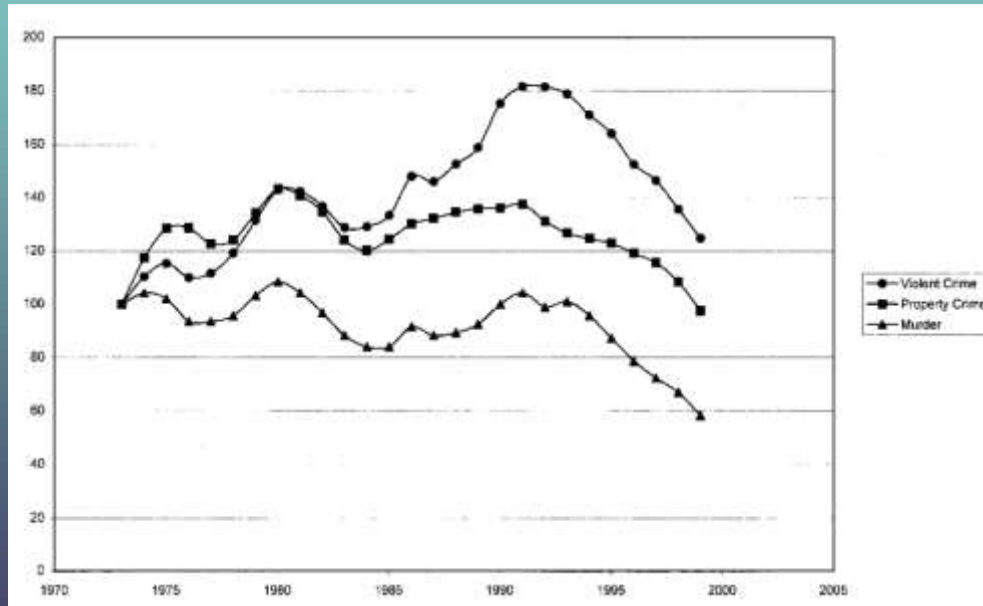# Causality

Identifying causal relationship

# Question

Starting from 1992, there is a sharp decline in the U.S. crime rate.



In your opinion, why?

# Question

The Clinton Administration: Because of the strong economy!

But economy was even stronger in 1980s…

# Question

Police Force: The increasing use of incarceration…

But incarceration started to increase 20 years ago…

# Question

The real reason: Legalized Abortion! ---- Steven Levitt

In 1973, abortion becomes legal in the whole US. Before 1973, many impoverished women became pregnant and had to give birth to their children. When these children grew up in financially disadvantaged situations, many of them later turned to a life of crime. However, this trend underwent a significant shift in 1973.

# Question

Why we see a decline in crime rate in 1992?

Many potential individuals who may have turned to a life of crime were not born as a result of the legalization of abortion in 1973. Had they been born, they would have reached the age of around 18-19 in 1992, which is considered a prime time for criminal activity to begin.

# Quick Review of Regression

Consider simple linear regression: $Y = a + bX$.

$Y$ is your dependent variable and $X$ is your independent variable.

We want to know how $X$ affects the value of $Y$.

# A Motivating Example

Suppose that you want to answer the following question: "Does eating ice-creams increase a person's weight?"

To answer this question, which regression should you run? What data should be collected?

# A Motivating Example

Suppose that you want to answer the following question: "Does eating ice-creams increase a person's weight?"

Here, our dependent variable (Y) is a person' current weight (e.g., Y = 50 means the weight is 50 kg).

Our independent variable (X) is the consumption of ice-cream in the previous week (e.g., X = 4 means you consumed 4 ice-creams in the last week).

# A Motivating Example

We sampled a large number of individuals (say, 1,000 people) and recorded their ice-cream consumption and weight at different time of year, and construct our dataset. Then, we regress Y on X.

# A Motivating Example

Surprisingly, you find that

$$Y = 60 - 4X$$

which suggests that ice-cream consumption *decreases* your weight!

You double and triple checked the data and your regression and cannot find any mistakes. Now, the question is, *what is wrong with our regression*?

# Answer

The regression ignores a key feature: the current season.

During the hot summer, people consume a lot of ice-creams. And also most people lose weight in summer.

During the cold winter, people do not consumer ice-creams. And also most people gain weight in winter.

In other words, in summer both X is large and Y is small, while in winter X is small and Y is large. It seems that X negatively affects Y, but in the end it is the season that makes the difference.

# A Motivating Example

You are paying high tuitions to receive education.  Why are you willing to pay so much money to take courses here?

The reason is that education is an investment. By investing in education, you will be rewarded in the future (e.g., you can make a better salary).

We call this "return to education."

# The Next Problem

A key issue in economics is estimating the return to education.

More specifically, how much more money can you make by taking one extra year's education?

How would you answer this question?

# The Simple Idea

One simple idea is to collect data from different individuals. For example, suppose that we collect data from $N$ individuals. For each one, we know his or her income (which we call $Y$) as well as his or her year of education (which we call $X$).

Then, we can run the regression $Y = a + bX$.

Here, if you take one extra year education, your income will increase by $b$. We are done!

# The Simple Idea

Then, we can run the regression $Y = a + bX$.

Here, if you take one extra year education, your income will increase by $b$. We are done!

Any issues here? Discuss it with your classmates!

# The Issue

Yes, we can run the regression $Y = a + bX$, and find that by taking more education, people make higher salaries.

However, this *does not* necessarily imply education helps you make more money?

But why?

# The Issue

Consider one issue: people have family background.

If your parents are rich, they will send you to the best schools and pay for your education. Then, your education level (X) becomes higher.

At the same time, they can help you find better jobs. In this case, your income (Y) will also be higher.

So, there is a positive relationship between X and Y; but we cannot yet say X has caused Y.

# The Issue

Let us incorporate family background into the regression as well and use $T_1$ to represent the income of a person's parents.

We do the following regression:

$$Y = a + bX + c_1 T_1$$

Would this be fine? Discuss this issue with your classmates.

# The Issue

There may still be many other factors…

For example, gender.

 In many places, parents are more willing to send boys to school, and are reluctant to send girls to school.

 At the same time, gender discrimination leads firms to pay more to males for the same work.

 So, it may be gender that matters.

# The Issue

Even for this simple problem, we have a lot of issues…

Gender, Family Background, IQ level, Geographic Location….

We can try to control for these factors in our regression, but there are so many things to control for and we are never sure if we have controlled for everything…

# The Issue

We call the variable that you miss the "omitted variable", and the issue the "omitted variable bias".

It means that you omit one or more relevant variables in your model specification, and this variable is a determinant of the dependent variable and correlated with one or more of the included independent variables.

Omitted Variable Bias

# Exercise I

Suppose that we are regressing Y (the sales of a smartphone) on X (the price of a smartphone) to see how demand changes with price.

What can be an omitted variable in the above regression?

# Exercise 2

Suppose that we are regressing Y (life expectancy) on X (the air quality) to see how better air makes people live longer.

What can be an omitted variable in the above regression?

# On Fixed Effects

Suppose that we want to investigate the relationship between Grades and Hours of Study.

We know that the grades are affected not only by hours of study, but also by IQ. However, we only observe a student's hours of study but do not observe a student's IQ (you cannot force the students to take an IQ test).

What should we do?

# More on Fixed Effects

Suppose that the actual regression equation is

$$\text{Grades} = a + b_1 \text{IQ} + b_2 \text{Hours}$$

How can we get the value of $b_2$ without observing the IQ of each person?

# More on Fixed Effects

Suppose that the actual regression equation is

$$Grades = a + b_1 IQ + b_2 Hours$$

Consider a student, Alice. Alice takes many classes, denoted by class 1, 2, … J. Then, for class j, we have

$$Grades_{Alice,j} = a + b_1 IQ_{Alice} + b_2 Hours_{Alice,j}$$

# More on Fixed Effects

Consider a student, Alice. Alice takes many classes, denoted by class 1, 2, … J. Then, for class j, we have

$$\text{Grades}_{\text{Alice},j} = a + b_1 \text{IQ}_{\text{Alice}} + b_2 \text{Hours}_{\text{Alice},j}$$

For all classes Alice has taken, we can derive that

$$\overline{\text{Grades}}_{\text{Alice}} = a + b_1 \text{IQ}_{\text{Alice}} + b_2 \overline{\text{Hours}}_{\text{Alice}}$$

# More on Fixed Effects

$$\text{Grades}_{\text{Alice,j}} = a + b_1 \text{IQ}_{\text{Alice}} + b_2 \text{Hours}_{\text{Alice,j}}$$

$$\overline{\text{Grades}}_{\text{Alice}} = a + b_1 \text{IQ}_{\text{Alice}} + b_2 \overline{\text{Hours}}_{\text{Alice}}$$

Using the above two equations, we can derive that

$$\text{Grades}_{\text{Alice,j}} - \overline{\text{Grades}}_{\text{Alice}} = b_2 \left( \text{Hours}_{\text{Alice,j}} - \overline{\text{Hours}}_{\text{Alice}} \right)$$

Note that all parameters in the above equation is known. We are done! Here, we say that we control for the "individual fixed effect".

# More on Fixed Effects

Let us make the problem more realistic. Again, we want to evaluate the effect of hours of study on the final grades. However, your final grades are also affected by

Your IQ, which is not observed.

The hardness of the exam, which is not observed either. Some classes have easy exams while other classes have difficult exams. What should you do then?

# More on Fixed Effects

$$\text{Grades}_{i,j} = a + b_1 \text{IQ}_i + b_2 \text{Easiness}_j + b_3 \text{Hours}_{i,j}$$

$$\left(\text{Grades}_{A,1} - \text{Grades}_{A,2}\right) = b_2\left(\text{Easiness}_1 - \text{Easiness}_2\right) + b_3\left(\text{Hours}_{A,1} - \text{Hours}_{A,2}\right)$$

$$\left(\text{Grades}_{B,1} - \text{Grades}_{B,2}\right) = b_2\left(\text{Easiness}_1 - \text{Easiness}_2\right) + b_3\left(\text{Hours}_{B,1} - \text{Hours}_{B,2}\right)$$

Then, we come up with

$$\left(\text{Grades}_{A,1} - \text{Grades}_{A,2}\right) - \left(\text{Grades}_{B,1} - \text{Grades}_{B,2}\right)$$
$$= b_3\left(\left(\text{Hours}_{A,1} - \text{Hours}_{A,2}\right) - \left(\text{Hours}_{B,1} - \text{Hours}_{B,2}\right)\right)$$

# More on Fixed Effects

Common types of fixed effects:

Year / Month fixed effects: 2021 and 2020 may be different.

Weekday/ Weekends fixed effects: Monday may be different from Sunday.

Individual fixed effects: Alice is different from Bob.

Location fixed effects: China may be different from Japan.

# A Crowdfunding Example

Suppose that you want to investigate how the target of a crowdfunding campaign affects the total funding raised. Which types of fixed effects may come into play?

Individual fixed effects --- some people are more successful than others.

Location fixed effects --- The state can make a difference.

Subtype fixed effects --- Smartwatches may be different from software.

# A Crowdfunding Example

Question: In the Kickstarter dataset, is it possible for you to control for these fixed effects? Why?

Individual fixed effects --- some people are more successful than others.

Location fixed effects --- The state can make a difference.

Subtype fixed effects --- Smartwatches may be different from software.

# A Crowdfunding Example

Question: In the Kickstarter dataset, is it possible for you to control for these fixed effects? Why?

Individual fixed effects --- some people are more successful than others.

To be able to control for the individual fixed effect, we must have at least two observations for each entrepreneur, so we can calculate the difference. However, in the crowdfunding dataset, each entrepreneur typically only has one project.

# A Crowdfunding Example

Now consider the following regression:

Here, we do not control for any fixed effects.

|              | Estimate | Std. Error | t value | Pr(>\|t\|) |     |
|--------------|----------|------------|---------|-----------|-----|
| (Intercept)  | 3.43693  | 0.27031    | 12.71   | <2e-16    | *** |
| LogTarget    | 0.30421  | 0.02726    | 11.16   | <2e-16    | *** |

# A Crowdfunding Example

Now consider the following regression:

Here, we control for fixed effects.

```
result = lm(LogFunding ~ LogTarget +
factor(Subtype) + factor(Location), data = mydata)
summary(result)
```

# A Crowdfunding Example

Now consider the following regression: Here, we control for fixed effects.

```
Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                         6.66186    0.40466  16.463  < 2e-16 ***
LogTarget                           0.20560    0.02403   8.556  < 2e-16 ***
factor(Subtype)Apps                -4.19174    0.33504 -12.511  < 2e-16 ***
factor(Subtype)CameraEquipment      1.30956    0.45434   2.882 0.003960 **
factor(Subtype)DIYElectronics      -0.20024    0.40840  -0.490 0.623940
factor(Subtype)FabricationTools    -2.53178    0.56363  -4.492 7.17e-06 ***
factor(Subtype)Flight              -1.50850    0.47699  -3.163 0.001571 **
factor(Subtype)Gadgets             -0.33102    0.34636  -0.956 0.339246
factor(Subtype)Hardware             0.18633    0.33815   0.551 0.581627
factor(Subtype)Makerspaces         -1.07444    0.56385  -1.906 0.056752 .
factor(Subtype)Robots              -0.10786    0.43235  -0.249 0.803008
factor(Subtype)Software            -3.07865    0.34560  -8.908  < 2e-16 ***
factor(Subtype)Sound                0.05697    0.42930   0.133 0.894436
factor(Subtype)SpaceExploration    -0.88965    0.50991  -1.745 0.081076 .
factor(Subtype)Technology          -1.70381    0.33201  -5.132 2.95e-07 ***
factor(Subtype)Wearables            0.19453    0.38017   0.512 0.608882
factor(Subtype)Web                 -4.43896    0.34505 -12.865  < 2e-16 ***
factor(Location)IL                 -0.96675    0.16151  -5.986 2.26e-09 ***
factor(Location)MA                  0.07324    0.15776   0.464 0.642465
factor(Location)NY                 -0.36623    0.11099  -3.300 0.000973 ***
factor(Location)TX                 -1.31772    0.11932 -11.043  < 2e-16 ***
factor(Location)WA                 -0.63908    0.16752  -3.815 0.000137 ***
```

# Difference-in-Difference Analysis

David Card
The Sveriges Riksbank Prize in Economic Sciences in
Memory of Alfred Nobel 2021

Born: 1956, Guelph, Canada

Affiliation at the time of the award: University of California,
Berkeley, CA, USA

Prize motivation: "for his empirical contributions to labour
economics."

Prize share: 1/2

# Diff-in-Diff

David Card and Alan Krueger wanted to investigate the effect of minimum wage on unemployment rate. They consider two states in the US: New Jersey (NJ) and Pennsylvania (PA).

| Minimum Wage | March 1992 | December 1992 |
|---|---|---|
| New Jersey | 4.05 | 5.05 |
| Pennsylvania | 4.05 | 4.05 |

# Diff-in-Diff

| Minimum Wage | March 1992 | December 1992 |
|---|---|---|
| New Jersey | 4.05 | 5.05 |
| Pennsylvania | 4.05 | 4.05 |

| Unemployment Rate | March 1992 | December 1992 |
|---|---|---|
| New Jersey | A | B |
| Pennsylvania | C | D |

# Diff-in-Diff

Note that unemployment rates are determined not only by the minimum wage, but also by the following factors:

The location: New Jersey may have better economy and lower unemployment rate

The time: The economy may be stronger in December 1992 so that unemployment rate is lower then.

# Diff-in-Diff

Economics Model:

$$\text{Unemployment}_{it} = a + \text{State}_i + \text{Month}_t + b\text{Wage}_{it}$$

$$A = a + NJ + Mar + b \times 4.05$$
$$B = a + NJ + Dec + b \times 5.05$$
$$C = a + PA + Mar + b \times 4.05$$
$$D = a + PA + Dec + b \times 4.05$$

# Diff-in-Diff

$$A = a + NJ + Mar + b \times 4.05$$
$$B = a + NJ + Dec + b \times 5.05$$
$$C = a + PA + Mar + b \times 4.05$$
$$D = a + PA + Dec + b \times 4.05$$

$$B - A = (Dec - Mar) + b \times 1.0$$
$$D - C = (Dec - Mar)$$

$$(B - A) - (D - C) = b \times 1.0$$

The Economist

Does minimum wage hurt the economy?

# Example

We all know that smoking is bad for health, and quitting smoking is the best thing that a smoker can do.

# Example

As data scientists, we want to estimate the effect of quit smoking. For instance, we want to answer the following question: Can a lifetime smoker reduce the chance of developing lung cancer by quitting smoking?

What would you do to answer this question?

# Example

The simplest idea: Collect data from a large number of lifetime smokers. Let Y be whether or not a smoker develops lung cancer ($Y = 1$ when developing cancer and $Y = 0$ when no cancer).

Let X denote whether or not the smoker has quitted smoking. That is, $X = 1$ if he or she quitted smoking and $X = 0$ if he or she did not quit smoking.

# Example

Then, we can run the following regression:

$$\Pr[Y = 1] = \text{Logit}(a + bX + c_1 X_1 + c_2 X_2 \dots)$$

$X_1, X_2 \dots$ are other control variables such as gender, income, location, exercise habits, occupation,….

We find that $b$ is positive and significant. What does it mean? Discuss it with your classmates!

# Example

Surprisingly, when data scientists first ran this regression, they found out that $b > 0$, that is, quitting smoking *increases* the change of developing lung cancer!

Ironically, this result suggests that we should encourage smokers to continue smoking....

# Example

This result, of course, contradicts with our knowledge…
Then, the question is,

What is wrong here?

Discuss this with your classmates!

# Example

This is a long story…

We know that quitting smoking is very difficult for most smokers. Usually, people start considering quitting smoking when they are suffering from certain disease…

In other words, many smokers quit smoking *after* they know that they have lung cancer!

# Example

Now the answer is clear…

Y (cancer) and X (quit smoking) are positively correlated because Y causes X, not X causes Y.

This is a common issue called "reverse causality."

# Example

Consider another example. In US, people want to answer the following question: "Does the police reduce the crime rate?"

Here, our dependent variable Y is the crime rate, and our independent variable X is the size of the police force. By running regression, we find that Y increases with X.

We should defund the police!

# Example

Consider another example. In US, people want to answer the following question: "Does the police reduce the crime rate?"

The real reason is, when there is more crime, the city tends to hire more police officers. It is Y that causes X, not X causes Y!

# Reverse Causality

In many cases X and Y can affect each other.

In other words, X causes Y and Y causes X.

For example, let X be police force and Y be crime rate.

Y affects X: when there is more crime, there will be more police.

X affects Y: when there is more police, there is less crime.

# DISCUSSION

Try to find other common examples of reverse causality.

# OUR TASK

---

We want to know how the change of X affects the value of Y. We want to exlude other factors and reserved causality.

# Two Common Causes

Omitted Variable

Reserve Causality

# Two Common Solutions

Experiments

Instrumental Variable

# Hypothetical Example

Which of the following schools should you join if you want to make more money in the future?
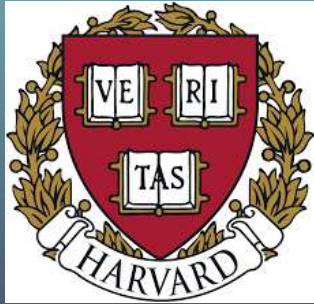
# Hypothetical Example

Which of the following schools can help you make more money in the future?

Suppose that you analyze the data and find that the average salary for the graduates of the three schools are (HKD):

Harvard:        1M
HKU:            500K
Lanxiang (蓝翔):    200K

Can we say that Harvard is the best school that helps you make more money?

# Hypothetical Example

Of course not.

Reason: Harvard/HKU is a famous school, and it has very admission standard. You must have a high IQ to be admitted to Harvard/HKU.

It is hard to say whether Harvard/HKU graduates are making more money because of their Harvard degree or because of their high IQ.

# Hypothetical Example

We are more interested in the following question:

Holding all other things equal (e.g., IQ, family background, education background, gender, …) would a person make a higher salary if he/she chooses Harvard instead of HKU or Lanxiang?

# American Business Schools Sorted by Average Starting Salary and Bonus

| School | Average Starting Salary and Bonus | Percent Employed at Graduation | Average GMAT Score (full-time) | Acceptance Rate (full-time) |
|---|---|---|---|---|
| University of Pennsylvania (Wharton) | $159,815 | 82.3% | 730 | 19.2% |
| Stanford Univeristy | $159,440 | 63.9% | 737 | 5.7% |
| Harvard University | $158,049 | 78.9% | 731 | 9.9% |
| University of Virginia (Darden) | $153,576 | 83.4% | 713 | 24.5% |
| Dartmouth College (Tuck) | $152,805 | 80.2% | 722 | 23.0% |
| Cornell University (Johnson) | $152,207 | 80.3% | 700 | 29.9% |
| Columbia University | $151,849 | 69.9% | 727 | 14.0% |
| University of Chicago (Booth) | $151,085 | 88.0% | 730 | 23.5% |
| University of Michigan, Ann Arbor (Ross) | $150,052 | 89.7% | 716 | 25.3% |

# Hypothetical Example

Idea:

1. We recruit a large number of high school students.

2. We randomly send the students to the three schools, Harvard, HKU and Lanxiang.

3. After they graduate, we compare which school's graduates make a higher salary.

# Hypothetical Example

Note that now, students are randomly allocated to the three schools. In this case, which school you enter is independent of your personal information: The school you enter has nothing to do with your IQ, your family background, …

Harvard will get both high IQ and low IQ students. Similarly, Lanxiang will also get both high IQ and low IQ students. Their students' backgrounds are comparable.

Then, if their graduates make different salaries, it is the school that makes the difference and we can rule out all other factors!

# One Solution

Of course, this example is more or less impractical…

    You cannot force Harvard to admit a student, or force a student to join Lanxiang.

    This type of experiment is also illegal…

But the idea is great! We can use it elsewhere.

# One Solution

Similarly, we can do this for the smoking problem.

Recall that we want to know if quitting smoking reduces the chance of developing cancer.

# One Solution

First, we recruit many smokers to do the experiment.

Second, we randomly assign them into two groups: a control group that continues smoking, and a treatment group.

Individuals in the treatment group are forced to quit smoking.

Later, we examine whether each individual develops cancer in each group.

# Pharmaceutical

In the pharmaceutical industry, researchers need to prove that a new drug (X) is effective in curing or alleviating a certain disease (Y). To establish a credible causal relationship, they use "randomized controlled trials (RCTs)"

# Pharmaceutical

- We randomly assign participants into two groups, a treatment group and a control group.

- The treatment group takes the medication ($X=1$) while the control group takes nothing ($X=0$).

- Finally, we calculate the value of Y (seriousness of disease) for each participant in the study.

# Pharmaceutical

However, this alone is *not* enough!

Why? In addition to the medical effect, there may be a placebo effect that comes into play.

When you take the real medication, you feel better, and your mood is good. However, mood can also affect Y.

We do no really know if the medication works, or patient gets a better mood.

# Pharmaceutical

- So, in addition to random trials, we also make the trials "blind": you do not know which group (i.e., treatment versus control group) that you belong to.

- In practice, we offer a fake medication (placebo) which resembles the real medication to individuals in the control group. So, everyone gets a pill, and you don't know whether or not you are getting the real medication.

# AB Testing

The most commonly used
experimental method

"Whenever we launch a new APP, we run many AB tests in different platforms to find a best name for it. Even though you are 99.9% sure one name is the best, it does not hurt running an AB test."
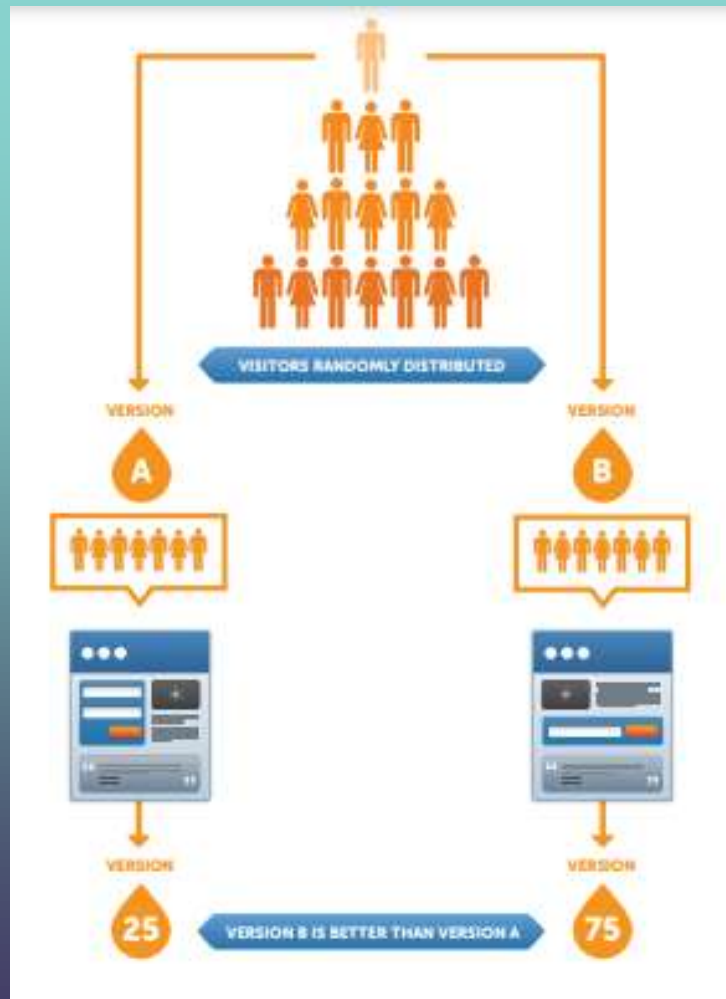
---Zhang Yiming, ByteDance CEO
(字节跳动张一鸣)

- AB testing was inspired by RCT used in pharmaceutical industry; it is commonly used in the Internet age.

- It is commonly used for Website and APPs.

  - Which website design is better?
  - Which APP design is better?
  - Which image is more attractive?

# Experiment

Unfortunately, we are not always able to run our experiments or AB testing. For example, experiments can be illegal, too costly, too time consuming, or simply impossible.

For example, suppose that you want to understand how smoking affect one's health; however, you cannot run an AB test forcing some people to smoke/not to smoke. So, how to obtain the results?

# Two Common Solutions

Experiments

Instrumental Variable

# Course Overview

2021 Nobel Prize in Economics



Joshua D. Angrist and Guido W. Imbens

"for their methodological contributions to the analysis of causal relationships"

# Instrumental Variable

When running an experiment is impossible, we may also consider the instrumental variable approach.

Idea: Find a new variable that affects your X but does not affect your Y directly.

# Theory (Optional)

Suppose that you want to estimate how X affects the value of Y. Mathematically, suppose that when X increases by 1, Y will increase by b. We want to find out the value of b.

You find a variable Z that affects X but does not affect Y directly.

People have proved that

$$b = \frac{Cov(Y,Z)}{Cov(X,Z)}$$

where cov stands for covariance.

# Example

Let us go back to the smoking example.

Here, we want to find out an instrumental variable $Z$ such that (1) when $Z$ changes, people's smoking behavior changes; (2) $Z$ does not affect people's health directly.

Any ideas?

# Example

One plausible instrumental variable is tobacco tax, which is changing over time. It has the following two features:

(1)  A tobacco tax affects the amount that one smokes.
(2)  Tobacco tax does not affect one's health directly. It only affects one's health through how people smoke.

# A FACT

---

Someone may be telling you that he/she is an economist. However, what he/she does everyday is simply looking for instrumental variables.

# Discussion

Recall that earlier in this class, we want to understand how education affects one's income. However, we cannot easily run an experiment on this.

So, we may consider finding an instrument. Here, the instrument should (1) affect one's year of education but (2) does not affect one's income directly.

Any ideas?

# Discussion

This is really a famous problem! And there is also a famous instrument – the month of birth.

In the US, children have to stay in school until a certain age (e.g., 18 years old). For example, if you are born in Jan 2000, you can leave school in Jan 2018, and so on.

# Discussion

However, everyone joins school in September.

Then, two persons born in Jan 2000 and June 2000 join school at the same time but can leave school at different times (Jan 2018 vs. June 2018). So the latter individual takes more education than the former one does.

In this case, your month of birth affects your year of education.

# Discussion

While your month of birth affects your year of education, it does not affect your income through other channels.

In this way, your month of birth can be an instrumental variable.

# One of the most classic research!

## DOES COMPULSORY SCHOOL ATTENDANCE AFFECT SCHOOLING AND EARNINGS?*

### JOSHUA D. ANGRIST AND ALAN B. KRUEGER

We establish that season of birth is related to educational attainment because of school start age policy and compulsory school attendance laws. Individuals born in the beginning of the year start school at an older age, and can therefore drop out after completing less schooling than individuals born near the end of the year. Roughly 25 percent of potential dropouts remain in school because of compulsory schooling laws. We estimate the impact of compulsory schooling on earnings by using quarter of birth as an instrument for education. The instrumental variables estimate of the return to education is close to the ordinary least squares estimate, suggesting that there is little bias in conventional estimates.

# One of the most classic research!



The association between education and income

There is a clear correlation between years of education and income. But does this mean there is also causation?

12%

65%

Income

Years of education
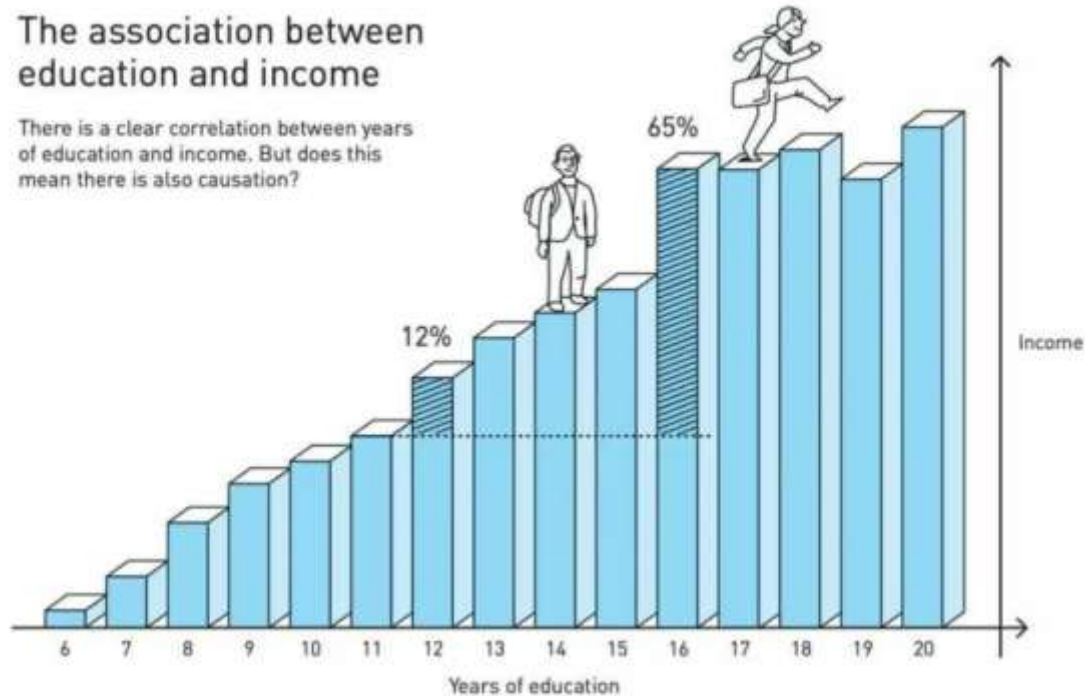
# Any Other Instrumental Variable?



Instrumental Variables In Action: Education & Wages (graphs)
CAUSAL INFERENCE BOOTCAMP

# THE ELITE ILLUSION: ACHIEVEMENT EFFECTS AT BOSTON AND NEW YORK EXAM SCHOOLS

BY ATILA ABDULKADIROĞLU, JOSHUA ANGRIST, AND PARAG PATHAK[1]

Parents gauge school quality in part by the level of student achievement and a school's racial and socioeconomic mix. The importance of school characteristics in the housing market can be seen in the jump in house prices at school district boundaries where peer characteristics change. The question of whether schools with more attractive peers are really better in a value-added sense remains open, however. This paper uses a fuzzy regression-discontinuity design to evaluate the causal effects of peer characteristics. Our design exploits admissions cutoffs at Boston and New York City's heavily over-subscribed exam schools. Successful applicants near admissions cutoffs for the least selective of these schools move from schools with scores near the bottom of the state SAT score distribution to schools with scores near the median. Successful applicants near admissions cutoffs for the most selective of these schools move from above-average schools to schools with students whose scores fall in the extreme upper tail. Exam school students can also expect to study with fewer nonwhite classmates than unsuccessful applicants. Our estimates suggest that the marked changes in peer characteristics at exam school admissions cutoffs have little causal effect on test scores or college quality.

# Comment on Machine Learning

Nowadays, machine learning is very powerful in making predictions. That is, given $X$, machine learning algorithms predict the value of $Y$.

Many algorithms such as neural networks, naïve Bayesian, matrix factorization etc.

Basically, machine learning algorithms generate a function $f$ such that $Y \approx f(X)$.

# Comment on Machine Learning

There is a frequent challenge on machine learning:

*"Machine Learning-based projects focus on predicting outcomes rather than understanding causality."*

For this reason, machine learning algorithms are often called "black box" --- we know it works, but we don't know how it works (只知其然，不知其所以然).

# Why should we care about causality?

In an e-commerce context, we could determine which specific factor impacts the decision to purchase a product. With this information, we could better allocate resources to improve a specific KPI. We could also rank the impact of different factors on the purchasing decision. We could determine if a given customer would have purchased a specific product if he/she had not bought other products for the last two years.

# Why should we care about causality?

In the agricultural field, we often try to predict if a farmer's crop yield will be lower this year. However, using casual inference, it will become to better understand what steps should we take to increase the harvest.