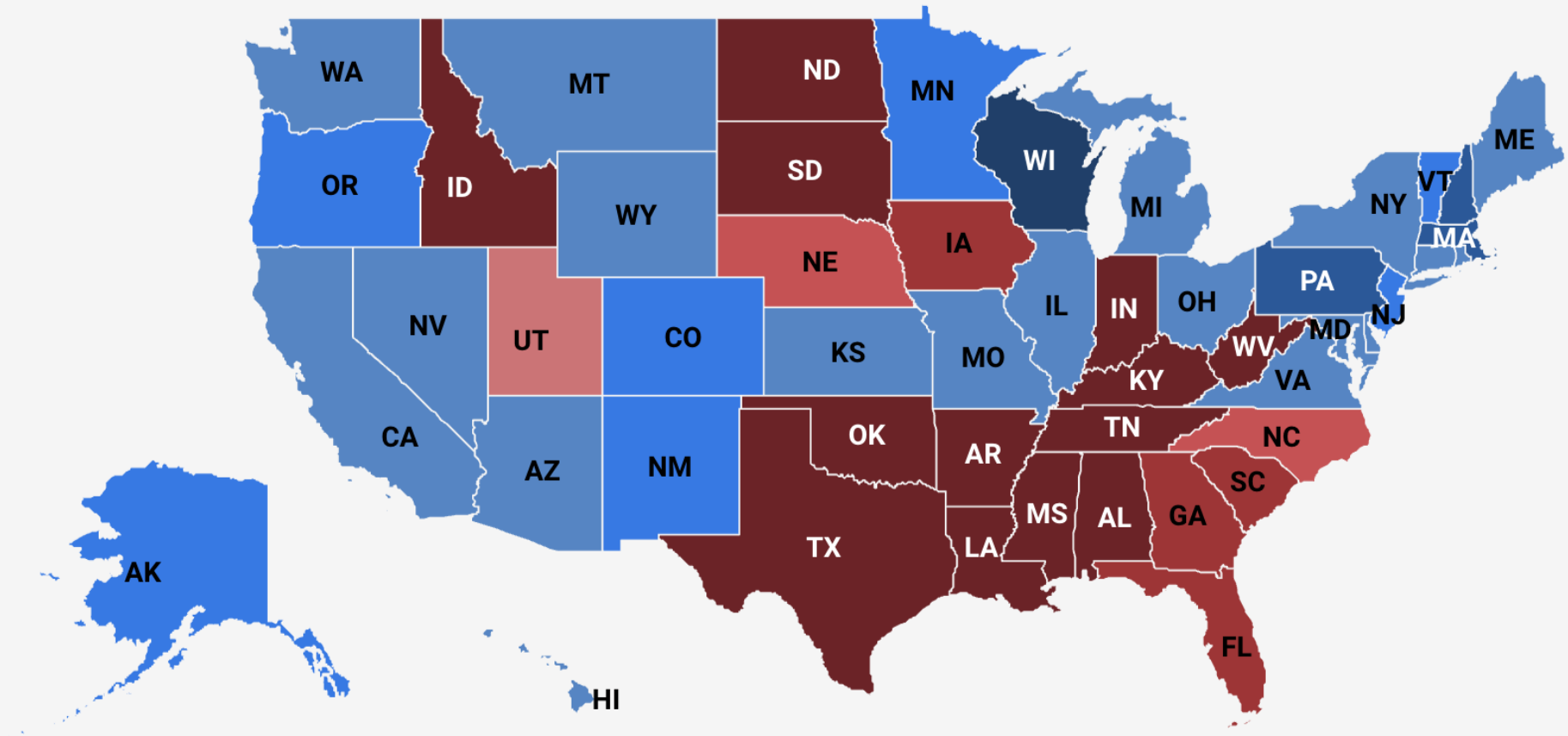
Causality


WA
OR
CA
ID
NV
MT
WY
UT
AZ
CO
NM
ND
SD
NE
KS
OK
TX
MN
IA
MO
AR
LA
WI
IL
MS
MI
IN
KY
TN
AL
OH
WV
GA
FL
SC
NC
VA
PA
NY
MD
NJ
VT
ME
MA
AK
HI

A Puzzle

Through data analysis, a high school teacher finds that students who use blue pens score lower compared to their classmates who use black pens. He concludes that blue color reduces students' performance.

Can you explain this puzzling finding? Why is this so?

You're paying high tuition to receive an education. Why are you willing to spend so much money to take courses here?

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, Y_i, as well as his or her years of education, X_i.

Next, we run a linear regression $Y_i = a + bX_i$: one extra year's education can increase your salary by b.

What's wrong with the above approach?

There are so many other variables that influence both your education and salary. Here is an incomplete list:

- IQ
- Gender
- Health condition
- Geographic location

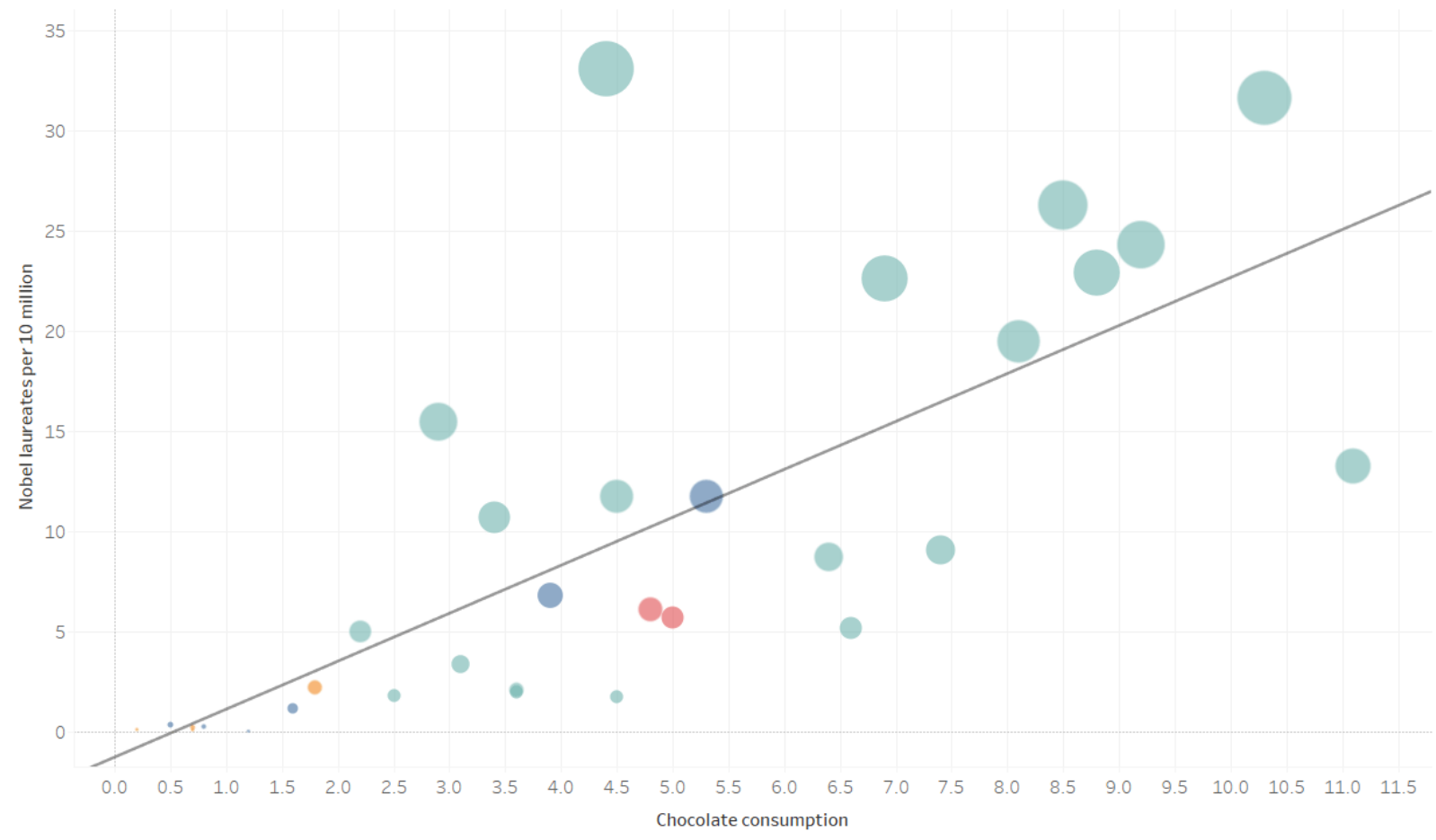
And you cannot include everything into your regression. On the one hand, there may be 100 factors in the list; on the other hand, even though you know IQ affects both X and Y, it is almost impossible for you to know the IQ of each individual.

We call the variable(s) that you ignore 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.

https://www.youtube.com/embed/b4jhrK03zhs?enablejsapi=1

Exercise: What can be an omitted variable?



Suppose that we want to know how hours of study affects your grades in the final exam. Here, IQ is an issue:

- If you have a high IQ, you will study less.
- If you have a high IQ, you will do better in the exam.

Ideally, we want to run the following regression:

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

But we do not know the IQ of anyone. What can we do to figure out the value of b_1 without knowing IQ?

In statistics, this is equivalent to taking fixed effects! In the first example, we take the "individual fixed effect," and in the second example, we take both "individual fixed effect" and "exam fixed effect."

The above approach is also known as the "difference-in-difference" or simply the "DID" approach.

https://www.youtube.com/embed/8H4yp8Fbi-Y?enablejsapi=1

A study shows that cat owners are more likely to be single. Does it mean cats scare away partners?

As discussed above, two issues make it difficult for us to figure out causal relationships:

(1) omitted variable bias and (2) reversed causality.

We propose two ways to fix the issue:

(1) running experiments and (2) using instrumental variables.

Experiments (AB Test)

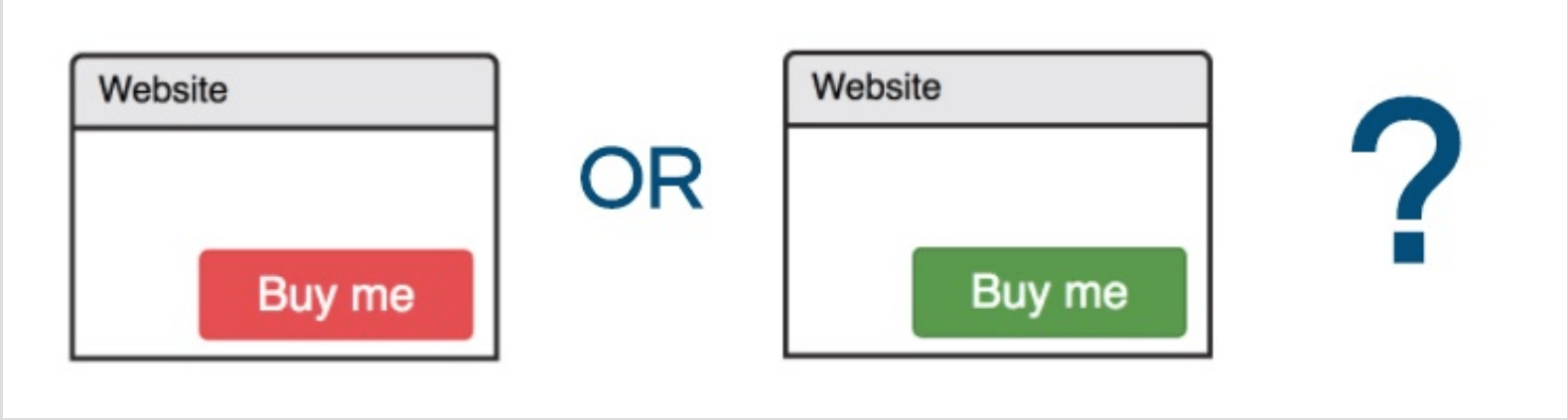
A hypothetical example

You are admitted to two universities, HKU (QS #11) and Columbia (QS #38). You want to make more money in the future. Which one should you join?

We cannot state that Columbia > HKU. There are many reasons:

- Columbia may attract more talented students.
- Columbia students may come from richer families.

AB Tests: Examples



Testing the color of your Call-to-Action button.

Which one is better?

How do you chat with AI?

Very Polite / Polite / Neutral / Rude / Very Rude

Examples of very polite tones:

- Can you kindly consider the following problem and provide your answer.
- Can I request your assistance with this question.
- Would you be so kind as to solve the following question?

Examples of polite tones:

- Please answer the following question:
- Could you please solve this problem:

Examples of rude tones:

- If you're not completely clueless, answer this:
- I doubt you can even solve this. Try to focus and try to answer this question:

Examples of very rude tones:

- You poor creature, do you even know how to solve this?
- Hey gofer (马仔), figure this out.
- I know you are not smart, but try this.

Guess what happened?

Free frame

* Limited offer *

Only 100 buyers get a bottle of tomato pasta sauce for free.

Get this for free !!!

$0 frame

* Limited offer *

Only 100 buyers get a bottle of tomato pasta sauce for $0.

Get this for $0 !!!

Free frame

$2 gift certificate for Starbucks

List price $2.00
Discount -$2.00
FREE

Add to Cart

$0 frame

$2 gift certificate for Starbucks

List price $2.00
Discount -$2.00
$0.00

Add to Cart

$0 vs. Free: Which one is better?

Which display performs best?

Analyzing Data from AB Tests

Suppose that we want to test the effectiveness of two banner ads:

A: **Enjoy 15% for your car insurance!**

B: **Last-minute deals for your car insurance!**

Our outcome is whether a user clicks through with ad A versus ad B. How do we tell if one ad is more effective than the other?

Suppose that:

45 out of 856 [5.25%] users clicked through on ad A;
71 out of 1,298 [5.47%] users clicked through on ad B.

Can you say that ad B is more effective than ad A?

Suppose that:

45 out of 856 [5.25%] users clicked through on ad A;
71 out of 1,298 [5.47%] users clicked through on ad B.

Can you say that ad B is more effective than ad A?

Perhaps not. Ad B may be just lucky enough to have a few more accidental clicks.

Running χ^2 Tests with LLM

Suppose that we want to test the effectiveness of two banner ads. We run an AB test. Out of 856 consumers who saw ad A, 45 clicked through. Out of 1289 consumers who saw ad B, 71 clicked through. Run a Chi-squared test to compare the two ads.

There is no significant difference between Ad A and Ad B. The click-through rates are statistically indistinguishable ($p \approx 0.80$).

Suppose that:

489 out of 2,000 [24.45%] users clicked through on ad A;
585 out of 2,000 [29.25%] users clicked through on ad B.

Can you say that ad B is more effective than ad A?

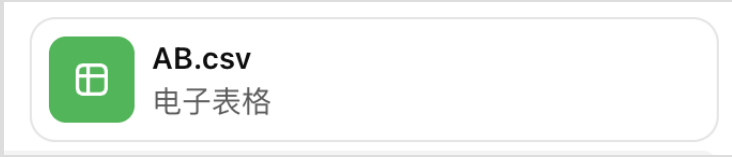
Running χ^2 Tests with LLM

Suppose that we want to test the effectiveness of two banner ads. We run an AB test. Among consumers who saw ad A, 489 clicked through and 1,511 did not click. Among consumers who saw ad B, 585 clicked through and 1,415 did not click. Run a Chi-squared test to compare the two ads.

The chi-square test shows a statistically significant difference between Ad A and Ad B ($p < 0.001$). Ad B performs significantly better than Ad A.

Compare revenues (Data)

This is data from an AB test. Run a t-test to compare the revenues of group A and group B.



The p-value is extremely small, far below any common significance threshold. This means revenue in group A and group B is statistically significantly different. The negative t-statistic indicates that group A has lower mean revenue than group B.

Question: What is the difference between the χ-squared test and the t-test?

Question: What is the difference between the χ-squared test and the t-test?

t-test is used to compare the means of two continuous variables. χ-squared test, by contrast, demonstrates whether there is an association between two categorical variables.

https://www.youtube.com/embed/7_lZpPO-Vxg?enablejsapi=1

Question

> *The dangers of a slow web site: frustrated users, negative brand perception, increased operating expenses, and loss of revenue.*

——Steve Souders

Question

Of course, faster is better, but how important is it to improve performance by 0.1 second? Should you have a person focused on performance? Maybe a team of five? The return-on-investment (ROI) of such efforts can be quantified by running a simple experiment.

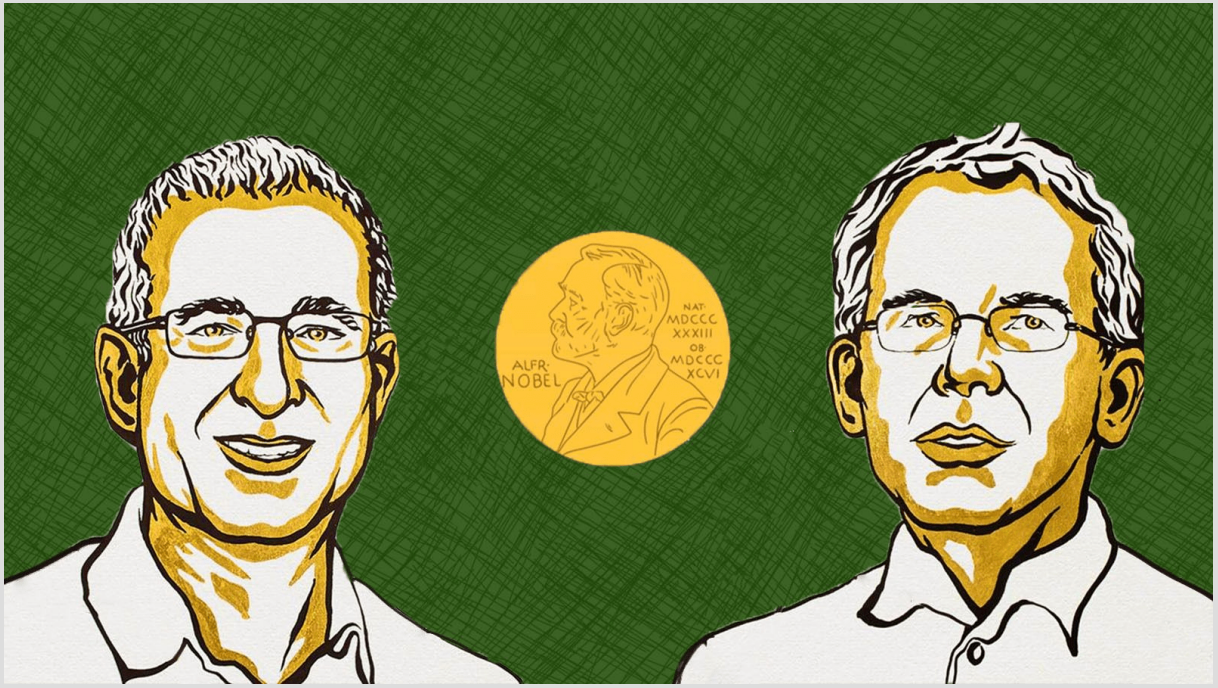
The problem is, we cannot easily increase the speed. It costs millions of dollars. What should we do then?

Question

You want to study the effect of Uber driver supply on the consumer demand. You want to change the number of Uber drivers to see how the number of orders change. In some (randomly assigned) conditions you have more drivers and in some (randomly assigned) conditions you have fewer drivers.

But you cannot force drivers to work in certain hours. What could you do in this case?

Instrumental Variable



In 2021, Joshua Angrist (MIT) and Guido Imbens (Stanford) won the Nobel Prize in economics “for their methodological contributions to the analysis of causal relationships.”

Instrumental Variables

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 through any other channel.

Instrumental Variables

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. Statisticians have proved that

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

Instrumental Variables

Let's consider the coffee example. We want to show whether coffee intake can reduce anxiety.

A valid instrumental variable should (1) affect a person's coffee intake but (2) is not related to a person's anxiety level through any other channels.

Do you have any idea?

Instrumental Variables

Next, we want to examine how education affects one's income. However, we cannot easily run an experiment.

So, we may consider finding an instrument. Here, the instrument should (1) affect one's year of education but (2) is not related to one's income through any other channels.

Any ideas?

https://www.youtube.com/embed/vacBsxBgFMY?enablejsapi=1

Please click the following link to do a peer evaluation of the group AI project.

- You will evaluate 6 randomly chosen stories
- Skip the story if it is created by your own team
- Skip the story if you cannot understand the language