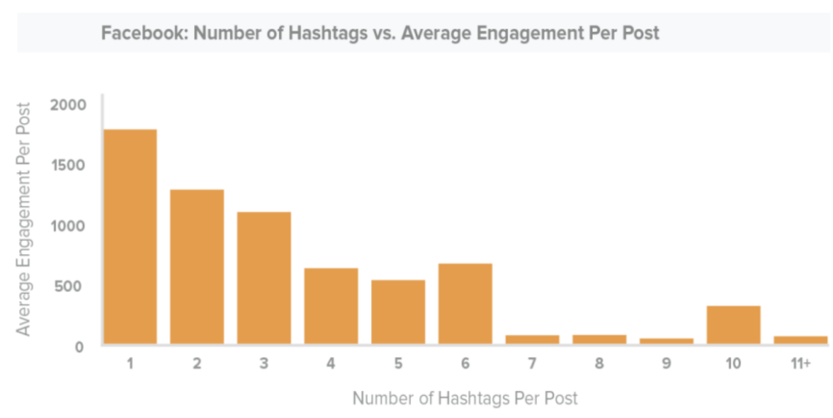
More on Regression Analysis

Question

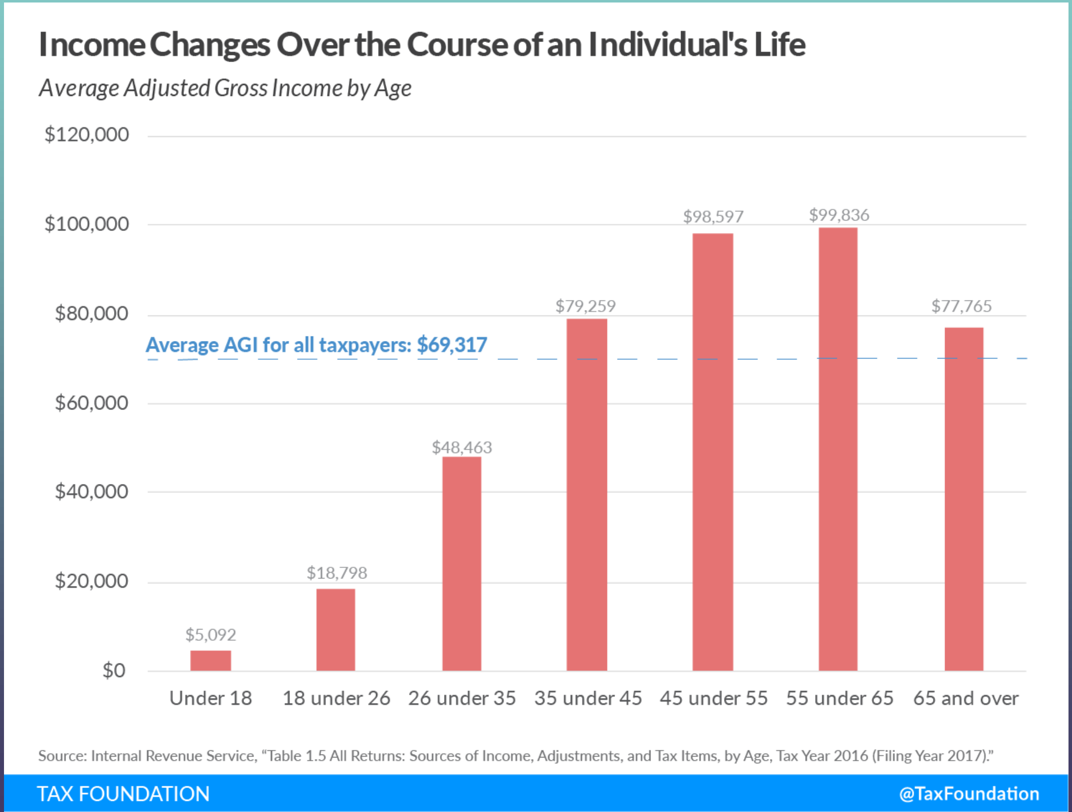
Suppose that we want to estimate the relationship between the number of Facebook hashtags of a post (X) and the level user engagement (e.g., number of shares or likes, Y). What type of analysis should you do?

Question



Suppose this is your data. You may regress engagement on the number of hashtags, and find that the coefficient is negative: Having more hashtags reduces user engagement.

Another Example



Income Changes Over the Course of an Individual's Life

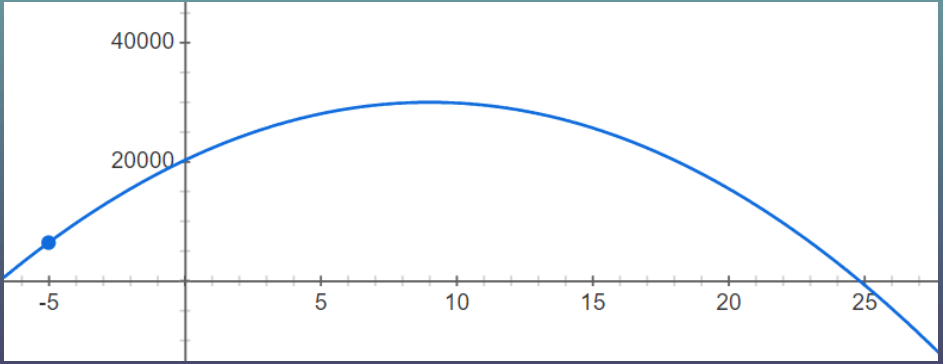
Average Adjusted Gross Income by Age

Source: Internal Revenue Service, “Table 1.5 All Returns: Sources of Income, Adjustments, and Tax Items, by Age, Tax Year 2016 (Filing Year 2017).”

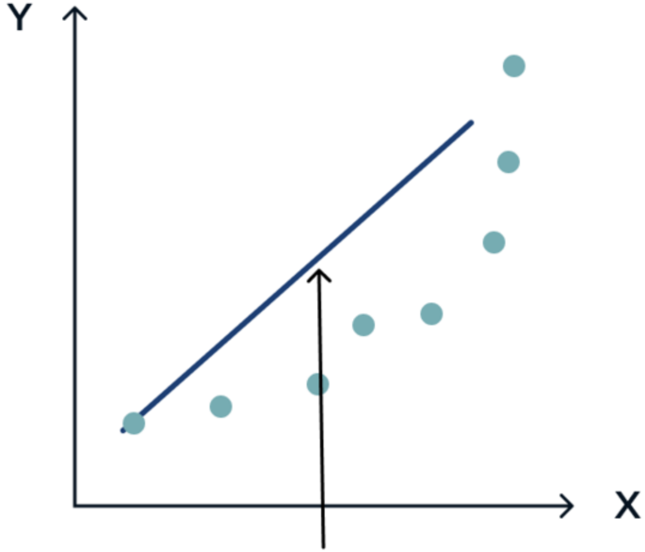
TAX FOUNDATION @TaxFoundation

Question

This is a nonlinear relationship between X and Y! This can be captured by a quadratic form. Consider the following example.

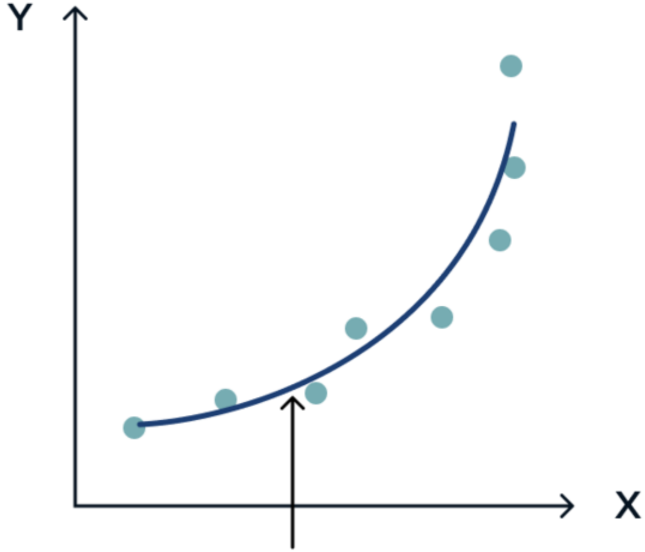


Simple linear model



$y=b_0+b_1x$

Polynomial model



$y=b_0+b_1x+b_2x_1^2$

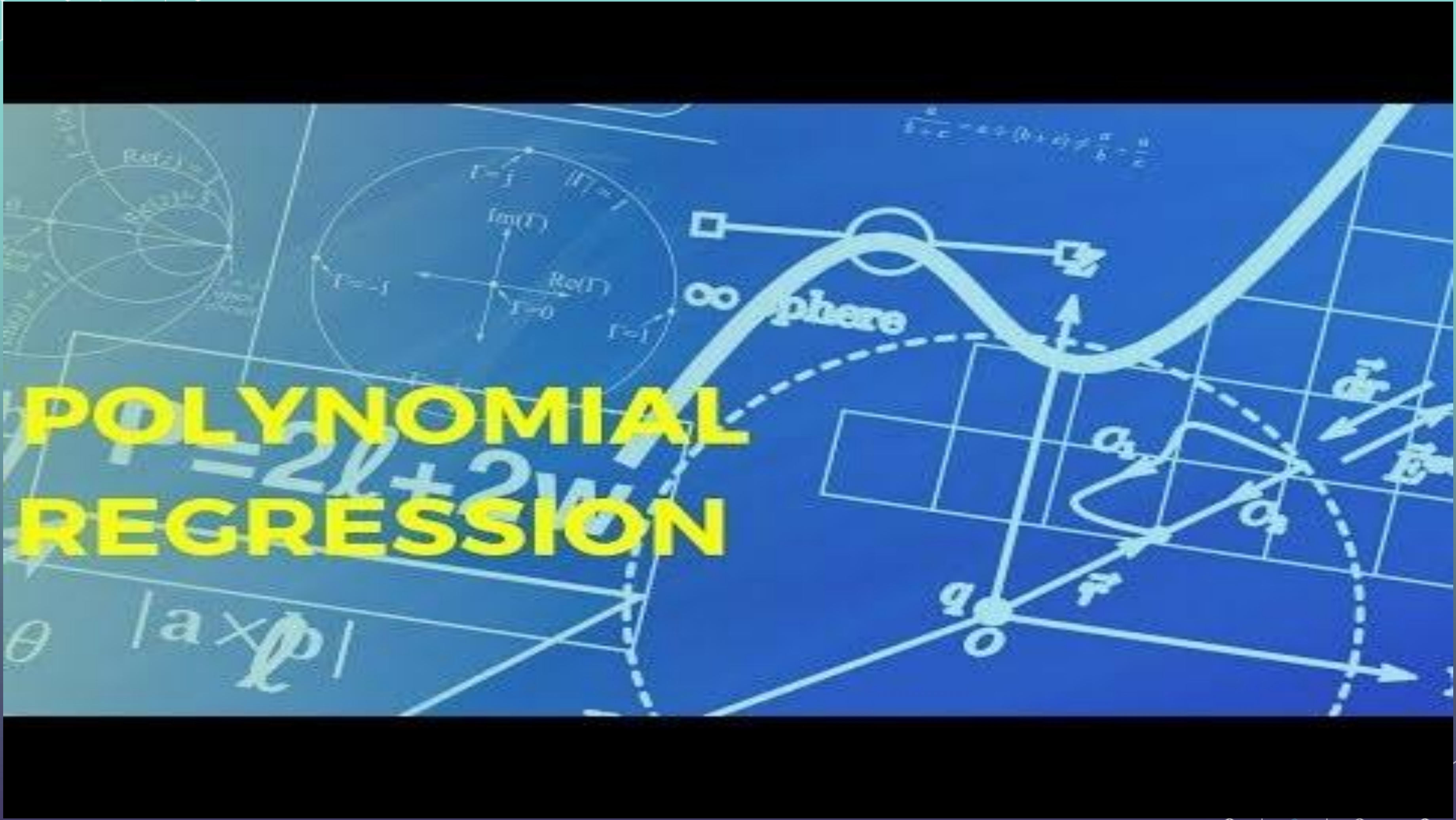
Quadratic Regression in R

Suppose that we believe that the relationship between X and Y is quadratic (as opposed to linear). Then, we would like to regress Y on both X and X^2. In this case, our regression question will be

$$Y = a + b_1X + b_2X^2$$

You can further extend the model to run cubic/polynomial regression…

$$Y = a + b_1X + b_2X^2 + b_3X^3 + \cdots$$


POLYNOMIAL
REGRESSION
P=2l+2w
|a×b|
θ
Re(Γ)
Im(Γ)
Γ=0
Γ=1
Γ=−1
∞
sphere
q
O
r⃗
dr⃗
E⃗

Crowdfunding: An example

We want to investigate the relationship between video length and the chance of success. Let us prepare the data:

```
mydata <-
read.csv("https://ximarketing.github.io/class/Kickstart
er-Project.csv", fileEncoding = "UTF-8-BOM")
```

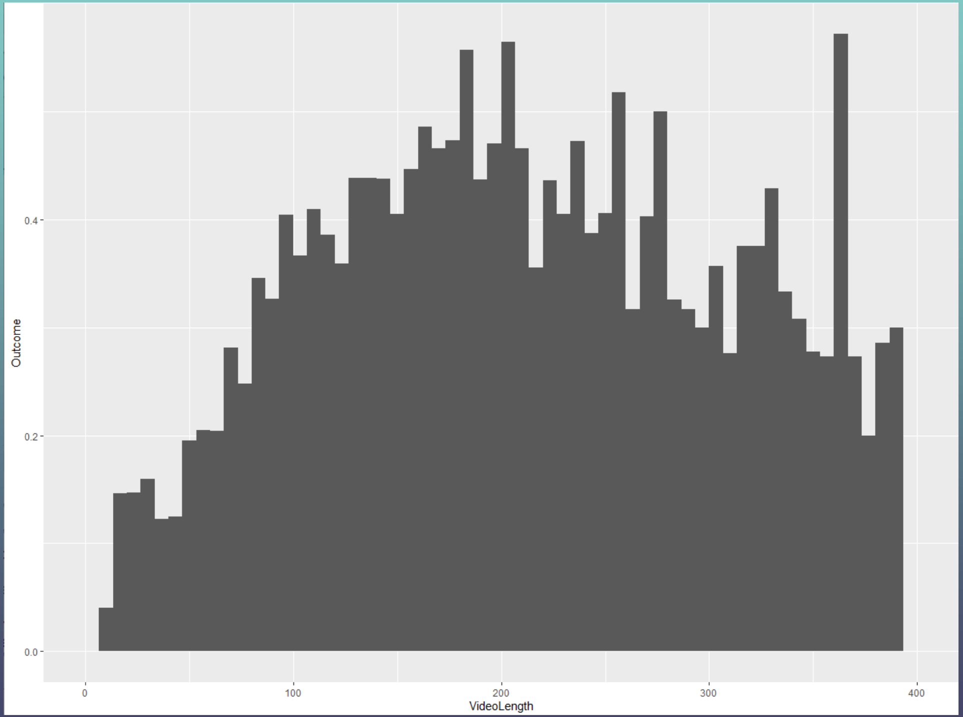
```
subdata = subset(mydata, IsVideoAvailable == 1)
```

Crowdfunding: An example

We want to investigate the relationship between video length and the chance of success. Let us prepare the data:

```
library(ggplot2)
ggplot(subdata,  mapping = aes(VideoLength, Outcome)) +
stat_summary_bin(fun.y="mean", geom="bar",
bins=60)+xlim(0, 400)
```

Crowdfunding: An example



Crowdfunding: An example

It seems that the relationship between the video length and project success is nonlinear: When video length is short, increasing video length improves the success rate. However, having a very lengthy video does not benefit the project either.

Crowdfunding: An example

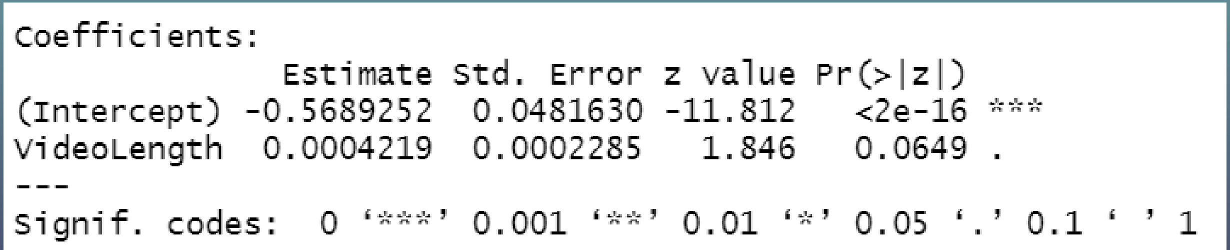
Let us try the following logistic regression:

$$\Pr(\text{Success}) = \frac{1}{\exp(-a - b_1\text{Length} - b_2\text{Length}^2)}$$

Crowdfunding: An example

Without quadratic term:

```
logit <- glm(Outcome ~ VideoLength, data = subdata,
family = "binomial")
summary(logit)
```



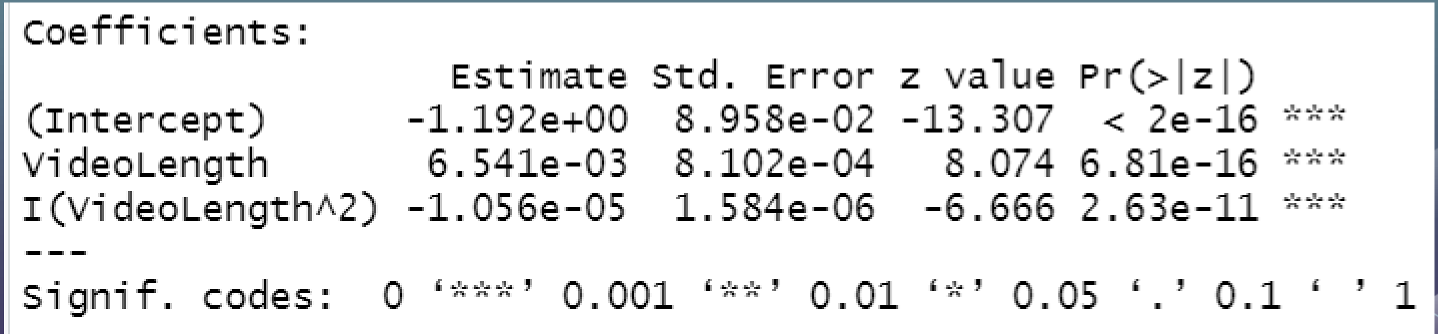
```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)    
(Intercept) -0.5689252  0.0481630 -11.812   <2e-16 ***
VideoLength  0.0004219  0.0002285   1.846   0.0649 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The longer the video is, the more successful the project will be.

Crowdfunding: An example

With quadratic term:

```
logit <- glm(Outcome ~ VideoLength + I(VideoLength^2),
data = subdata, family = "binomial")
summary(logit)
```



```
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)    
(Intercept)      -1.192e+00  8.958e-02 -13.307  < 2e-16 ***
VideoLength       6.541e-03  8.102e-04   8.074 6.81e-16 ***
I(VideoLength^2) -1.056e-05  1.584e-06  -6.666 2.63e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Crowdfunding: An example

We can further calculate the optimal length of the video: For a quadratic function $f = b_2 x^2 + b_1 x + a \quad (b_2 < 0)$, the function is maximized when

$$x = \frac{b_1}{2b_2}$$

If you forgot about it, please review your high school math. Link is here.

In our regression, $b_2 = -1.056 \times 10^{-5}$ and $b_1 = 6.541 \times 10^{-3}$. Then, we can calculate that the optimal length of the video is around 300 seconds (5 minutes).

Question

Suppose that you want to predict students' performance in exam. Two factors come into play: IQ and Hours of Study.

A student with a higher IQ is more clever, and gets higher grades on average.

A student who studies longer hours understands the content better, and gets higher grades on average.

Question

Let’s run the following linear regression:

$$\text{Grades} = \text{a} + \text{b}_1\text{IQ} + \text{b}_2\text{Hours}$$

Are we missing anything?

Examples of Interaction Effects

Suppose that your dependent variable is a programmer's salary.

Suppose that you have two independent variables: the programmer's knowledge of Python and his/her knowledge of R.

We find that

$$\text{Salary} = 1 + 3\text{Python} + 2\text{R} - 0.5\text{Python} \times \text{R}$$

How would you interpret this regression result?

Examples of Interaction Effects

$$\text{Salary} = 1 + 3\text{Python} + 2\text{R} - 0.5\text{Python} \times \text{R}$$

If you know more about Python, you can make a higher salary.

If you know more about R, you can maker a higher salary.

However, if you already know Python well, then knowing more about R does not help much, and vice versa.

This result suggests that Python and R are substitutes: After learning about one thing, learning about the other does not help you much.

Examples of Interaction Effects

Suppose that your dependent variable is a person's health score.

Suppose that you have two independent variables: the amount of swimming and running.

$$\text{Health} = 4 + 5\text{Running} + 3\text{Swimming} + 2\text{Running} \times \text{Swimming}$$

How would you interpret this regression result?

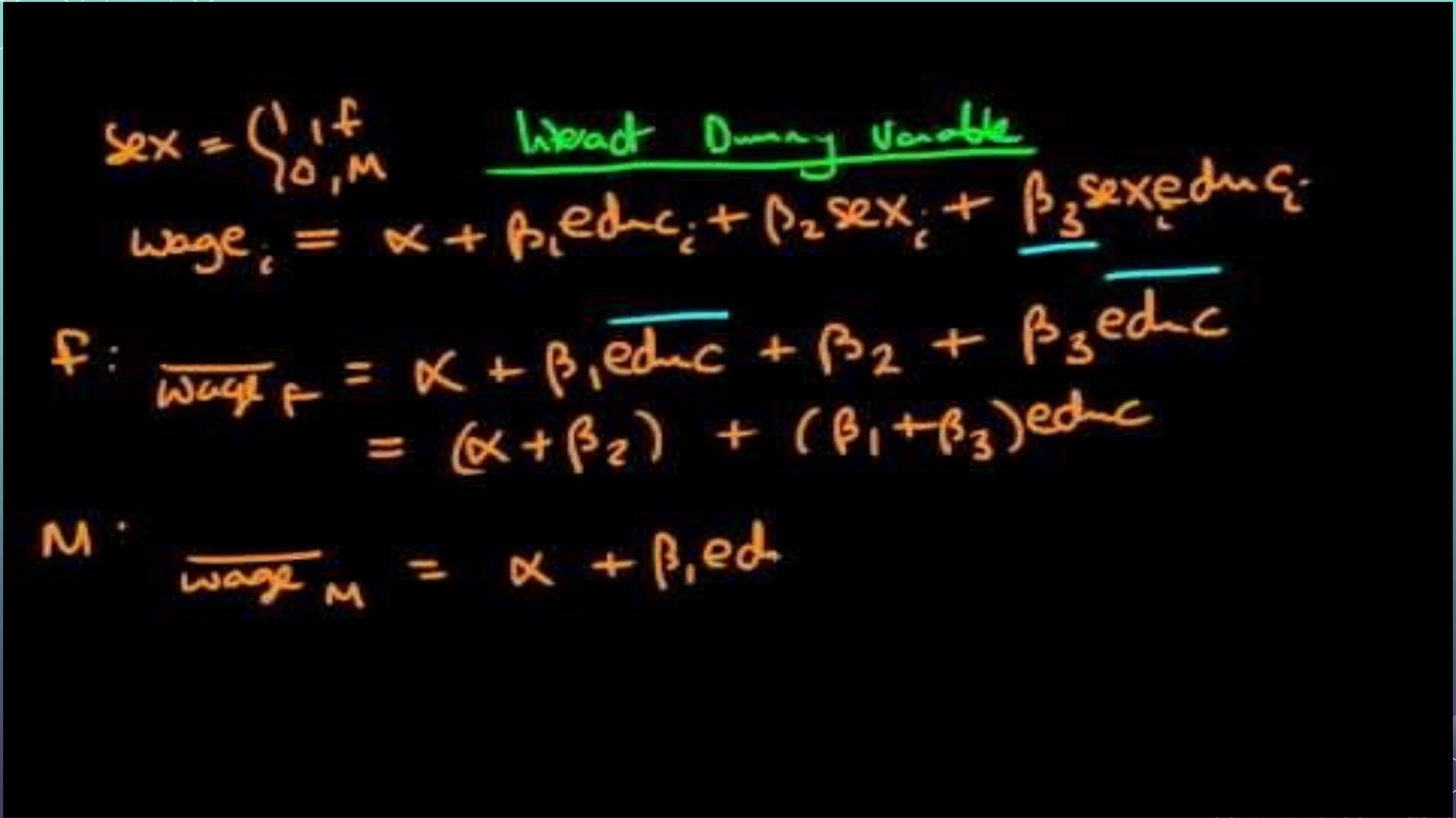
Examples of Interaction Effects

Suppose that your dependent variable is a person's health score.

Suppose that you have two independent variables: the amount of running exercise and whether or not the person is overweight.

$$\text{Health} = 4 + 5\text{Running} - 2\text{Overweight} + 3\text{Running} \times \text{Overweight}$$

How would you interpret this regression result?



Interact Dummy Variable

$$sex = \begin{cases} 1, f \\ 0, M \end{cases}$$

$$wage_i = \alpha + \beta_1 educ_i + \beta_2 sex_i + \underline{\beta_3} sex_i educ_i$$

f: $$\overline{wage}_F = \alpha + \beta_1 \overline{educ} + \beta_2 + \beta_3 \overline{educ}$$

$$= (\alpha + \beta_2) + (\beta_1 + \beta_3) educ$$

M: $$\overline{wage}_M = \alpha + \beta_1 ed$$

Crowdfunding: An example

We want to investigate the relationship between the total funding, the creators' experience and the number of products offered. Let us prepare the data:

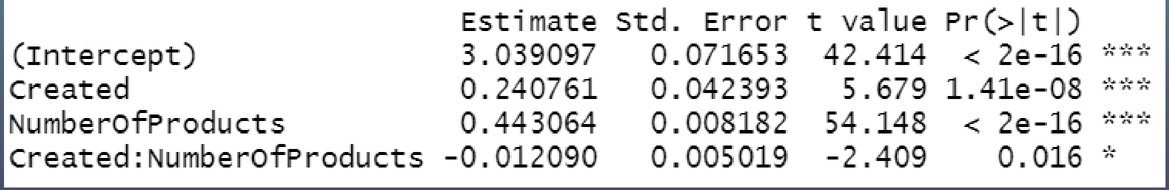
```r
mydata <-
read.csv("https://ximarketing.github.io/class/Kickstarter-Project.csv", fileEncoding = "UTF-8-BOM")

mydata$LogFunding = log(mydata$FundingRaised + 1)
```

Crowdfunding: An example

We want to investigate the relationship between the total funding, the creators' experience and the number of products offered. Let us run a regression with an interaction term:

```
result = lm(LogFunding ~ Created *
NumberOfProducts, data = mydata)
summary(result)
```



```
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)                3.039097   0.071653  42.414  < 2e-16 ***
Created                    0.240761   0.042393   5.679 1.41e-08 ***
NumberOfProducts           0.443064   0.008182  54.148  < 2e-16 ***
Created:NumberOfProducts -0.012090   0.005019  -2.409    0.016 *  
```

Crowdfunding: An example

$$\text{LogFunding} = 3.04 + 0.24\text{Created} + 0.44\text{Number of Products} - 0.012\text{Created} \times \text{Number of Products}$$

What does this result tell us?

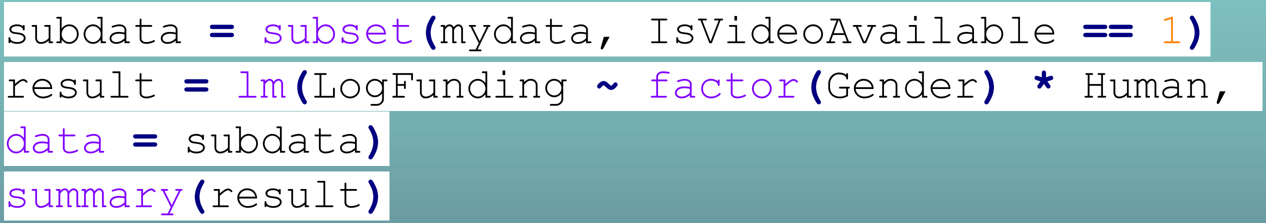
Crowdfunding: Another example

Let us explore something interesting.

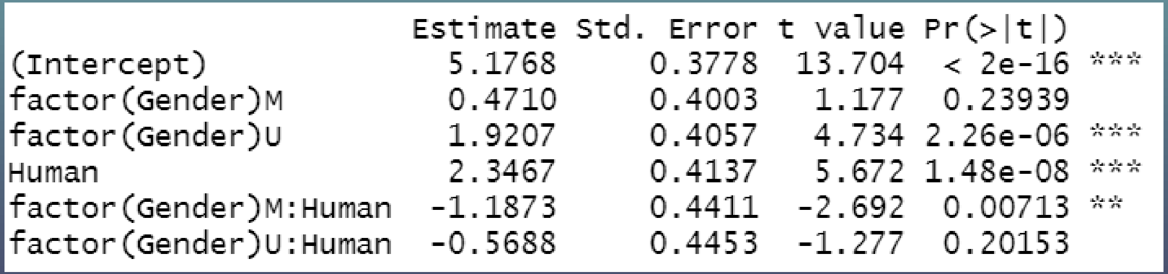
We already know that in a crowdfunding project, putting your face in front of the camera makes the project more successful.

However, does it make a difference whether this is a female face or a male face? What’s your intuition?

Crowdfunding: Another example



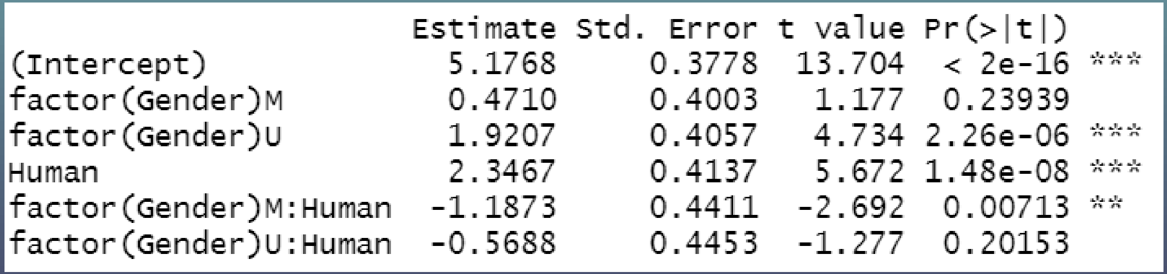
```
subdata = subset(mydata, IsVideoAvailable == 1)
result = lm(LogFunding ~ factor(Gender) * Human,
data = subdata)
summary(result)
```



```
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)              5.1768     0.3778  13.704  < 2e-16 ***
factor(Gender)M          0.4710     0.4003   1.177  0.23939    
factor(Gender)U          1.9207     0.4057   4.734 2.26e-06 ***
Human                    2.3467     0.4137   5.672 1.48e-08 ***
factor(Gender)M:Human   -1.1873     0.4411  -2.692  0.00713 ** 
factor(Gender)U:Human   -0.5688     0.4453  -1.277  0.20153    
```

Crowdfunding: Another example

This result tells us that, featuring a human in your video is beneficial. Nonetheless, featuring a male is less helpful compared to featuring a female.



```
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)               5.1768     0.3778  13.704  < 2e-16 ***
factor(Gender)M           0.4710     0.4003   1.177  0.23939    
factor(Gender)U           1.9207     0.4057   4.734 2.26e-06 ***
Human                     2.3467     0.4137   5.672 1.48e-08 ***
factor(Gender)M:Human    -1.1873     0.4411  -2.692  0.00713 ** 
factor(Gender)U:Human    -0.5688     0.4453  -1.277  0.20153    
```

Exercise

Play with the Kickstarter dataset yourself and see if you can find any interesting interaction effects. Share it with us!

More 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). Let us ignore the interaction effect at the moment.

What should we do?

More on Fixed Effects

Suppose that the actual regression equation is

$$\text{Grades} = \text{a} + \text{b}_1\text{IQ} + \text{b}_2\text{Hours}$$

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

$$\text{Grades} = \text{a} + \text{b}_1\text{IQ} + \text{b}_2\text{Hours}$$

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

$$\text{Grades}_{\text{Alice,j}} = \text{a} + \text{b}_1\text{IQ}_{\text{Alice}} + \text{b}_2\text{Hours}_{\text{Alice,j}}$$

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

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?

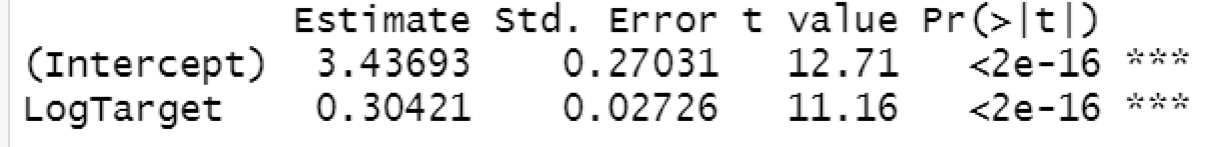
Individual fixed effects --- some people are more successful than others.

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

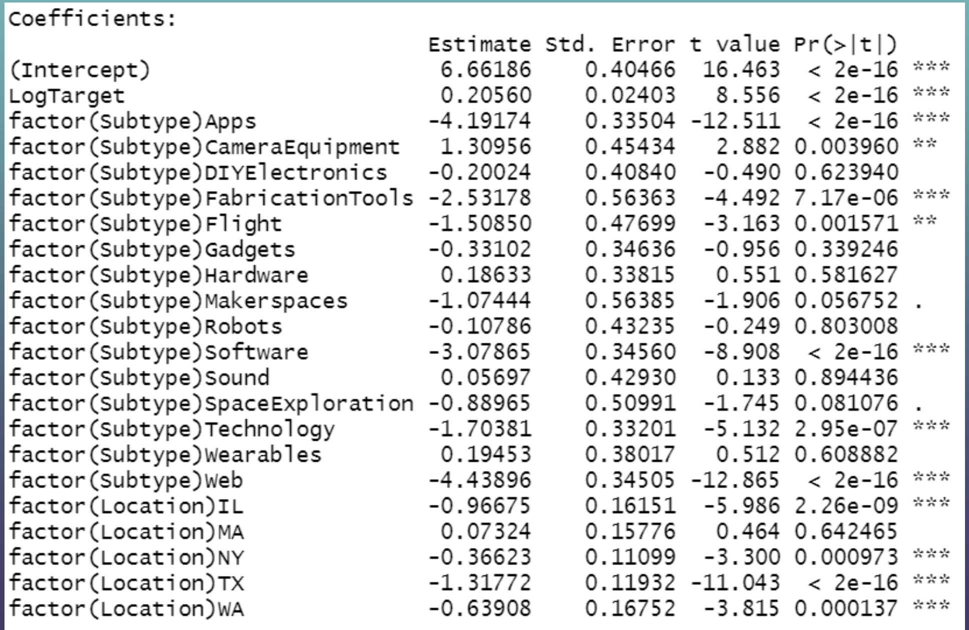
Now consider the following regression:

Here, we control for fixed effects.

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

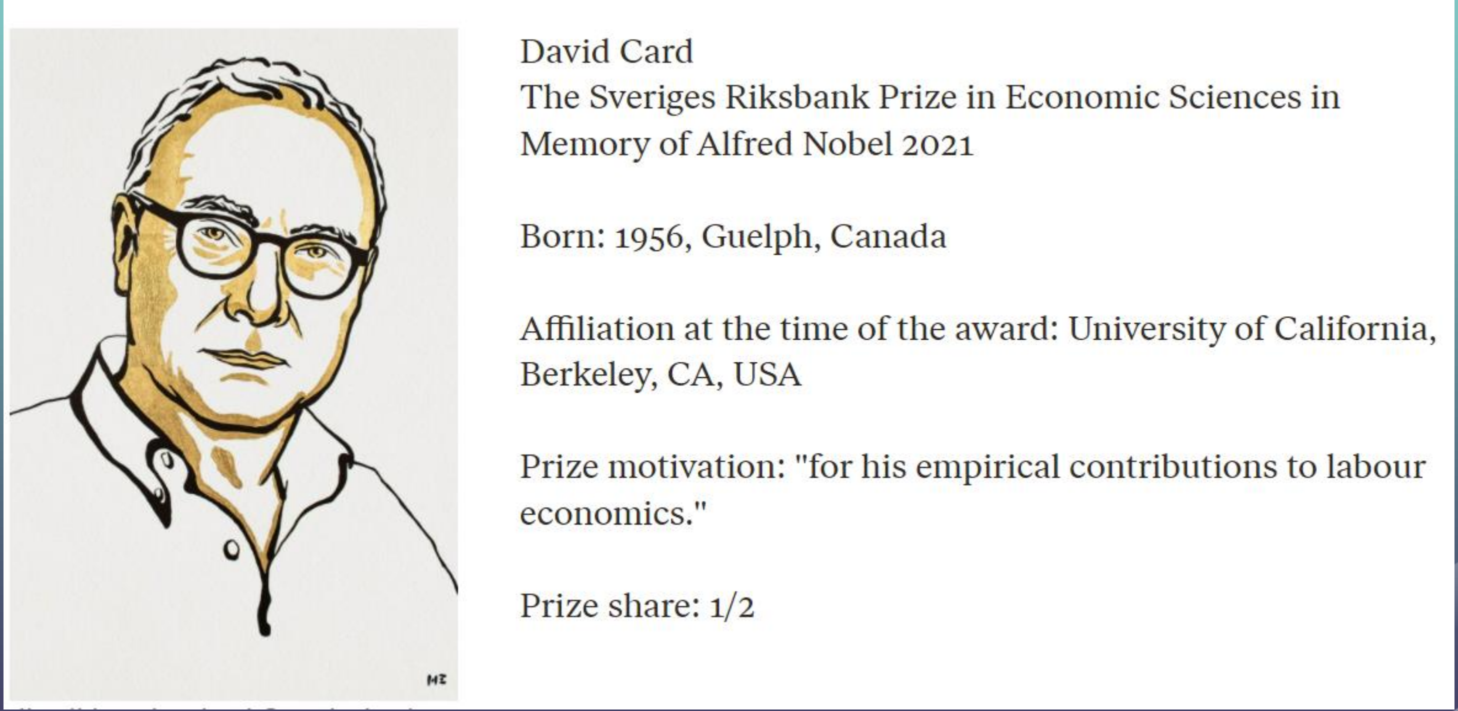
A Crowdfunding Example

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:

$$\mathrm{Unemployment}_{\mathrm{it}} = \mathrm{a} + \mathrm{State}_{\mathrm{i}} + \mathrm{Month}_{\mathrm{t}} + \mathrm{bWage}_{\mathrm{it}}$$

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

Diff-in-Diff

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

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

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

Diff-in-Diff

Online Review

Playing with TripAdvisor Data

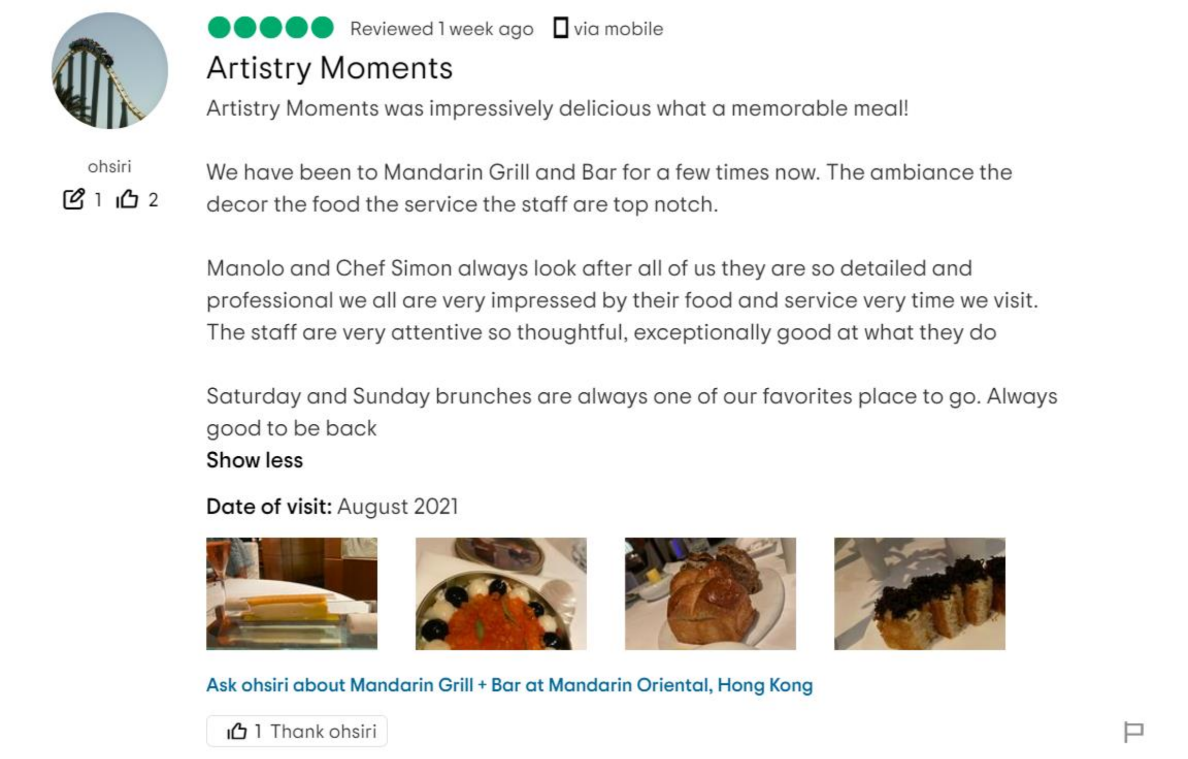
Background

We all know that online reviews are important, and our purchase decisions are likely to be influenced by online reviews.

TripAdvisor

Founded in 2000, TripAdvisor is one of the leading online review platforms. It mainly focuses on hotel and restaurant reviews.

A Typical TripAdvisor Review



●●●●● Reviewed 1 week ago via mobile

ohsiri

1 2

Artistry Moments

Artistry Moments was impressively delicious what a memorable meal!

We have been to Mandarin Grill and Bar for a few times now. The ambiance the decor the food the service the staff are top notch.

Manolo and Chef Simon always look after all of us they are so detailed and professional we all are very impressed by their food and service very time we visit. The staff are very attentive so thoughtful, exceptionally good at what they do

Saturday and Sunday brunches are always one of our favorites place to go. Always good to be back

Show less

Date of visit: August 2021

Ask ohsiri about Mandarin Grill + Bar at Mandarin Oriental, Hong Kong

1 Thank ohsiri

Data We Have

Reviews of Top 30 restaurants in the following cities:

New York, Las Vegas, Los Angeles, Chicago, Toronto, Vancouver, London, Sydney. 240 restaurants in total.

These are all English-speaking cities, and we only focus on reviews written in English that can be analyzed automatically.

Almost 150K reviews.

Data about the reviewer

Local: Whether or not the reviewer is a local resident (1 = local; 0 = nonlocal).

We incorporate this variables because local people’s preference may be different from that of visitors.

Data about the reviewer

CountResataurant: How many restaurants the reviewer has been to.

This variable captures the experience of the reviewers. An experienced foodies may differ from a normal person (e.g., a foodie may be more critical).

Data about the reviewer

CountReview: How many reviews the reviewer has written.

This variable also captures the experience of the reviewers.
The more reviews written, the more experienced the reviewer.

Data about the reviewer

CountVotes: How many helpful votes the reviewer has received.

This variable also captures the experience of the reviewers. The more votes, the more popular the reviewer.

Data about the review

Rating: The rating assigned by the reviewer.

TripAdvisor uses a 5-point ratings, with 5 being the best and 1 being the worst.

Data about the review

Helpful: The number of helpful votes the review received.

When a review receives more helpful votes, the review is more popular.

Data about the review

Mobile: Whether or not the review is typed from a mobile device (1 = mobile, 0 = nonmobile).

Mobile devices are small, and reviewers’ behavior may be different when using mobile devices.

Data about the review

Photo: Number of photos in the review.

Date: The date the review was posted on TripAdvisor.

Data has format YYYY-MM-DD.

Data about the review

Title Length: The length of the review title.

Length: The length of the review body.

Both are measured in number of characters.

Data about the review

We also use sentiment analysis to capture the sentiment of the review:

Sentiment: the polarization of the review (-1 to 1)
Subjectivity: the subjectivity of the review (0 = objective, 1 = subjective)

Data about the review

We also use sentiment analysis to capture the sentiment of the review:

Happy / Angry / Sad / Surprise: the emotion that is captured from consumer review. Each of them is a value between 0 to 1, and when the value is larger, it means the emotion is stronger.

Data about the review

We also analyze the content of photos of the review, if the review has at least one photo:

Menu: Whether or not there is a photo of the restaurant menu.
Building: Whether or not there is a photo of the restaurant building.
Meat: Whether or not there is a photo of meat.
Vegetable: Whether or not there is a photo of vegetable.
Person: Whether or not there is a photo of a person (e.g., a selfie)

Sample Questions

What makes a helpful review?

How does the reviewer's experience affect the characteristics of a review?

Is it true that "one picture is worth 1000 words?"

On Your Data Analysis

Try to incorporate at least one interaction effect in your data analysis.

Use quadratic terms and fixed effects whenever necessary.

Again, it would be desirable if you can come up with something surprising.

On Your Data Analysis

As we learned last week, when running regression analysis, we cannot easily make claims that X causes Y.

In most cases you are not able to run an experiment or find an instrumental variable.

It is fine that you cannot prove causality. Just be careful with how you draw your conclusions (do not easily conclude that something has caused the other; you can claim that they are correlated, though).

On Your Data Analysis

As before, choose 1 or 2 research questions and upload your slides to the course website (up to 20 slides).

Deadline: Jan 22, 5pm (Class A) 9:30pm (Class B) – one week from now.

On Your Final

The final exam will take place on Jan 26, Wednesday.

Online exam.

Any print materials are allowed. It can be slides, cheat sheets, books, or anything else that you want to bring. No electronic devices are permitted.

You can bring a calculator with you. However, I don’t think you will need to use one (we don’t really have questions that need to be answered with a calculator).

On Your Final

The purpose of the final is not to make your life difficult; it is intended to assign your grades in a more objective way.