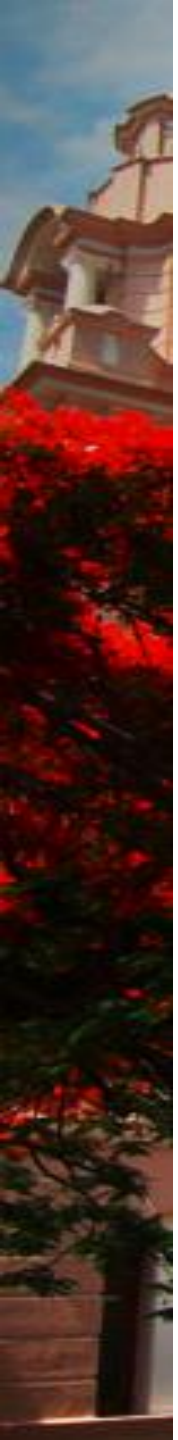
Matching

- Estimating the impact of a college degree
- $\hat{\delta}_{\text{naive}} = 10.2 - 4.4 = 5.8$

- The unconditional average treatment effect
  $= (4 - 2)(0.44) + (8 - 6)(0.24) + (14 - 10)(0.32) = 2.64$

- The average treatment effect among the treated = ?
  - $(4 - 2)(0.08/0.4) + (8 - 6)(0.12/0.4) + (14 - 10)(0.2/0.4) = 3$

- The average treatment effect among the untreated = ?
  - $(4 - 2)(0.36/0.6) + (8 - 6)(0.12/0.6) + (14 - 10)(0.12/0.6) = 2.4$



Matching

- Now suppose $\Pr(D = 1, S = 1) = 0$ and $\Pr(D = 0, S = 1) = 0.4$

Joint probability	D = 0	D = 1	Marginal
S = 1	0.40	0.00	0.40
S = 2	0.12	0.12	0.24
S = 3	0.12	0.24	0.36
Marginal	0.64	0.36	

- Due to lack of data, $E[Y^1|S = 1]$ cannot be estimated. Hence, the unconditional average treatment effect and the average treatment effect among the untreated cannot be estimated.
- However, the average treatment effect among the treated can be estimated: $(8 - 6)(0.12/0.36) + (14 - 10)(0.24/0.36) = 3.33$.

Case: Kjell and Company

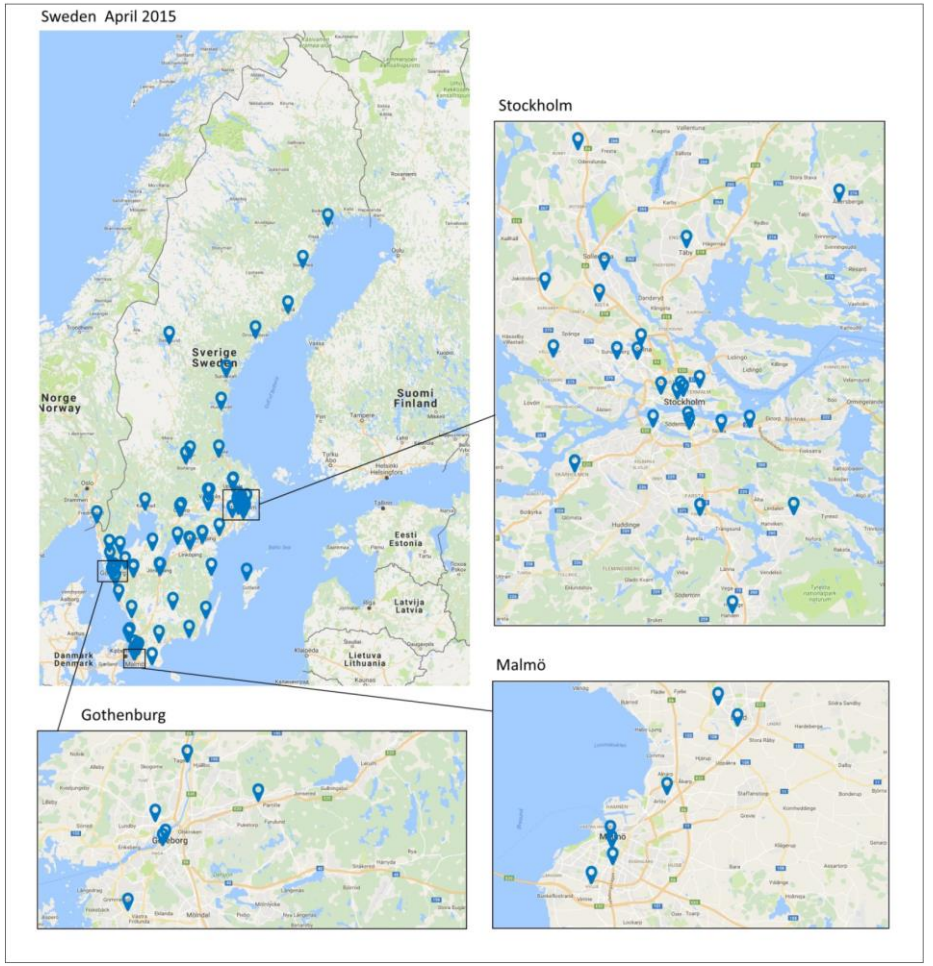
- Facts about Kjell:
  - A Swedish retail electronics chain …

Which of the following statement is true?
(A) The company had 90 stores in 2015.
(B) Kjell is the second-largest in Sweden and its products consisted mostly of smartphones.
(C) If a salesperson worked for 100 hours a month and generated sales of SEK 250,000, his commission would be SEK 1,651.
(D) In April 2015, the average SPH nationally is about SEK 1,491.

- Managerial Challenge:
  - Whether to change the compensation structure of Kjell's in-store salespeople
  - Whether to use a control group

Case: Kjell and Company

- The experiment design:
  - Treatment group: stores across the country
  - Control group: 5 stores located in the metropolitan areas of the three major cities, resembling their neighboring counterparts, but at a distance from other stores.
- Potential problems:
  - Can customers from metropolitan areas represent the whole country?
  - Can salespeople in the control group represent the whole country?

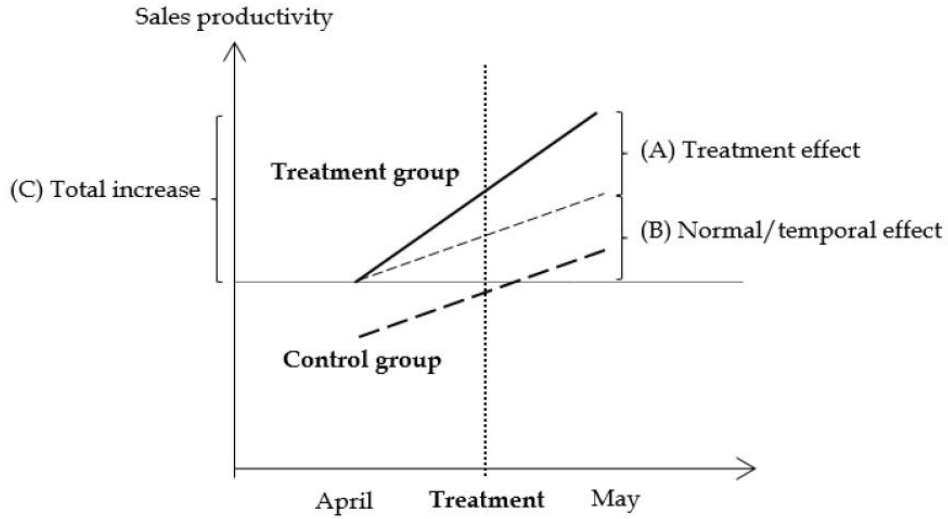


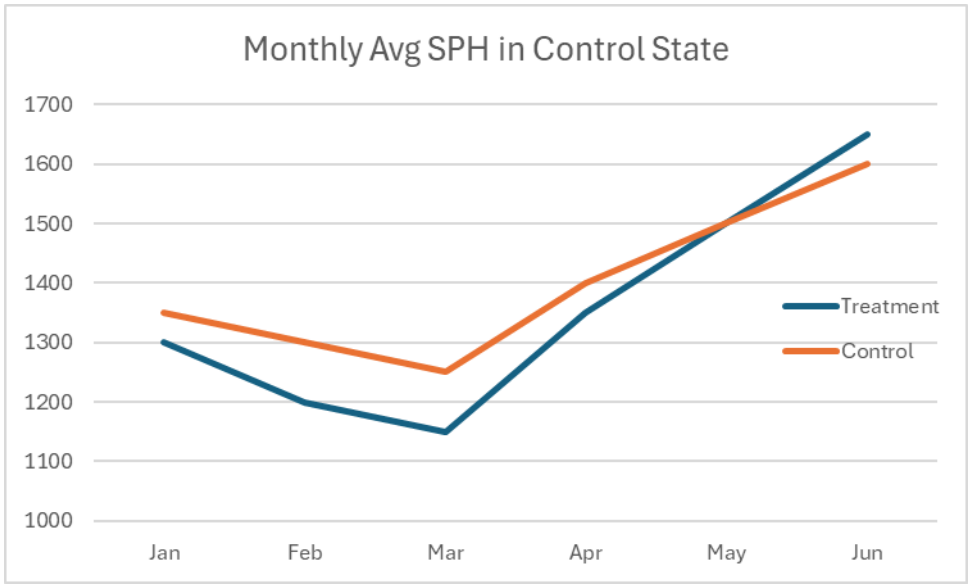
Case: Kjell and Company

- How to interpret the first month's results?

Table 1 Average SPH, Control Group and Treatment Group (in SEK)

Group/Month	April	May	% Change
Control	1,491.72	1,627.48	9.10%
Treatment	1,490.64	1,639.61	9.99%





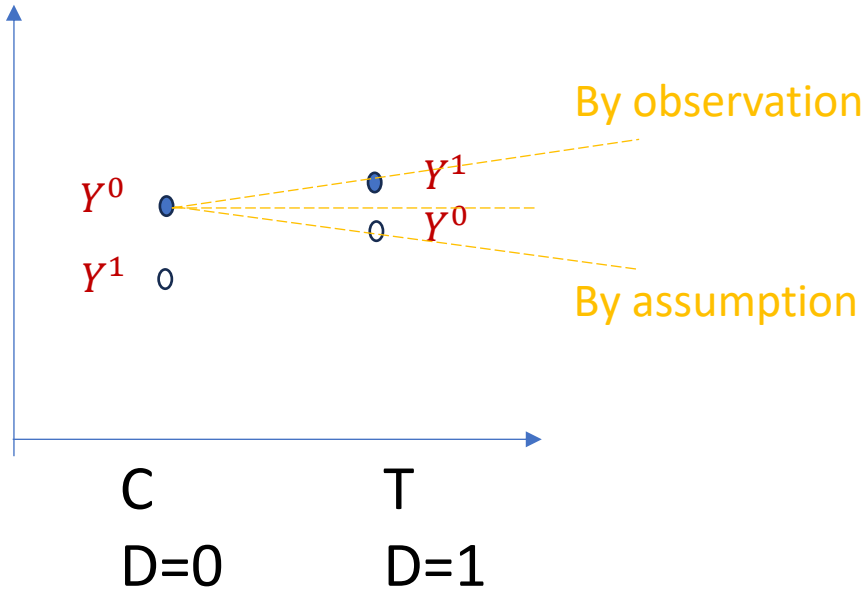
What if the control states are like this?

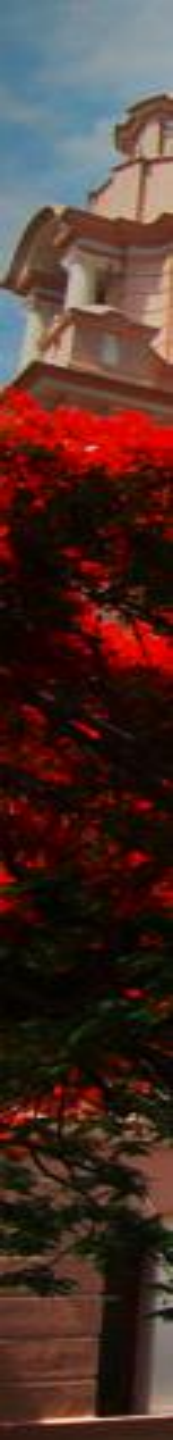
Case: Kjell and Company

- A biased estimation?
  - What if the control group stores are more likely to have Segment 3 and 4 salespeople?

Table 2 The Change in SPH, by Segment

Type	% Change
Segment 1	11.80%
Segment 2	2.00%
Segment 3	-3.70%
Segment 4	-8.10%





Case: Kjell and Company

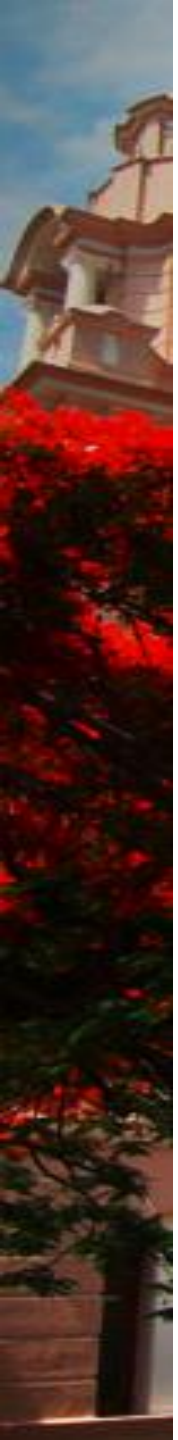
- Other possible impacts
  - The increase in sales under daily-quota plan might push salespeople to be overaggressive, pushing customers to buy unnecessary products. But this is not supported by the data on average.
  - Most salespeople sold low-margin products under daily-quota plan. Why?

Table 1 The Change in the Returns-to-Sales

Type	% Change
Segment 1	0.71%
Segment 2	-0.24%
Segment 3	-0.99%
Segment 4	-0.77%

Table 1 The Change in Product Quantity and Price, by Segment

Type	% Change in sales quantity (in units)	% Change in price per unit
Segment 1	13.64%	-0.14%
Segment 2	2.25%	-1.18%
Segment 3	-0.17%	-1.12%
Segment 4	-0.40%	-4.57%



Quiz

- Based on information below, the average treatment effect among the untreated is 0.30. True or False?

Joint probability	D = 0	D = 1	Marginal
S = 1	0.22	0.18	0.40
S = 2	0.22	0.28	0.50
S = 3	0.06	0.04	0.10
Marginal	0.50	0.50	

Potential outcome	D = 0	D = 1	
S = 1	$E[Y^0\|S]$ = 3	$E[Y^1\|S]$ = 4	$E[Y^1 - Y^0\|S]$ = 1
S = 2	$E[Y^0\|S]$ = 6	$E[Y^1\|S]$ = 6	$E[Y^1 - Y^0\|S]$ = 0
S = 3	$E[Y^0\|S]$ = 9	$E[Y^1\|S]$ = 8	$E[Y^1 - Y^0\|S]$ = -1