Causality

Back to Linear Regression

Suppose that you want to know whether eating ice creams increases one's weight. We collect the following data:

- An individual's ice cream consumption (e.g., in past month), denoted by X_i
- An individual's current weight, denoted by Y_i
- And you specify the linear regression $Y_i = a + bX_i$.

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

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

Exercise

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

What can be an omitted variable in the above regression?

Exercise

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?

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

In a study examing the relationship between coffee intake and the feeling an anxiety, scientists find that, as people take more coffee, they also feel more anxious. Does this mean that coffee has the side effect of causing anxiety?

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!

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

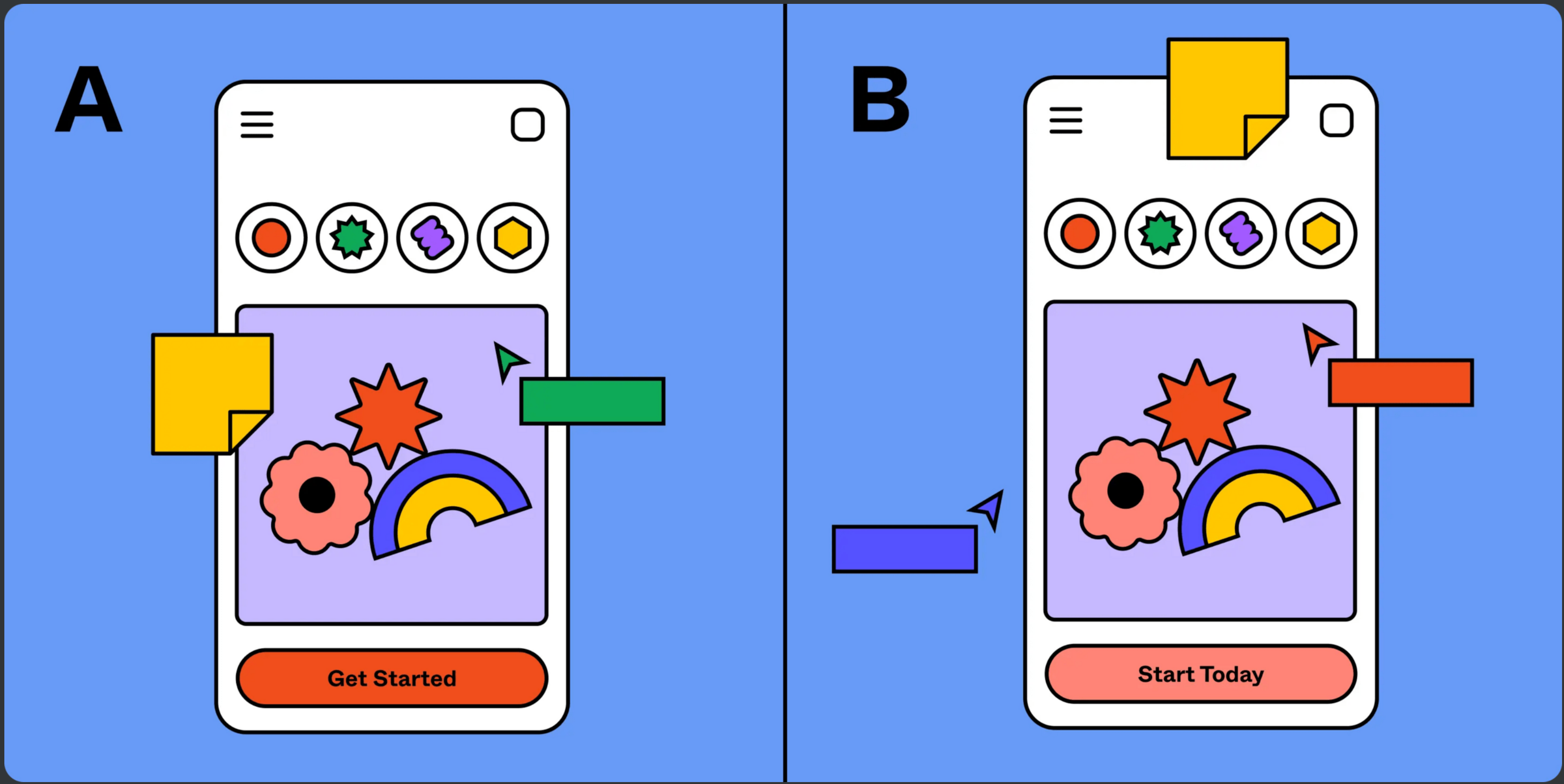
Consider three schools: Harvard, HKU, and Lanxing (蓝翔).
Which school helps you make most money?

Harvard	HKU	Lanxiang
$1 Million	$500 K	$200 K

This is the basic idea of AB testing.

When we want to compare two (or more) conditions to see which one works better, we can randomly assign participants into two (or more) groups, namely group A and group B. Since there are no other differences between the two groups, any difference in the outcome is driven by the difference in the conditions.

The key for successful A/B testing is random assignment. You must make sure that people in group A and group B are similar enough, ruling out other potential causes of the effects. AB testing is the gold standard for finding causal relationship. It is commonly adopted by big tech firms.



Why randomization is so important?

Consider an example in which you assign male MBAs to Harvard and female MBAs to HKU. If one university performs better, you don't know whether this is caused by gender difference or by difference in the schools.

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?

The χ-Squared Test

```r
library(readr)
mydata = read_csv("https://ximarketing.github.io/data/AB.csv")
head(mydata)
table(mydata$treated, mydata$CTR)
```

Treated: Which ad consumers are exposed to.

	No	Yes
A	1511	489
B	1415	585

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.

It seems that ad B is more effective than ad A.

The χ-Squared Test

```r
chisq.test(mydata$treated, mydata$CTR)
```

```
> chisq.test(mydata$treated, mydata$CTR)

	Pearson's Chi-squared test with Yates' continuity correction

data:  mydata$treated and mydata$CTR
X-squared = 11.488, df = 1, p-value = 0.0007006
```

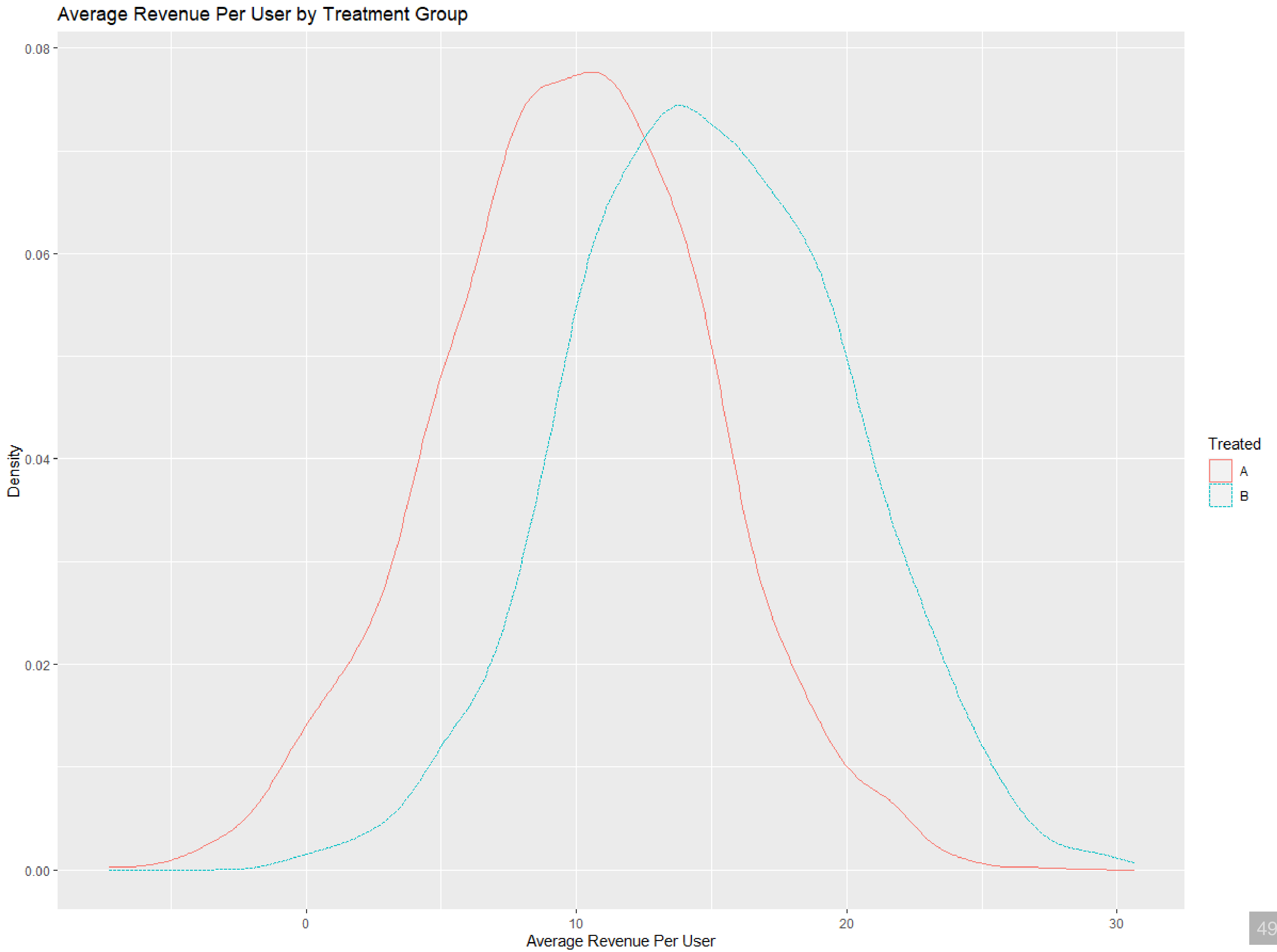
Here, we focus on the p-value. Typically, when $p < 0.05$, we claim the two conditions lead to significantly different outcomes; and in our case, $p < 0.001$, meaning that ad B is more effective than ad A.

The complete code is here.

```
library(readr)
mydata = read_csv("https://ximarketing.github.io/data/AB.csv")
head(mydata)
table(mydata$treated, mydata$CTR)
chisq.test(mydata$treated, mydata$CTR)
```

While we use χ-Square test to compare the click-through rates in the two groups, we now use t-test to compare the revenue per users in the two groups.

```
library(ggplot2)
ggplot(mydata,aes(x=revenue, color =treated))+
  geom_density(aes(linetype=treated))+
  labs(title="Average Revenue Per User by Treatment Group",
       x="Average Revenue Per User",
       y="Density", color ="Treated", linetype ="Treated")
        +theme(plot.title=elementtext(hjust=0.5))
```


Average Revenue Per User by Treatment Group
0.08
0.06
0.04
0.02
0.00
Density
0
10
20
30
Average Revenue Per User
Treated
A
B

While we use χ-Square test to compare the click-through rates in the two groups, we now use t-test to compare the revenue per users in the two groups.

```r
groupA = subset(mydata, treated == "A")
groupB = subset(mydata, treated == "B")
t.test(groupA$revenue, groupB$revenue)
```

```
> t.test(groupA$revenue, groupB$revenue)

        Welch Two Sample t-test

data:  groupA$revenue and groupB$revenue
t = -31.741, df = 3997.3, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.333737 -4.713168
sample estimates:
mean of x mean of y 
 9.896065 14.919518
```

The mean for group B is greater (14.91 vs. 9.89). Also, the *p*-value is highly significant (because $2.2 \times 10^{-16} \ll 0.05$), we can confidently claim that individuals in group B contribute a much higher revenue on average.

The complete code is here.

```r
library(readr)
library(ggplot2)
mydata = read_csv("https://ximarketing.github.io/data/AB.csv")
groupA = subset(mydata, treated == "A")
groupB = subset(mydata, treated == "B")
t.test(groupA$revenue, groupB$revenue)
```

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

Example of A/B test: Speed matters.

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

——Steve Souders

Example of A/B test: Speed matters.

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.

Example of A/B test: Speed matters.

While we may not be able to speed up the connection, it is rather easy to slow down. Consider the following two groups:

- Control group: The original speed
- Treatment group: Intentionally slow down by 100 msec.

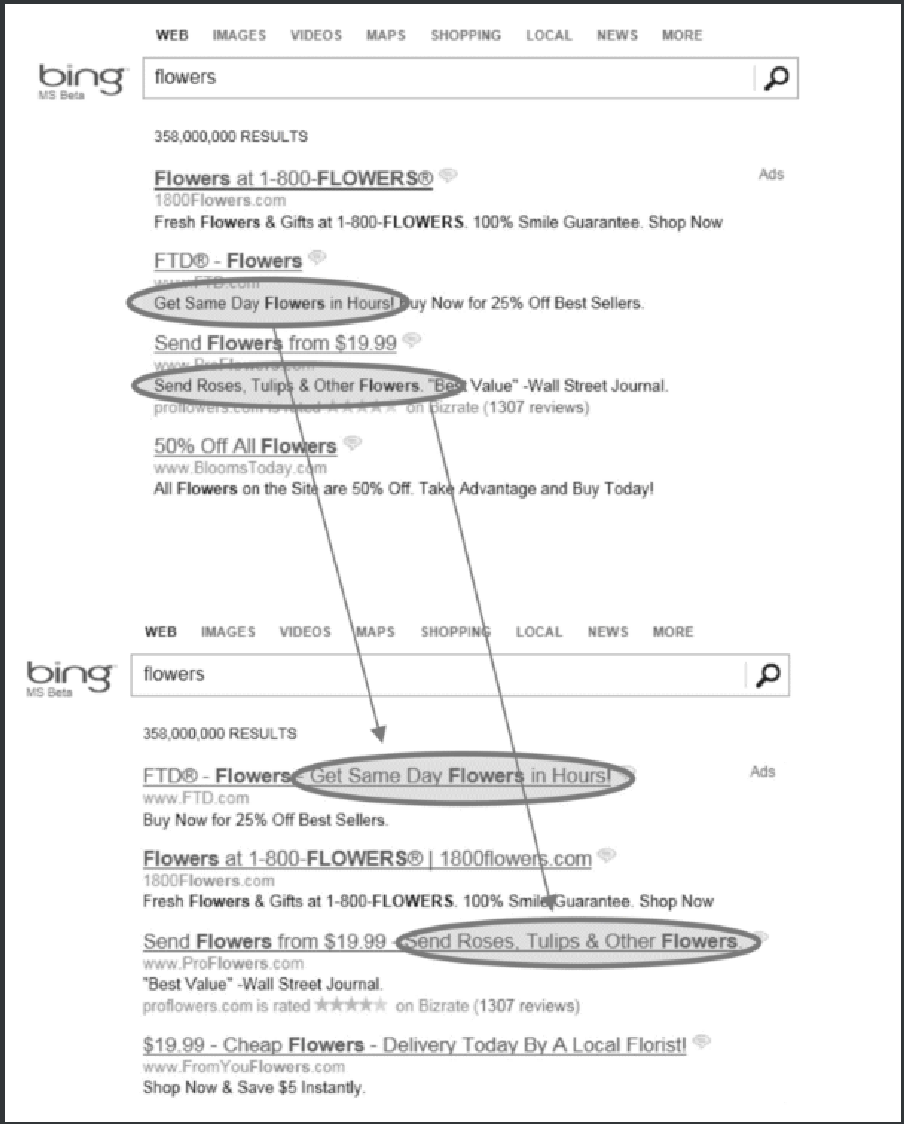
We then compare the performance of the two groups to see the effect of speed.

Example of A/B test: Speed matters.

At Amazon, a 100 msec slow down experiment decreased sales by 1% (Linden 2006).

An experiment at Bing revealed that a 100 msec slowdown is associated with a 0.6% change in revenue (Kohavi et al. 2013).

Example of A/B test



Example of A/B test

Nobody thought this simple change, among the hundreds suggested, would be the best revenue-generating idea in Bing's history! The feature was prioritized low and languished in the backlog for more than six months until a software developer decided to try the change, given how easy it was to code. An engineer implemented the idea and began evaluating the idea on real users, randomly showing some of them the new title layout and others the old one.

Example of A/B test

A few hours after starting the test, a revenue-too-high alert triggered, indicating that something was wrong with the experiment. The Treatment, that is, the new title layout, was generating too much money from ads.

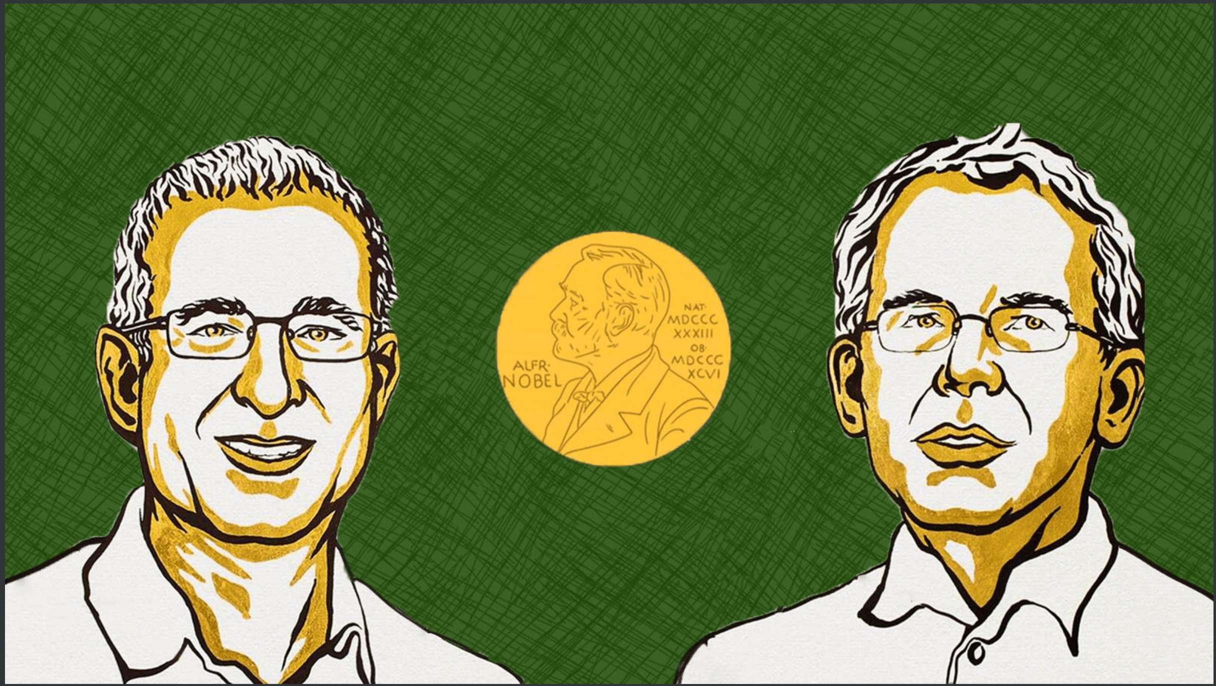
Bing's revenue increased by a whopping 12%, which at the time translated to over $100M annually in the US, without hurting key user-experience metrics. The experiment was replicated multiple times over a long period.

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 co
nsider 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) do not affect 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) do not affect one's income through any other channels.

Any ideas?

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