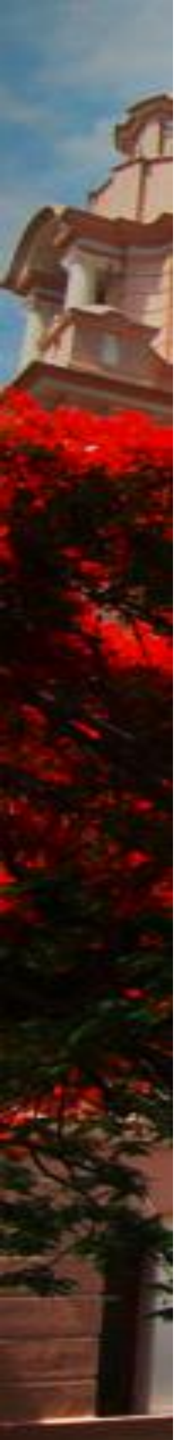
MSBA7003 Decision Analytics

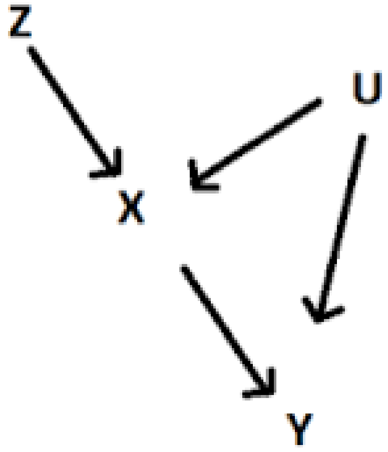
ZHANG, Wei
Associate Professor
HKU Business School

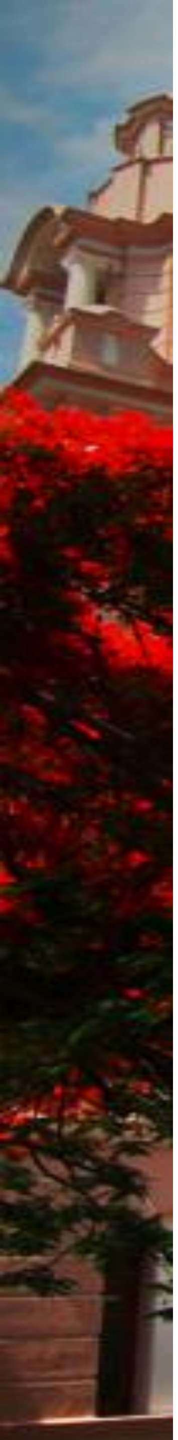
09 Causal Inference II



Agenda

- **Regression**
- **Causal Graphs**
- **Back-door-blocking Strategy**
- **Instrumental Variable**





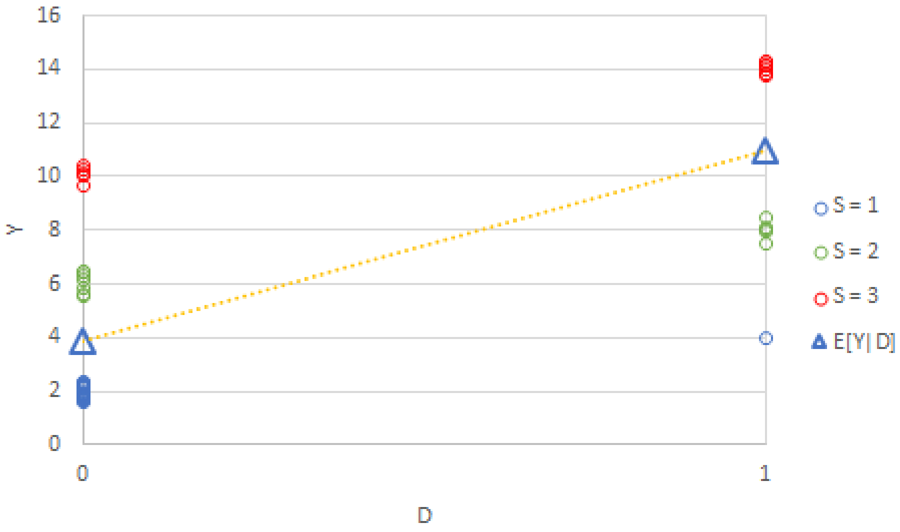
Regression

- **Theorem. Among all the functions of X, the least squared error predictor of Y given X is $E[Y|X]$.**
- Proof (*not required*):
- Suppose we want to find some function $\gamma(X)$ such that
$$\gamma(X) \equiv \arg\min_{g(X)} E\left[\left(Y - g(X)\right)^2\right]$$
- Let $\mu(X) \equiv E[Y|X]$
- Note that $E\left[\left(Y - g(X)\right)^2\right] = E\left[E\left[\left(Y - g(X)\right)^2 | X\right]\right]$
- And that $\left(Y - g(X)\right)^2 = \left(Y - \mu(X) + \mu(X) - g(X)\right)^2 = \left(Y - \mu(X)\right)^2 + 2\left(Y - \mu(X)\right)\left(\mu(X) - g(X)\right) + \left(\mu(X) - g(X)\right)^2$
- For the second term,
$$E\left[2\left(Y - \mu(X)\right)\left(\mu(X) - g(X)\right)|X\right] = 2\left(\mu(X) - g(X)\right)E\left[\left(Y - \mu(X)\right)|X\right] = 2\left(\mu(X) - g(X)\right)\left(E[Y|X] - \mu(X)\right) = 0$$
- It follows that
$$E\left[\left(Y - g(X)\right)^2\right] = E\left[\left(Y - \mu(X)\right)^2\right] + E\left[\left(\mu(X) - g(X)\right)^2\right]$$
- Hence, the expected squared error is minimized when $g(X) = \mu(X)$.
- Therefore, $\gamma(X) = E[Y|X]$ is the least squared error predictor.
- End of proof.

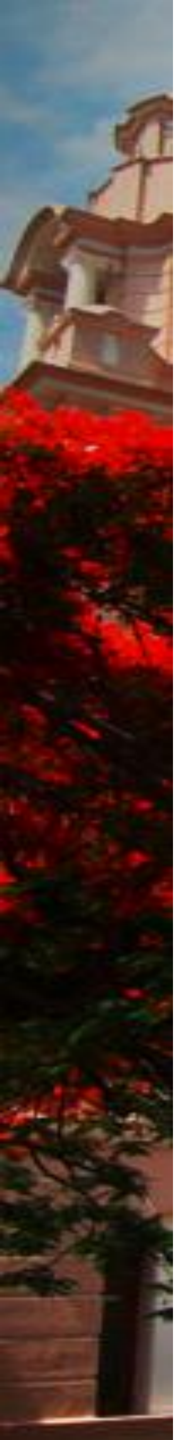
Regression

- If we regress Y against D, then the resulting predictive function
$$\hat{Y} = a + bD$$
is the least squared error predictor of Y given D.
- Suppose D can take only two values: 0 and 1. Then the regression line goes through the mean of Y given each value of D.
- Hence, $\hat{Y} = a + bD = E[Y|D]$ in this case.

- When D can take multiple values, the shape of $E[Y|D]$ could be nonlinear. A linear regression model only offers a linear approximation of $E[Y|D]$.



b represents the causal effect when independent condition holds!



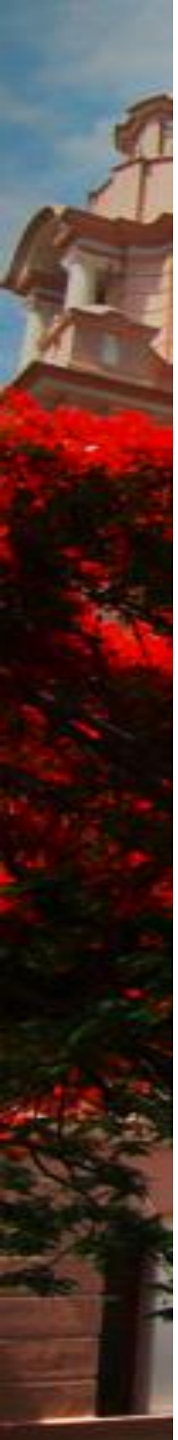
Regression

- Hence, a linear regression

$$\hat{Y} = \hat{\alpha} + \hat{\beta} S + \hat{\delta} D$$

only estimates the best linear approximation of the conditional expectation $E[Y|D,S]$.

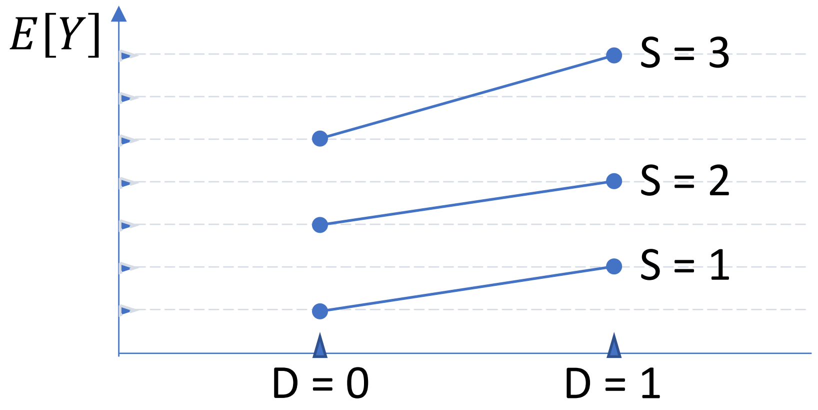
- In general, the coefficients have no causal interpretation. In particular, $\hat{\delta}$ can be interpreted as the naïve estimator for treatment D given stratum S.
  - $\hat{\delta} = E[Y|D=1,S] - E[Y|D=0,S]$
- However, when the following conditions are satisfied, regression can be used as a conditioning strategy to estimate causal effects.
  - (1) S allows a perfect stratification (i.e., treatment assignment in a given stratum is completely random)
  - (2) D is exogenous (i.e., Y does not cause D)
  - (3) The functional form is correctly specified (e.g., the causal effect is constant)



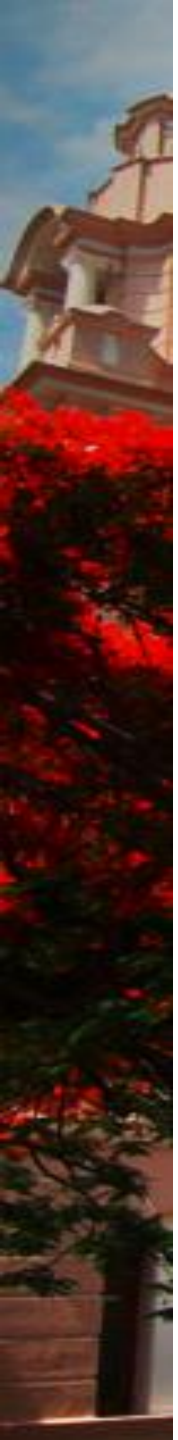
Regression

- How to build the regression model when the causal effect is not constant?

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



- Notice that different strata have different Y^1, Y^0, and $Y^1 - Y^0$.
- Can we correctly estimate the effect of D in each stratum?
- Can we correctly estimate the overall effect of D?

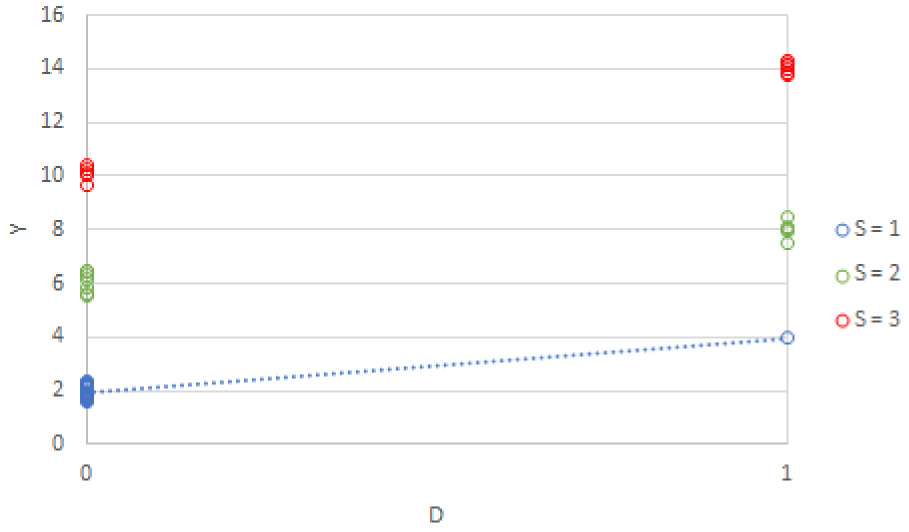


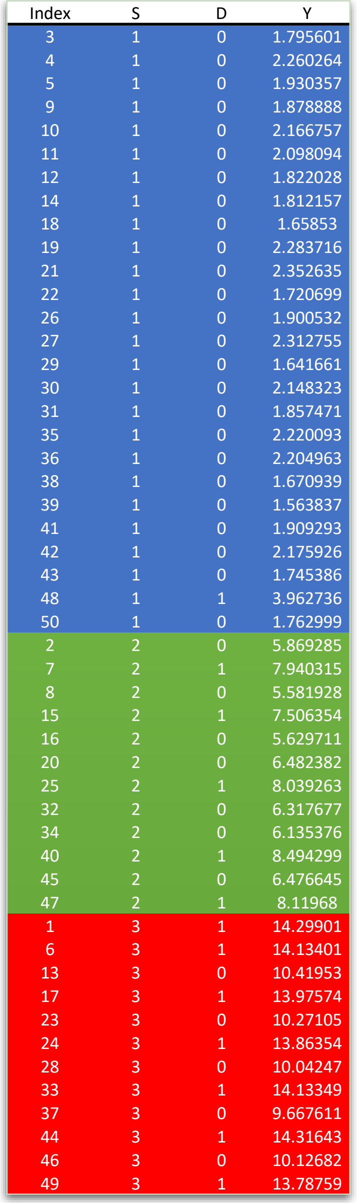
Regression

- To capture the different Y^1, Y^0, and $Y^1 - Y^0$, we need to
  - introduce dummy variables: S_2 and S_3 (representing whether the individual is $S = 2$ and $S = 3$, respectively, using $S = 1$ as baseline)
  - add interaction terms: $D \times S_2$ and $D \times S_3$
- The correct regression model is:

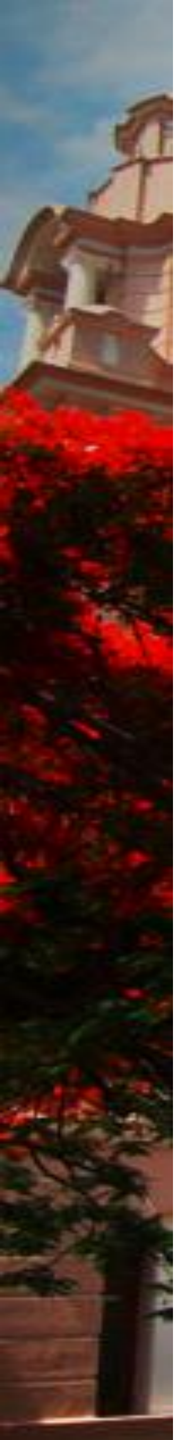
$$Y = \alpha + \beta_2 S_2 + \beta_3 S_3 + \delta_1 D + \delta_2 D \times S_2 + \delta_3 D \times S_3 + \epsilon$$

When $S_2 = S_3 = 0$, the regression line $\hat{Y} = \alpha + \delta_1 D$ goes through $E[Y^0|S = 1]$ and $E[Y^1|S = 1]$. The slope is $\delta_1 = E[Y^1 - Y^0|S = 1]$ = 2.



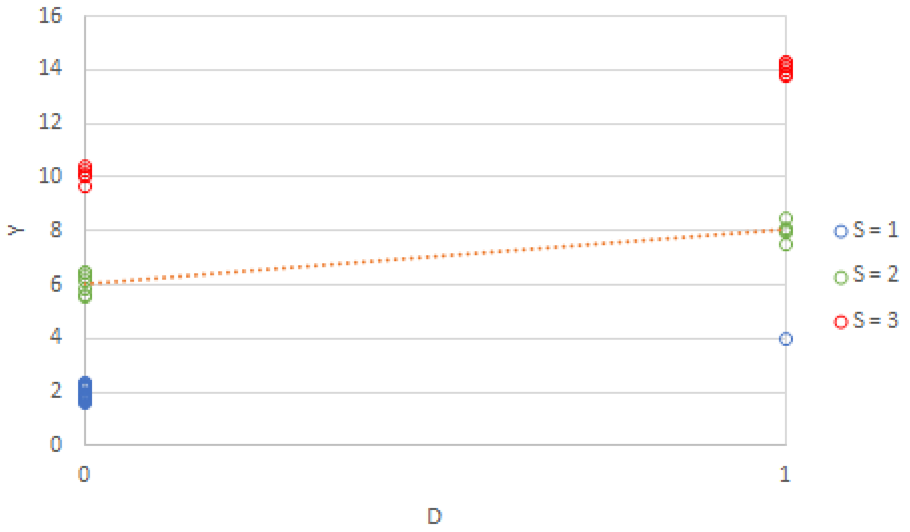


Index	S	D	Y
3	1	0	1.795601
4	1	0	2.260264
5	1	0	1.930357
9	1	0	1.878888
10	1	0	2.166757
11	1	0	2.098094
12	1	0	1.822028
14	1	0	1.812157
18	1	0	1.65853
19	1	0	2.283716
21	1	0	2.352635
22	1	0	1.720699
26	1	0	1.900532
27	1	0	2.312755
29	1	0	1.641661
30	1	0	2.148323
31	1	0	1.857471
35	1	0	2.220093
36	1	0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

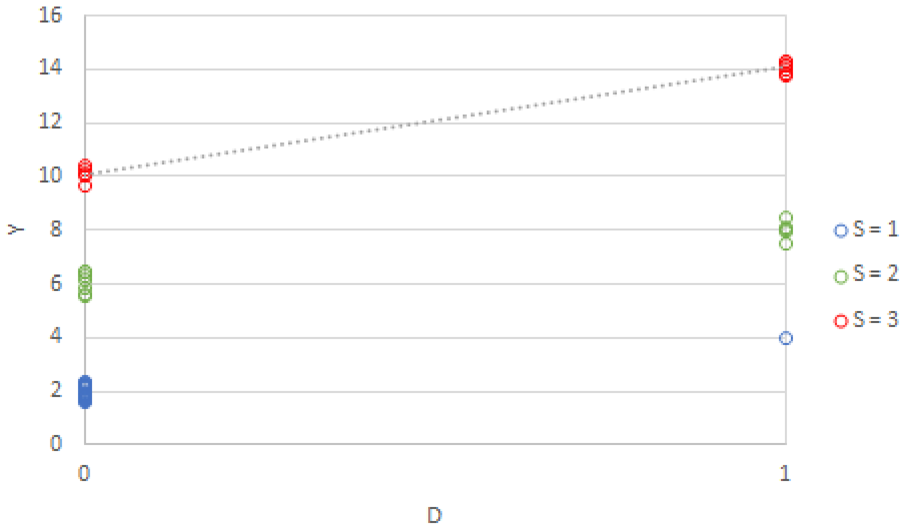


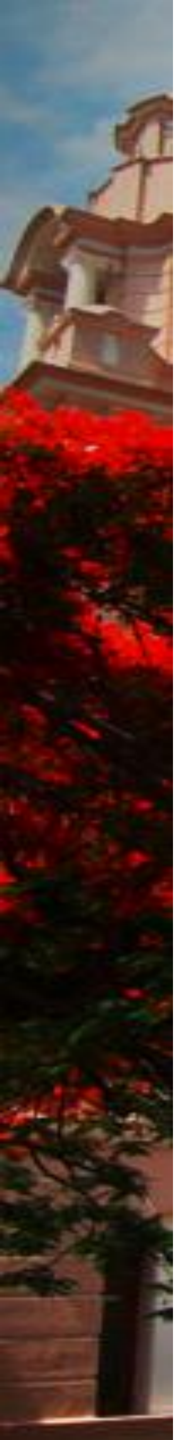
Regression

When $S_2 = 1$ and $S_3 = 0$, the regression line $\hat{Y} = \alpha + \beta_2 + (\delta_1 + \delta_2)D$ goes through $E[Y^0|S = 2]$ and $E[Y^1|S = 2]$. The slope is $\delta_1 + \delta_2 = E[Y^1 - Y^0|S = 2]$ = 2. Hence, $\delta_2 = 0$.



When $S_2 = 0$ and $S_3 = 1$, the regression line $\hat{Y} = \alpha + \beta_3 + (\delta_1 + \delta_3)D$ goes through $E[Y^0|S = 3]$ and $E[Y^1|S = 3]$. The slope is $\delta_1 + \delta_3 = E[Y^1 - Y^0|S = 3]$ = 4. Hence, $\delta_3 = 2$.



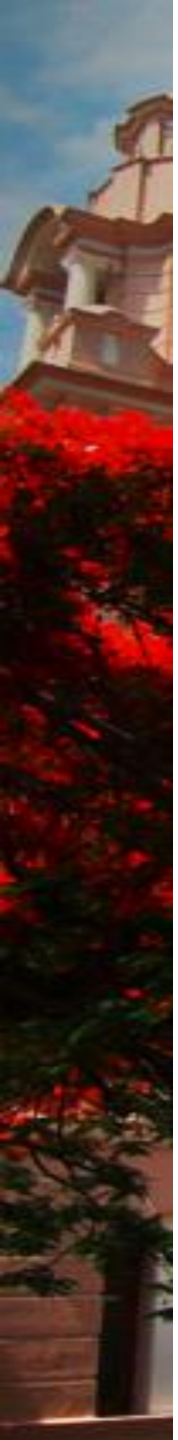


Regression

- Summary: with regression model

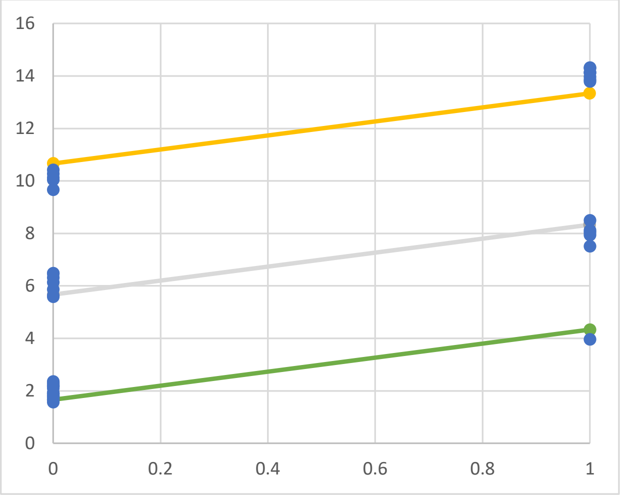
$$Y = \alpha + \beta_2 S_2 + \beta_3 S_3 + \delta_1 D + \delta_2 D \times S_2 + \delta_3 D \times S_3 + \epsilon$$

- α: the average performance of stratum 1 under the control state
- β_2: the additional average performance of stratum 2 against stratum 1 under the control state
- β_3: the additional average performance of stratum 3 against stratum 1 under the control state
- δ_1: the average treatment effect of the decision on stratum 1
- δ_2: the additional average treatment effect of the decision on stratum 2 against stratum 1
- δ_3: the additional average treatment effect of the decision on stratum 3 against stratum 1

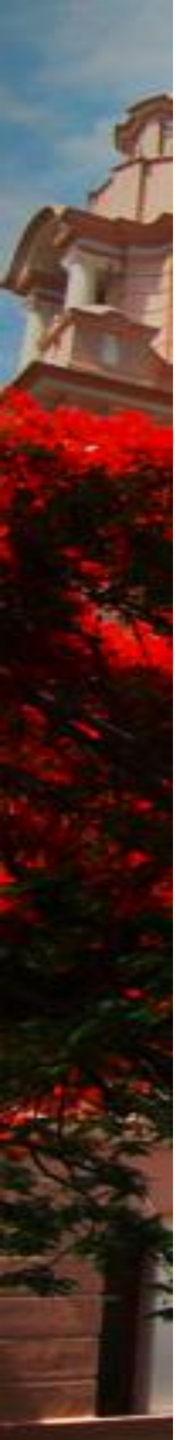


Regression

- If the regression model is $\hat{Y} = \alpha + \beta_2 S_2 + \beta_3 S_3 + \delta_1 D$, how to understand δ_1?
  - The model accommodates different Y^1 and Y^0,
  - but does not allow different $Y^1 - Y^0$.
  - Essentially, it uses three parallel lines to fit the data.
  - The slope depends on the distribution of data in each stratum.
  - Hence, δ_1 is not the overall effect of D in general.



- Regression specification is generally challenging.
- For binary causes, matching is in general preferred. In comparison with regression, matching is nonparametric.



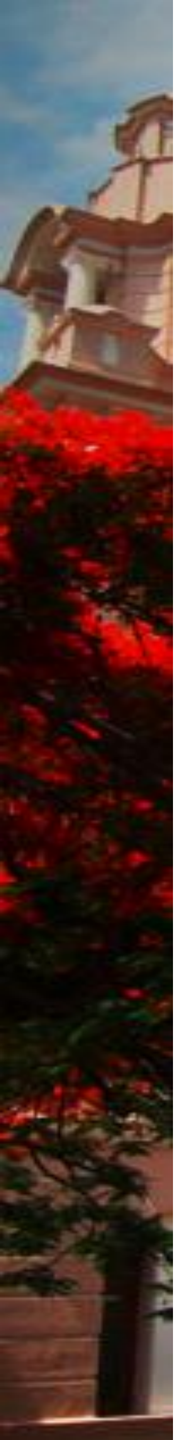
In-Class Exercise

- We would like to estimate the impact of smoking (D) for the general population with a random sample. In the sample, people can be classified along two dimensions. One is smoker versus non-smoker, and the other is high family income versus low family income. The number of people in each class and their average health condition are shown in the following tables.

Number	Smoker	Non-Smoker	Total
High Family Income	3	8	11
Low Family Income	7	3	10
Total	10	11	21

Health Condition	Smoker	Non-Smoker	Average
High Family Income	6	10	8.91
Low Family Income	5	8	5.90
Average	5.30	9.45	

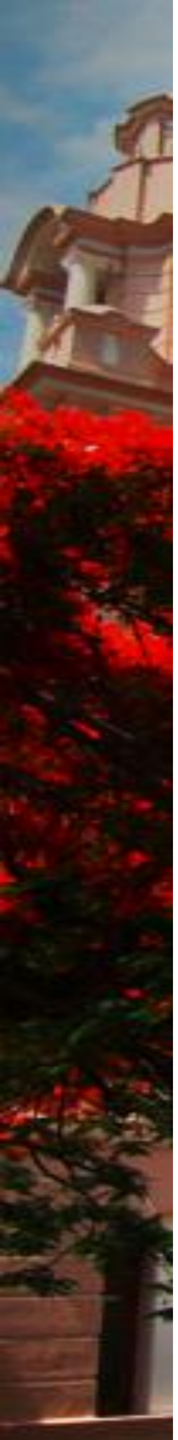
- What is the Naïve estimator of the impact of smoking on health?
- Why is the Naïve estimator biased?
- If we use regression to estimate the impact of smoking based on the given data, what could be the most appropriate model specification (or functional form)?
- If health conditions (Y^0, Y^1) are independent of smoking, conditioning on family income, what is the correct estimate of the impact of smoking on health for the general population?



In-Class Exercise

- True or False?

1. A restaurant analyzed its historical operational data (daily) through a linear regression as follows: E[profit] = 3.3 + 1.2*numberCustomer – 0.55*weekday + 0.43*avgPrice. Then, the expected profit can be increased by 0.43 if the average price of food on the menu is increased by 1.
2. We want to estimate the causal effect of D through regression, and S allows a perfect stratification. There are N strata in S, and Si (0-1 variable) represents stratum i (where i goes from 2 to N). Suppose the regression equation has all Si as well as the interaction terms Si*D. If the estimated coefficient of Si*D is A and is significant, then the treatment effect of D for stratum i is A.
3. If the treatment effect is the same among all strata, then we can just regress Y against D to estimate the treatment effect.

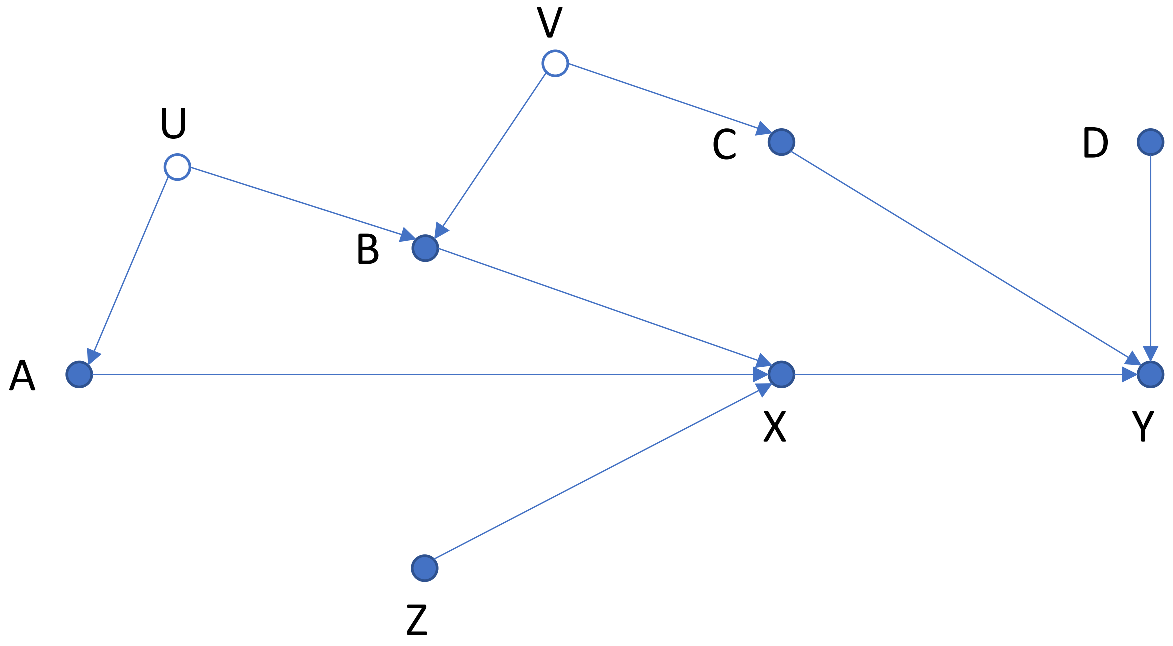


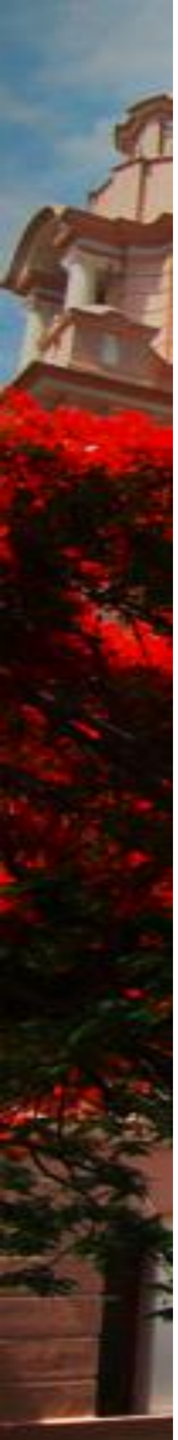
Causal Graphs

- How to make sure that the control variables allow a perfect stratification?
- We need causal graphs to conceptualize our problem.
- The goal is to represent explicitly all causes of the outcome.

- Each node of a causal graph represents a random variable
  - Labeled by a letter such as A, B, or C
- Observed variables are represented by a solid circle ●
- Unobserved variables are represented by a hollow circle ○
- Causes are represented by a directed edge → (i.e., single-headed arrow), such that an edge from one node to another signifies that the variable at the origin causes the variable at the terminus.

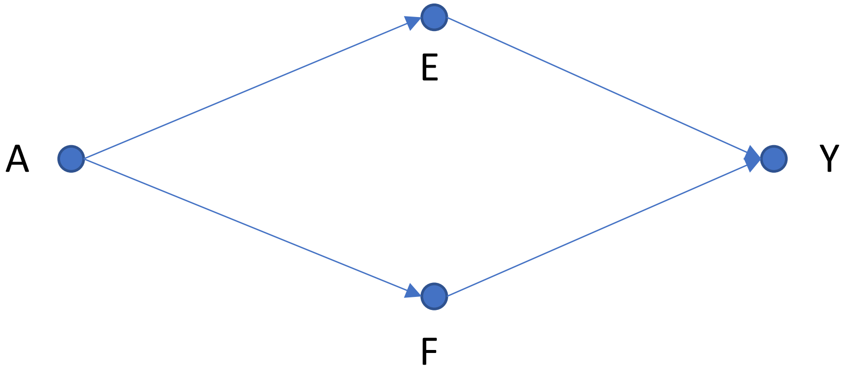
Causal Graphs

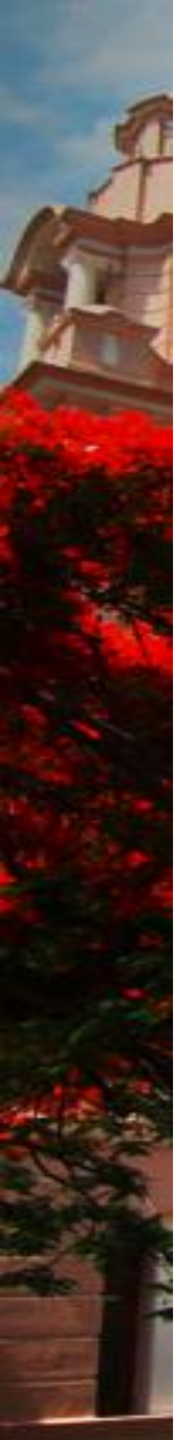




Transitivity of Causal Relations

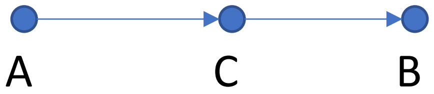
- For any three variables A, B, and C, if A causes B and B causes C, then A causes C.
  - This is an assumption. We assume it holds in most cases.
  - Counter-example: A = playing computer games; Y = academic performance; E = brain capability; F = study time (i.e., the effect of A on Y is cancelled out by different mechanisms.)



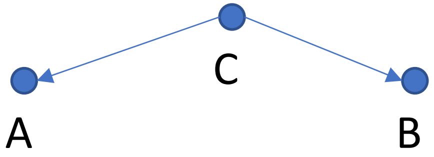


Causal Structures

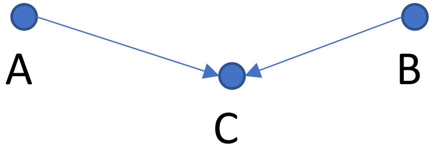
- There are three basic patterns of causal relationship that would exist for three variables that are structurally related to each other.
- (1) A chain of mediation:
  - Unconditional dependence: $P(AB)\neq P(A)P(B)$
  - Conditional independence: $P(AB|C)=P(A|C)P(B|C)$

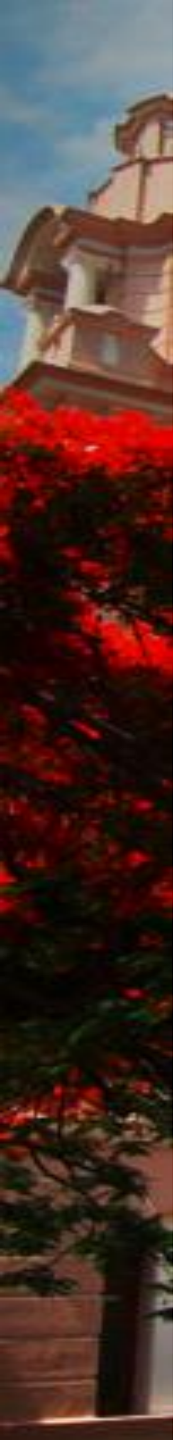


- (2) A fork of mutual dependence:
  - Unconditional dependence: $P(AB)\neq P(A)P(B)$
  - Conditional independence: $P(AB|C)=P(A|C)P(B|C)$



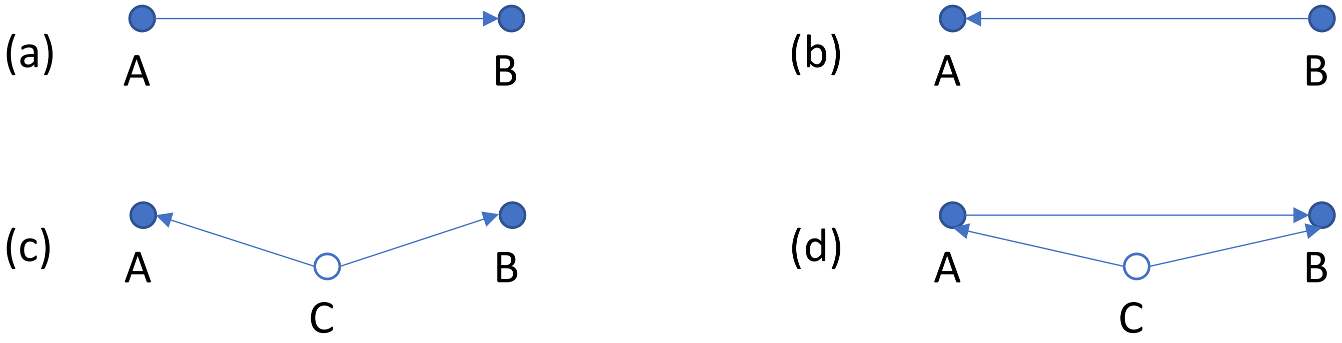
- (3) An inverted fork of mutual causation:
  - Unconditional independence: $P(AB)=P(A)P(B)$
  - Conditional dependence: $P(AB|C)\neq P(A|C)P(B|C)$



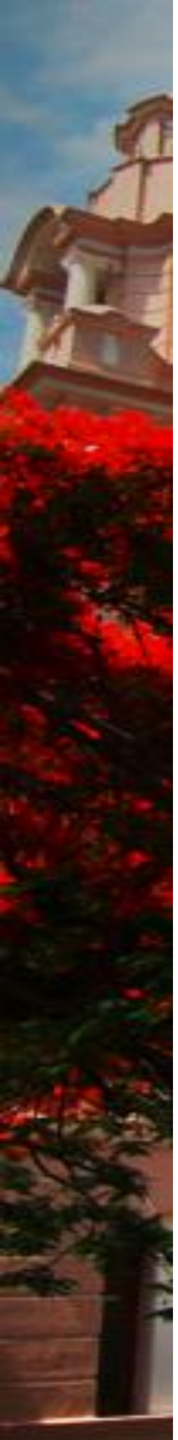


Unconditional Correlation

- The Reichenbach Principle: Any two variables A and B are unconditionally correlated if and only if either (a) A causes B, (b) B causes A, (c) a common cause C causes both A and B, or (d) any combination of (a)-(c).

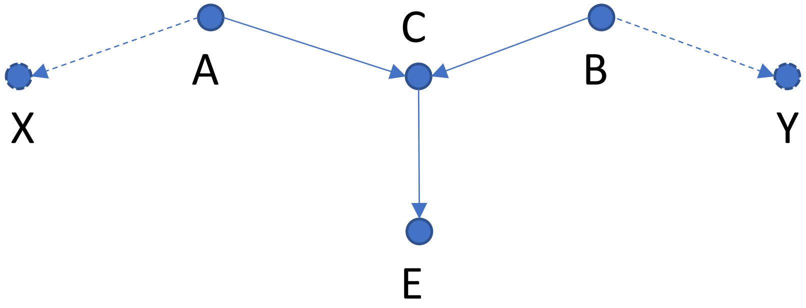


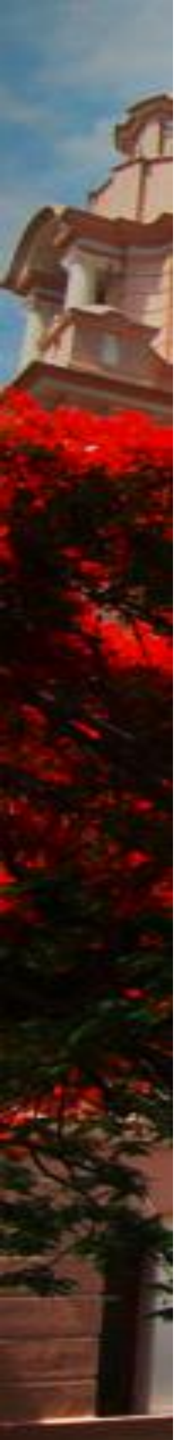
- Causes may be mediated.
- Correlation due to randomness is possible, but it will disappear as sample size increases.
- In sum, two variables are unconditionally correlated if and only if they share a driving factor.



Conditional Correlation

- Two variables are unconditionally independent, but are correlated conditioning on C (or E), if
  - (1) they are in the positions of A and B, or
  - (2) they are in the positions of X and B, or
  - (3) they are in the positions of X and Y.



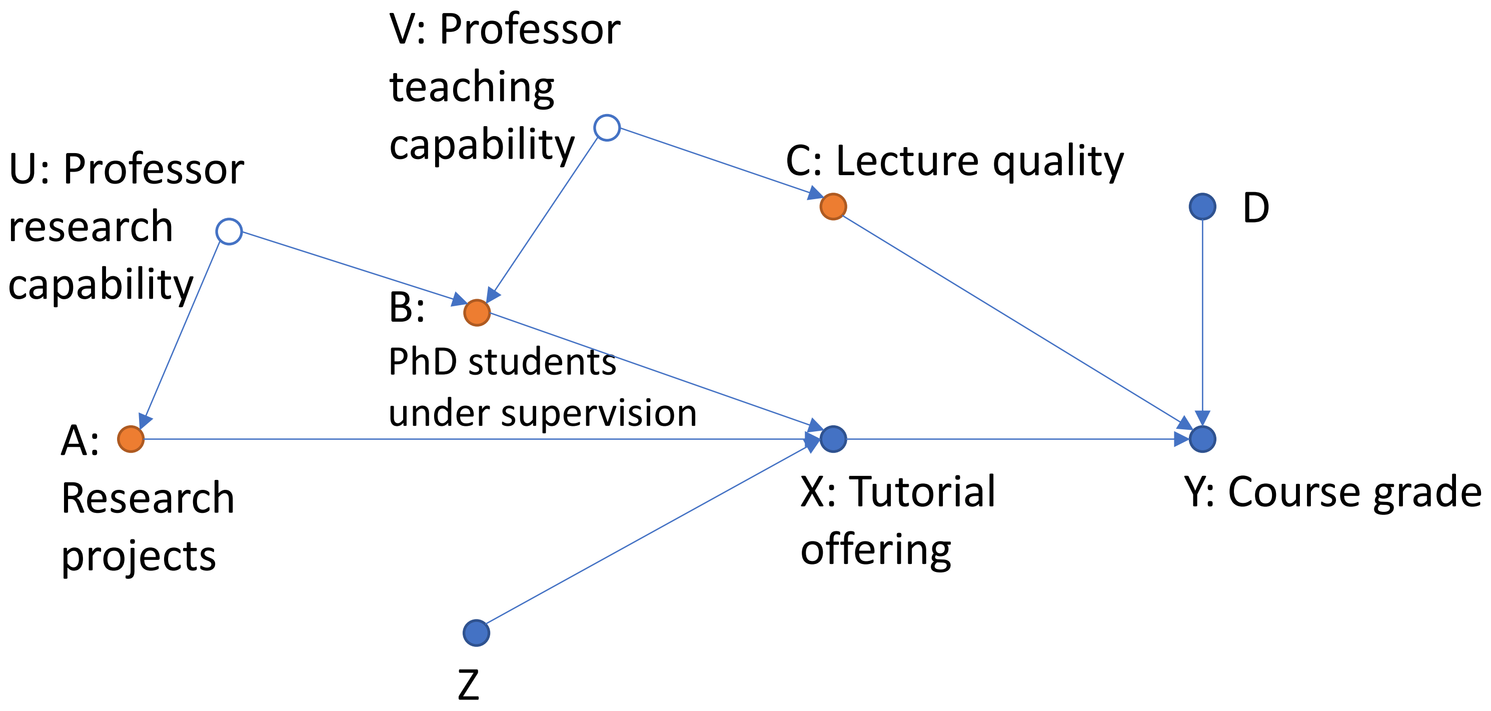


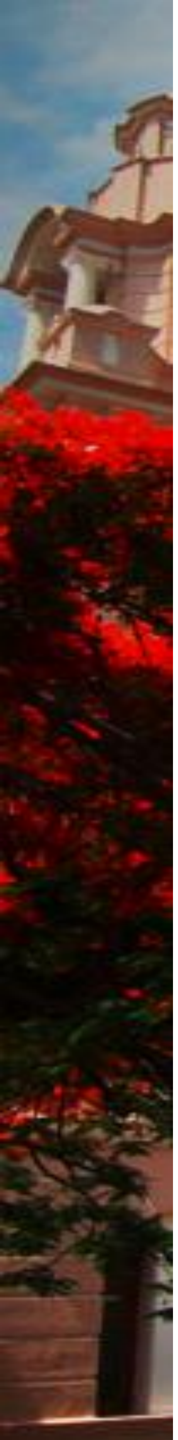
In-Class Exercise

- Which of the following statement is true?
  - If A and B are correlated, and B and C are correlated, then A and C must also be correlated.
  - If A and C are correlated, and B and C are correlated, then A and B could be independent.
  - If A and B are independent, and A and C are correlated, then B and C must be independent.
  - If A and B are independent unconditionally and are correlated conditioning on C, then A must be a cause of C.

Causal Inference Strategy

- Conditioning on “back-door-blocking” variables
  - Identify paths that unconditionally correlate X and Y; block the path by conditioning on a variable on the path
  - Condition on C, or
  - Condition on A and B


V: Professor teaching capability
U: Professor research capability
C: Lecture quality
D
B: PhD students under supervision
A: Research projects
X: Tutorial offering
Y: Course grade
Z

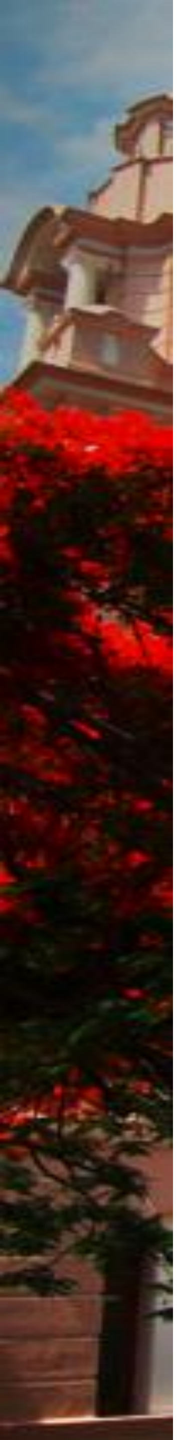


Back-door-blocking Strategy: An Example

- You are running an online shopping website that focuses on fashion apparel. Your company normally purchases a product from a supplier before the selling season starts.
- During the season, customers that purchase the product can give a rating of the product on the website.
- When the selling season ends, any leftovers will be shipped back to the supplier and partial refund will be provided.
- You want to use the historical data to estimate how customer rating of a product influences the sales of the product. You have prepared data for the following variables.

- Y: the total sales of a product
- X: the average customer rating
- P: the price of a product
- S: the set of dummies that indicate the category of a product

  (e.g., male versus female, season, and type)
- W: the beginning inventory level of a product
- C: the number of clicks associated with a product
- Q: product quality (measured by cost)

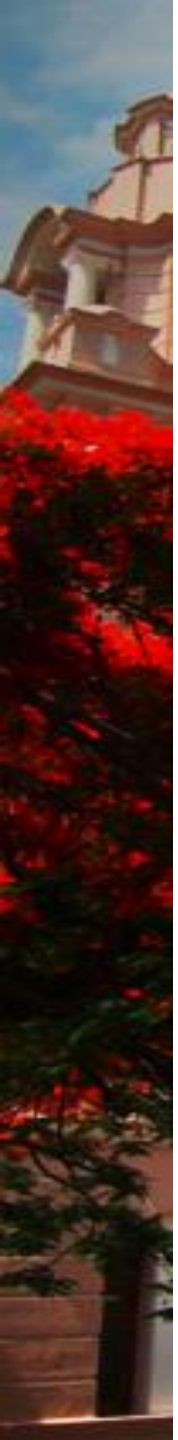


Back-door-blocking Strategy: An Example

- Draw the causal graph and find the BDB variable(s).

- Y: the total sales of a product
- X: the average customer rating
- P: the price of a product
- S: the set of dummies that indicate the category of a product

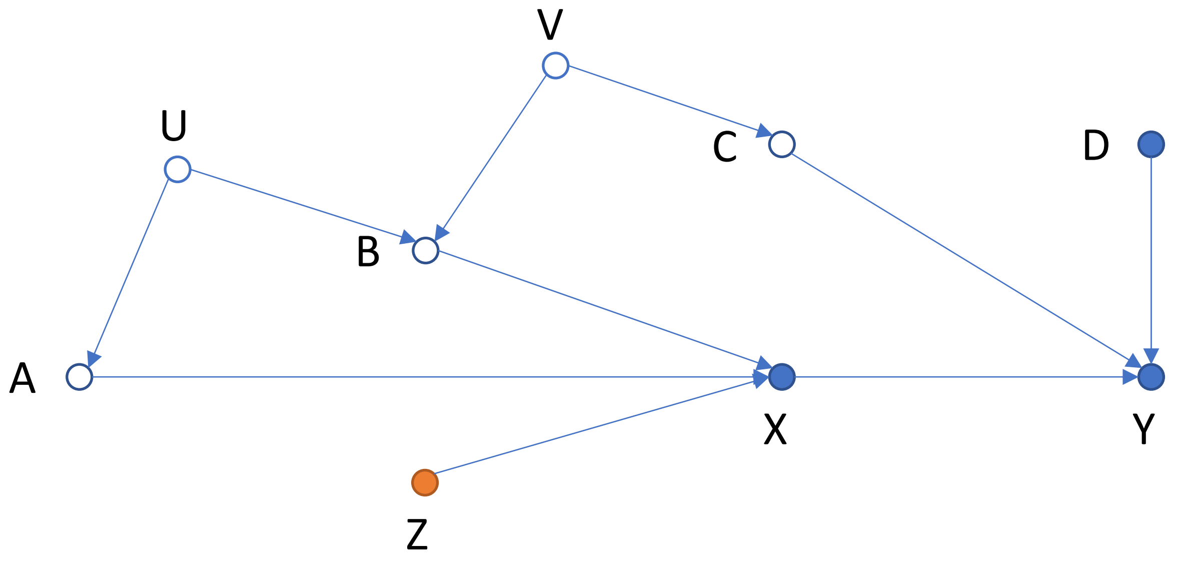
  (e.g., male versus female, season, and type)
- W: the beginning inventory level of a product
- C: the number of clicks associated with a product
- Q: product quality (measured by cost)

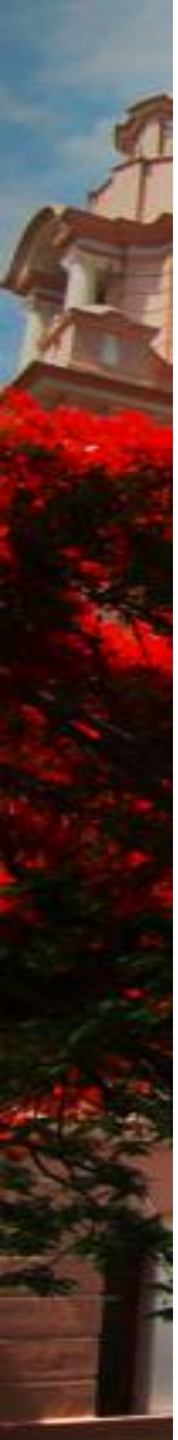


Quiz

When Conditioning Does Not Work

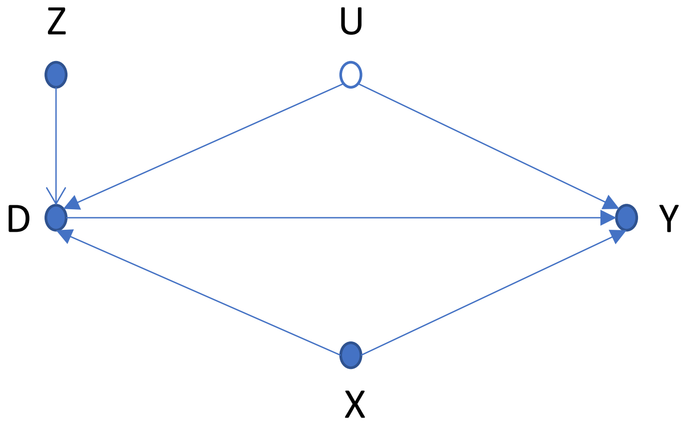
- When we cannot block all back-door paths, conditioning strategy does not work.
- We cannot use matching or simple regressions to estimate the effect of
  - Education on income with family background unknown
  - Screen size on online purchase behavior with individual income unknown
  - Firm revenue on default risk with economic condition unknown
- A possible remedy:
  - instrumental variable (IV) estimation
  - Z causes X
  - Z is related to Y only through X's effect
  - if Z and Y are correlated
  - it must be that X causes Y

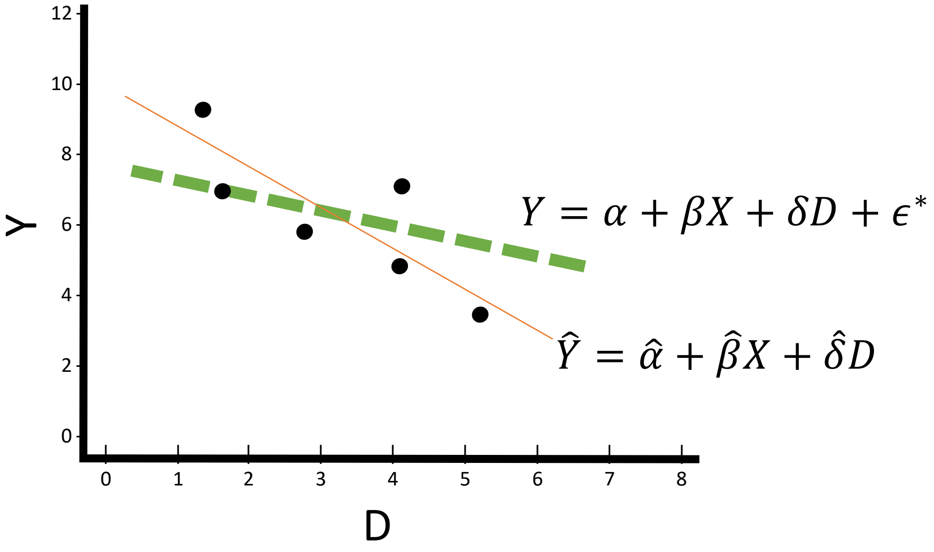


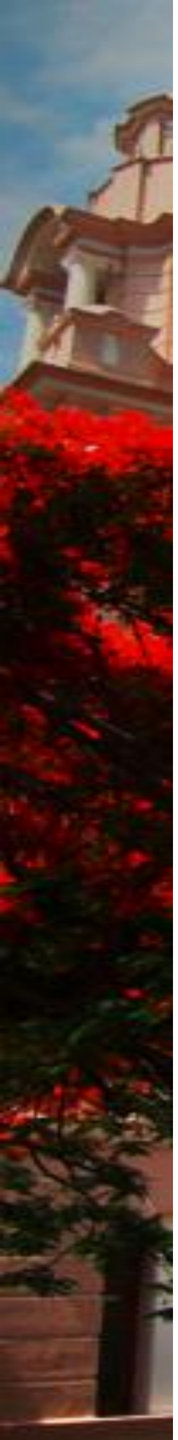


Instrumental Variable

- Suppose $Y = \alpha + \delta D + \beta X + \gamma U + \epsilon$ and the causal structure is given below. Assume that $\epsilon \perp (D, X)$ and that D can take many values.
  - E.g., Y is the total amount of smoking, D is years of education, U is family income, X is place of origin, and Z is the quarter of birth.
- What can go wrong if we use OLS method to estimate δ? Omitted Variable Bias.





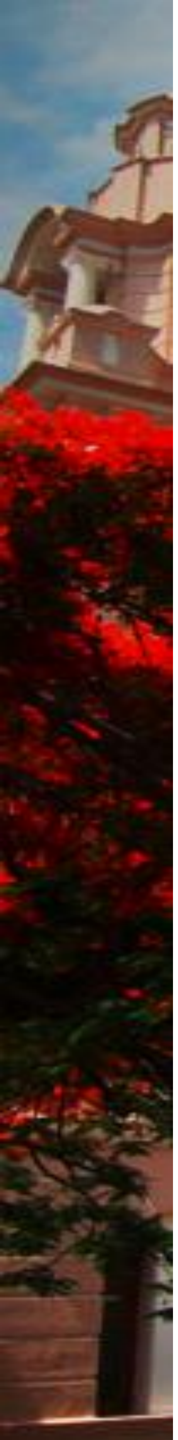


Instrumental Variable

- We can describe our causal system with two simultaneous equations:
- (1) $Y = \alpha + \delta D + \beta X + u$, where $u = \gamma U + \epsilon$.
- (2) $D = a + bZ + cX + v$, where v is correlated with u by U.

- If we develop a regression model based on equation (2), then b can be estimated without OVB. Let $\hat{b}$ denote the estimated b.
- If we replace equation (2) into (1), we have
- (3) $Y = (\alpha + \delta a) + \delta bZ + (\beta + \delta c)X + (u + \delta v) = A + BZ + CX + \varepsilon$

- Since Z is uncorrelated with ε , equation (3) and B can be estimated without OVB. Note that $B = \delta b$. Hence, δ can be consistently estimated by $\hat{B}/\hat{b}$ (in two steps).



Instrumental Variable

- Formally, for Z to be an instrumental variable for the causal relationship $(D \rightarrow Y)$, Z must satisfy two conditions:

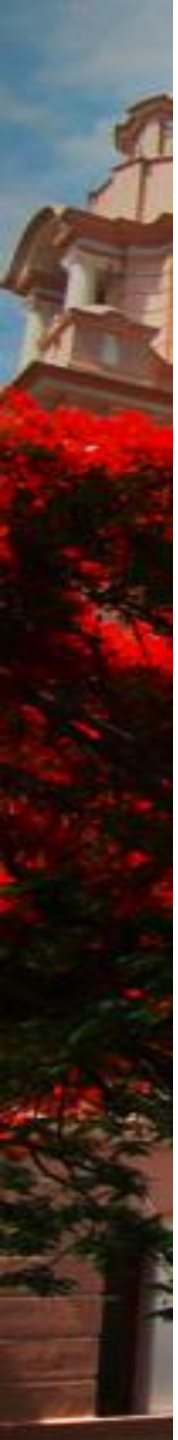
- (1) **Relevance:** $Cov(D, Z) \neq 0$
- The instrument must be correlated with the explanatory variable D.

Sources of variations of Y other than D.

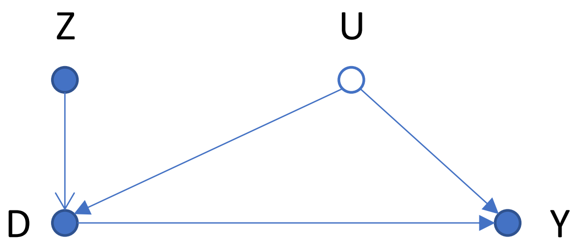
The true relationship of how Y depends on D.

- (2) **Exogeneity:** $Cov(\epsilon', Z) = 0$, where $\epsilon' = Y - \delta D$
- Z cannot be correlated with Y except through D's effect on Y.

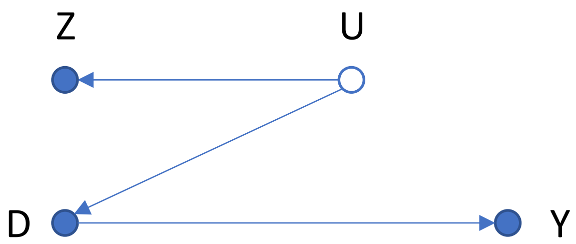
- (1) can be tested by regressing D on Z; (2) cannot be directly tested because δ cannot be correctly estimated without a valid IV.



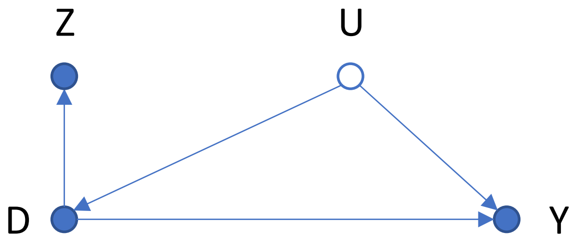
Instrumental Variable



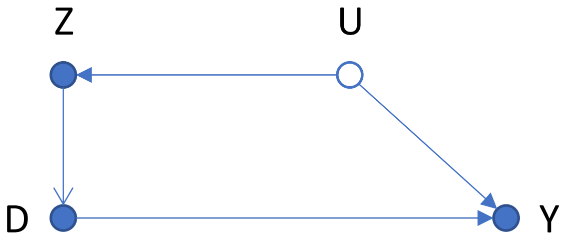
Z is a valid IV



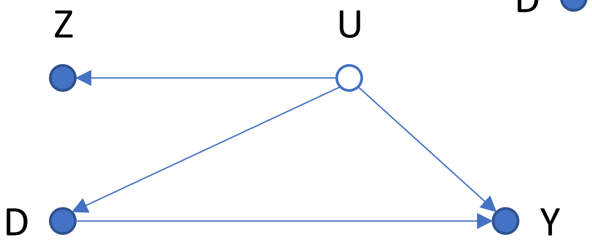
Z is a valid IV



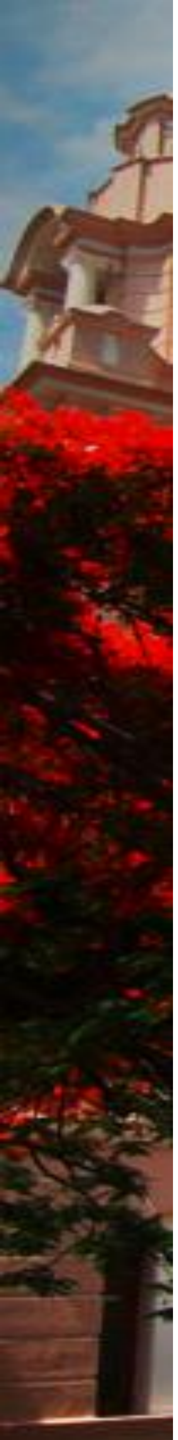
Is Z a valid IV?



Z is not a valid IV



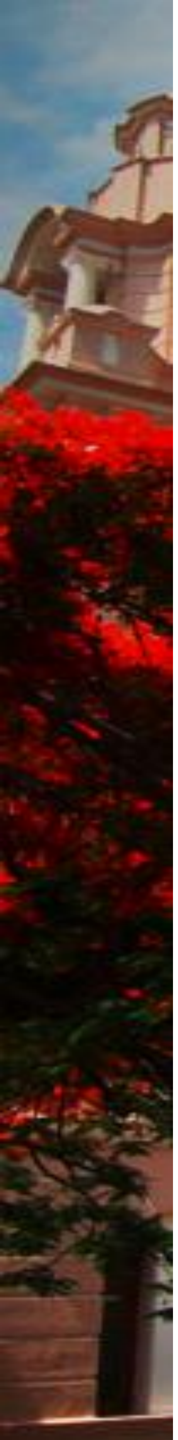
Z is not a valid IV



Examples of I.V.

- Using day of the week of hospital admission as an instrument for the effect of waiting time to surgery on length of stay and patient mortality. (Assume factures happen randomly. Many surgeons operate only on weekdays and, therefore, patients admitted on weekends may wait longer.)

- Using the quarter of birth as an instrument for the effect of years of schooling on subsequent earnings. (Students born in different months of the year start school at different ages. Students who are born early in the academic year are typically older when they enter school. If the fraction of students who desire to leave school after they reach the legal dropout age is constant across birthdays, those born in the beginning of the academic year will have less schooling. Caution: subpopulation & local average effect.)



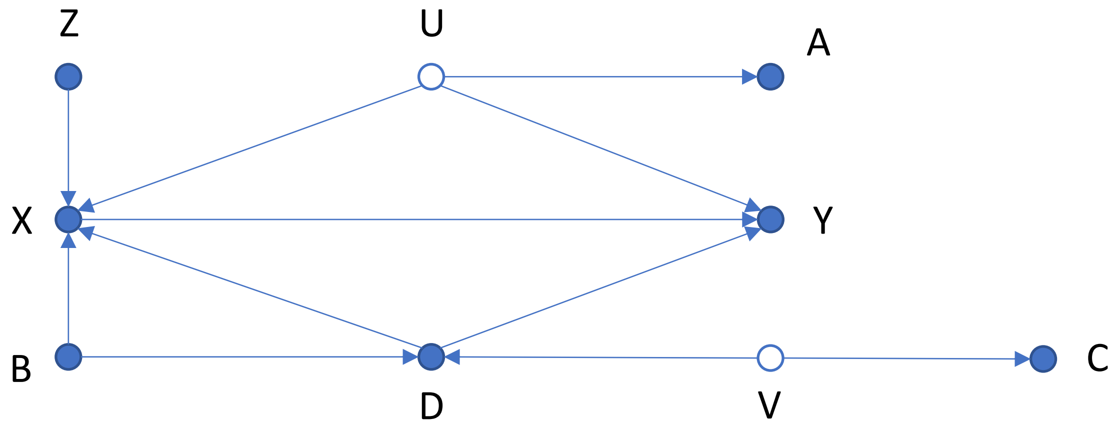
Examples of I.V.

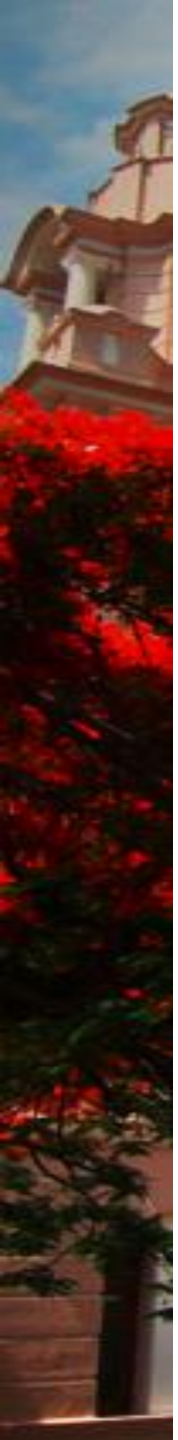
- Using favorable growing condition (e.g., the amount of rainfall) as an instrument for the effect of market price on demand of an agricultural product. (Favorable growing conditions will affect supply of the product but not the demand, and supply will affect the price.)

- Counter-example:
- Using tax rate for tobacco products as an instrument for the world-average effect of smoking on health. (The tax rate for tobacco products is determined mainly by political and economic factors. Economic factors could be related to the general health condition of the population. This IV is also invalid if the government considered general health condition when determining the tax rate.)

In-Class Exercise

- In the following causal system, which variable is a valid instrument for the aggregate causal effect of D on Y?





In-Class Exercise

- We want to investigate how Manning's staffing level affects store sales using OLS. What factors must be included in the regression? If the back-door path cannot be completely blocked, what can be a valid instrument variable?
- If we want to estimate the causal effect of skipping classes on final exam score, what factors should be included in the regression? What can be a valid instrument variable?
- Consider estimating the effect of PC ownership on college GPA for senior students in a large university. what factors should be included in the regression? What can be a valid instrument variable?