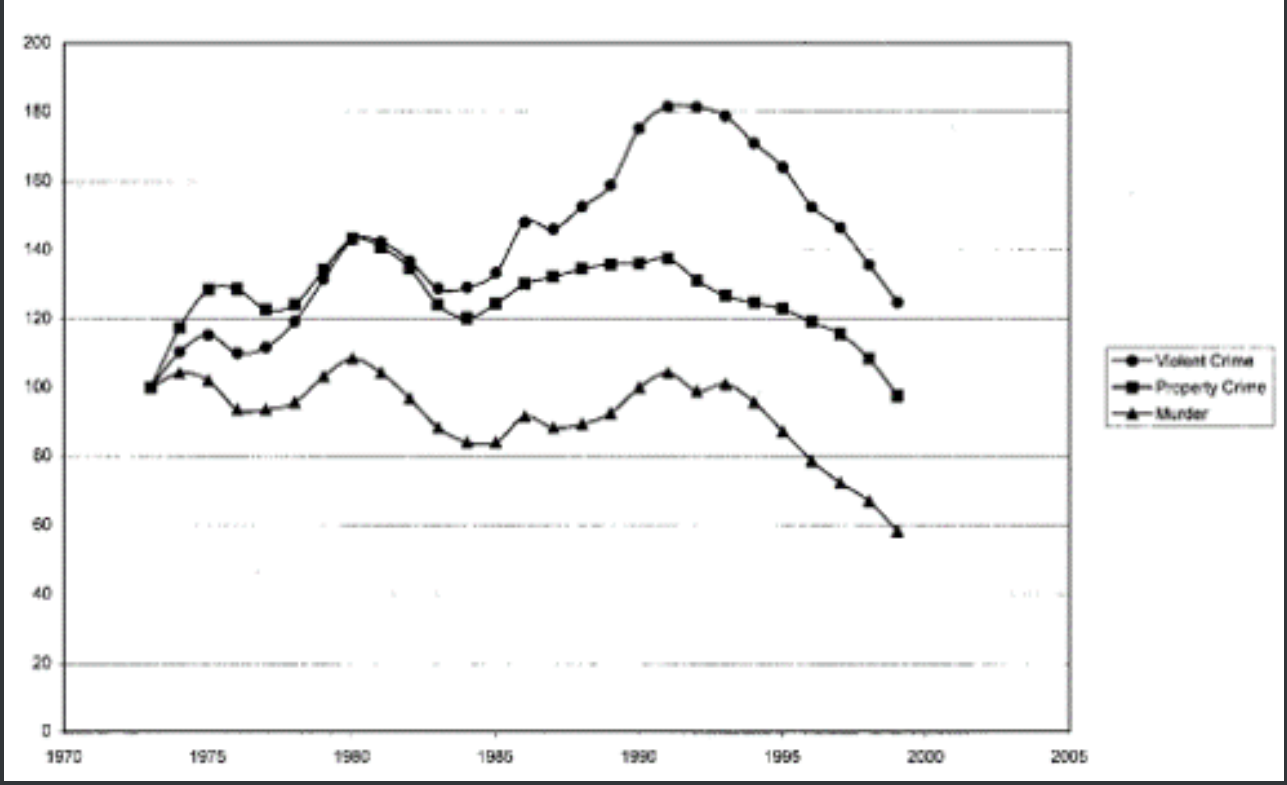
Causality

Since 1992, there is a sharp decline in the US crime rate. Can you guess the reason?



Since 1992, there is a sharp decline in the US crime rate. Can you guess the reason?

- Clinton administration: strong economy!
- Police force: high incarceration rate!
- Christian church: religious education!

Since 1992, there is a sharp decline in the US crime rate. Can you guess the reason?

- Steven Levitt (American economist): Legalized Abortion!
- 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.

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 $Y_i = a + bX_i$.

Back to Linear Regression

You run the regression and obtain the following result:

$$Y_i = 60 - 4X_i.$$

This result suggests that eating ice creams reduces one's weight! You double- and triple-checked your data and did not find any issues.

What can be wrong with your linear regression?

Back to Linear Regression

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 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.

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

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."

A key issue in economics is to estimate 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?

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.

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?

Suppose that you find $b > 0$. However, this does not necessarily mean more education helps you make more money. Consider one factor here: 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_i) becomes higher.
- At the same time, they can help you find better jobs. In this case, your income (Y_i) will also be higher.
- So, there is a positive relationship between X and Y; but we cannot yet say X has caused Y.

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

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?

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 do better in the final 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?

Let's take a student, Alice, as an example. Alice takes multiple classes, $j = 1, \ldots, n$, and her grade is $\text{Grade}_{\text{Alice},j}$ for j's class. Then, we can write

$$\text{Grade}_{\text{Alice},1} = a + b_1 \text{Hours}_{\text{Alice},1} + b_2 \text{IQ}_{\text{Alice}}$$
$$\text{Grade}_{\text{Alice},2} = a + b_1 \text{Hours}_{\text{Alice},2} + b_2 \text{IQ}_{\text{Alice}}$$

Let's take a student, Alice, as an example. Alice takes multiple classes, $j = 1, \ldots, n$, and her grade is $\text{Grade}_{\text{Alice},j}$ for j's class. Then, we can write

$$\text{Grade}_{\text{Alice},1} = a + b_1 \text{Hours}_{\text{Alice},1} + b_2 \text{IQ}_{\text{Alice}}$$
$$\text{Grade}_{\text{Alice},2} = a + b_1 \text{Hours}_{\text{Alice},2} + b_2 \text{IQ}_{\text{Alice}}$$

Taking the difference between the above two equations:

$$\text{Grade}_{\text{Alice},1} - \text{Grade}_{\text{Alice},2} = b_1 (\text{Hours}_{\text{Alice},1} - \text{Hours}_{\text{Alice},2})$$

You can then estimate the value of b_1!

Suppose that we want to know how hours of study affects your grades in the final exam. In addition to IQ, the easiness of the example is also an issue.

- If exam is easy, you will study less
- If exam is easy, you do better in the final exam

But we observe neither the IQ of students nor the easiness of exam for the classes. What can you do to answer the question?

$$\text{Grade}_{\text{Alice},1} = a + b_1\text{Hours}_{\text{Alice},1} + b_2\text{IQ}_{\text{Alice}} + b_3\text{Easiness}_1$$
$$\text{Grade}_{\text{Alice},2} = a + b_1\text{Hours}_{\text{Alice},2} + b_2\text{IQ}_{\text{Alice}} + b_3\text{Easiness}_2$$

The difference between the above two equations is:

$$\text{Grade}_{\text{Alice},1} - \text{Grade}_{\text{Alice},2} = b_1(\text{Hours}_{\text{Alice},1} - \text{Hours}_{\text{Alice},2}) + b_3(\text{Easiness}_1 - \text{Easiness}_2)$$

For Bob, we can derive the following equation:

$$\text{Grade}_{\text{Bob},1} - \text{Grade}_{\text{Bob},2} = b_1(\text{Hours}_{\text{Bob},1} - \text{Hours}_{\text{Bob},2}) + b_3(\text{Easiness}_1 - \text{Easiness}_2)$$

Taking differences again, we obtain that:

$$(\text{Grade}_{\text{Alice},1} - \text{Grade}_{\text{Alice},2}) - (\text{Grade}_{\text{Bob},1} - \text{Grade}_{\text{Bob},2}) =$$
$$b_1((\text{Hours}_{\text{Alice},1} - \text{Hours}_{\text{Alice},2}) - ((\text{Hours}_{\text{Bob},1} - \text{Hours}_{\text{Bob},2}))$$

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?

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?

This is not necessarily the case. There is another possibility: When people become anxious, they drink more coffee to calm down. In this case, it may be that anxiety causes coffee consumption, not the other way around.

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!

Actually, in many cases, X and Y affect each other:

- Anxiety leads to coffee consumption, but
- Coffee consumption leads to lowered anxiety.

In the example of police,

- High crime rates lead to an increase in police force,
- Police force helps reduce crime rate.

Exercise: Find other examples of reversed causality.

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

A hypothetical example

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

We cannot state that Harvard > HKU > Lanxiang. The reason is that, Harvard and HKU also attract better students, or your IQ is higher.

Are you making high salaries because of your school education, or simply because you are smarter?

A hypothetical example

In this example, we are interested in the more fundamental question: Holding other things (IQ, gender, ethnic, family background) equal, can Harvard students still make more money than HKU and Lanxiang students?

How to achieve that? We can use random assignment.

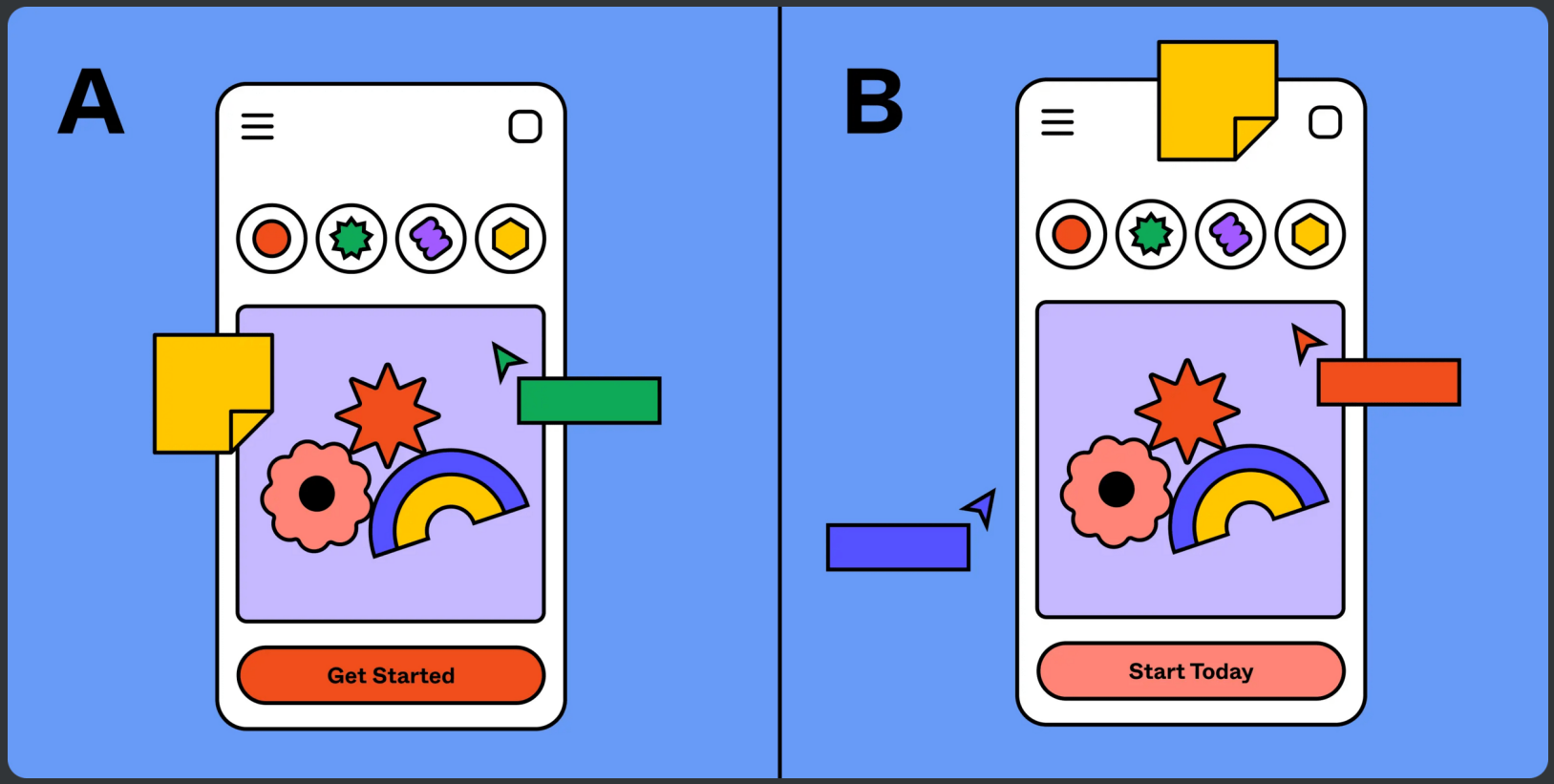
Suppose that there are many students applying to these schools, say 10,000. Then, we randomly admit 1,000 to Harvard, and randomly admit another 1,000 to HKU, and another 1,000 to Lanxiang. So, whether you are admitted to Harvard only depends on your luck, not your age, ability, IQ, talent, family background...

If there is difference in salary, the difference can only be driven by school education.

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?

Suppose that:

45 out of 856 [5.25%] users clicked through on ad A;
99 out of 1,298 [7.62%] 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?

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.

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

```
1 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
```

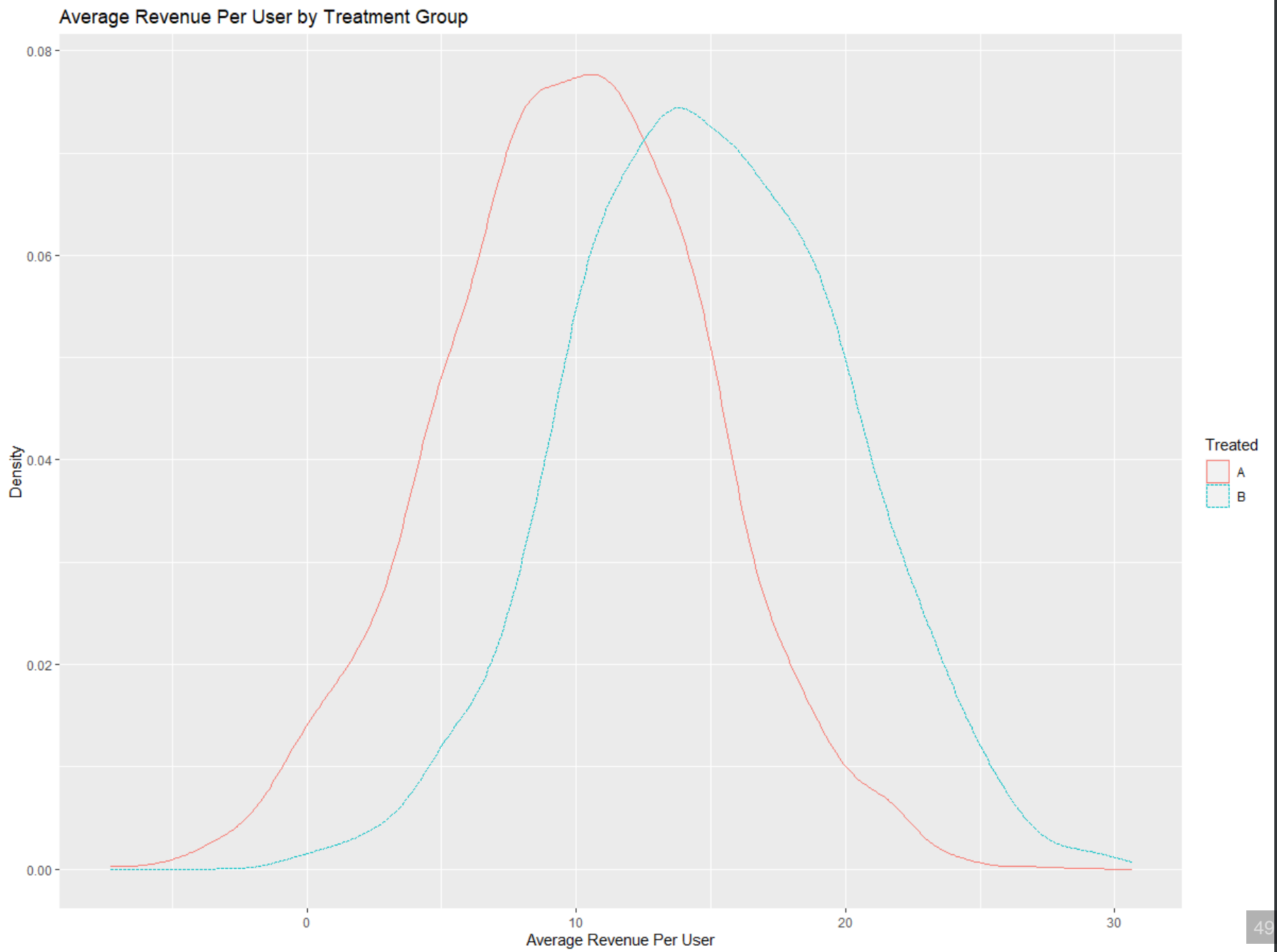
Here, we focus on the p-value. Typically, when $p < 0.05$, we claim the two conditions lead to significantly different outcomes; and in our case, $p < 0.001$, meaning that ad B is more effective than ad A.

The complete code is here.

```r
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.

```
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)
```

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

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.

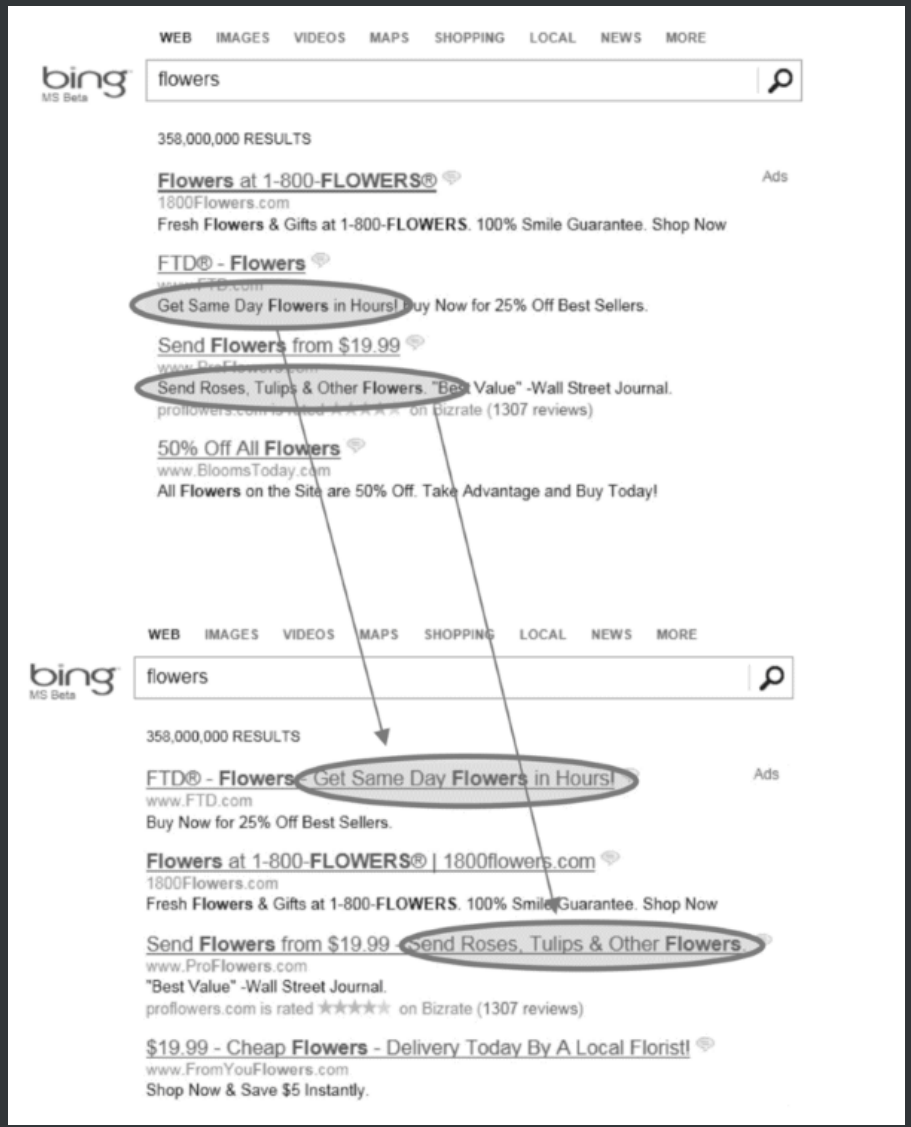
We then compare the performance of the two groups to see the effect of speed.

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.

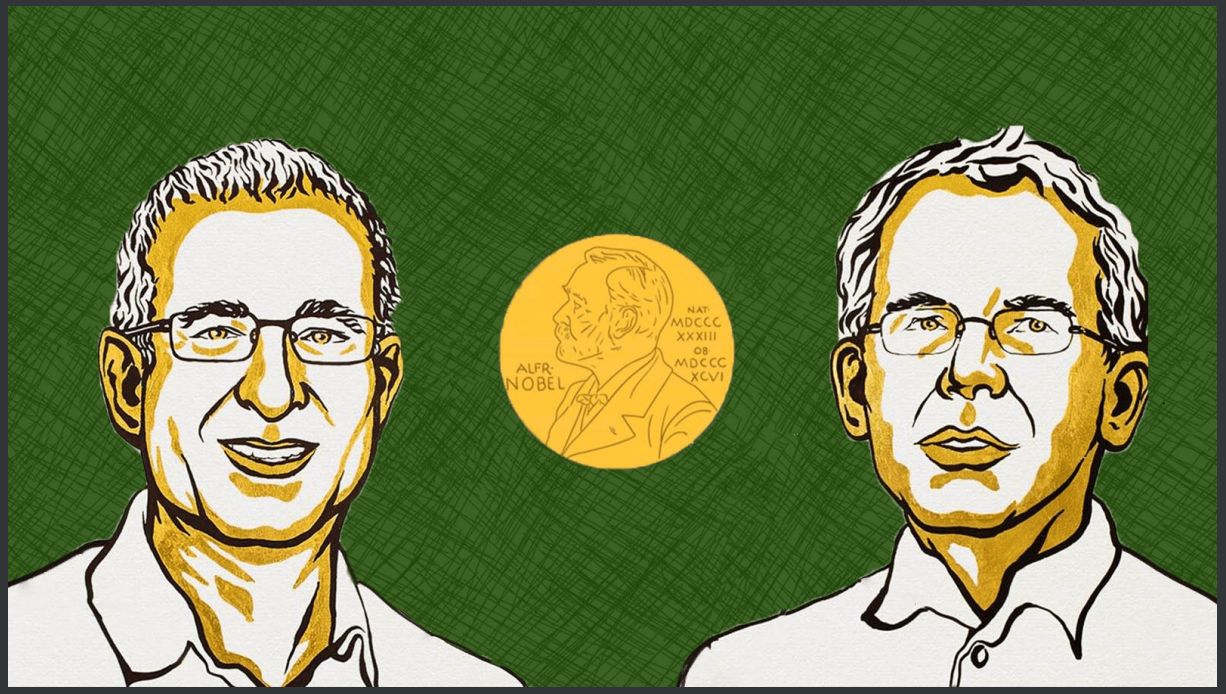
But you cannot force drivers to work in certain hours. What could you do in this case?

Question

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

Instead of forcing them to work, you can offer subsidies of random values to the drivers. When there is a high subsidy, more drivers will be willing to work for Uber.

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?

Instrumental Variables

This is really a famous problem! And there is also a famous instrument (which won the Nobel prize) – 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.

Instrumental Variables

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.

Instrumental Variables

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.

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.

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