# Log Transformation in Regression

#### **Linear Regression**

In a linear regression, we assume that the relationship between the dependent variable and independent variable is linear, i.e., we specify the following relationship:

$$Y = a + bX$$

#### **Log Transformations**

But sometimes we also take the log-transformation of the linear regression. For example, consider the following relationship:

$$\log Y = a + b \log X$$

Here, we typically use the natural logarithms (base is  $e \approx 2.718$ ) in log transformation.

#### The Log Function

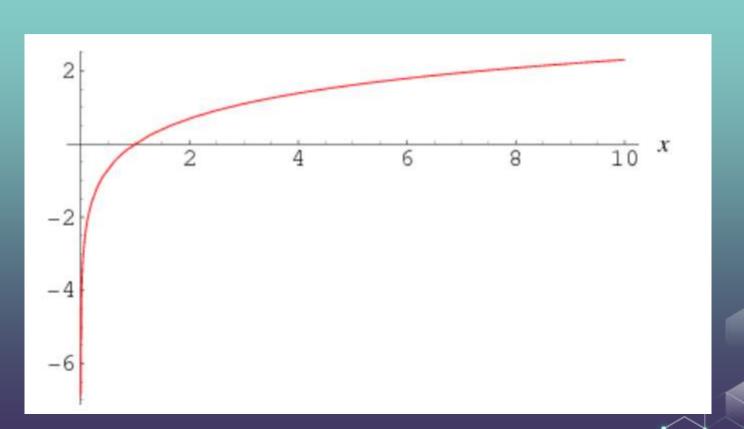
If  $e^a = b$ , then  $\log(b) = a$ , where  $e \approx 2.718$ 

For example,

 $log(10) \approx 2.3026$ 

 $\log(100) \approx 4.6052$ 

# **The Log Function**

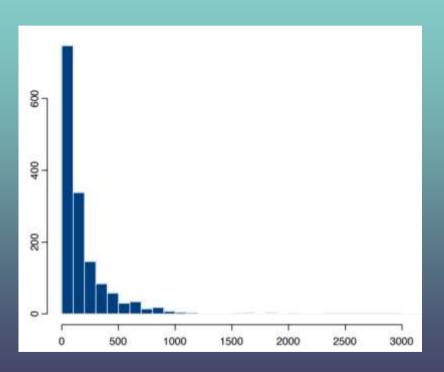


#### Question

When should we take log transformation instead of directly using the linear regression?

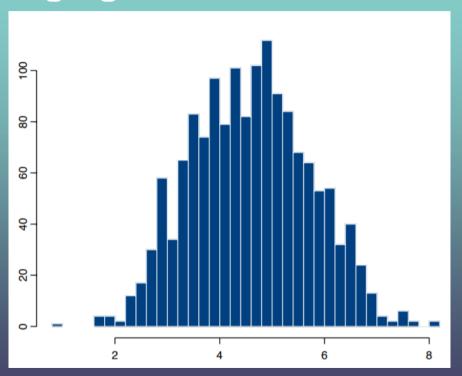
We typically take the log transformation when the original variable is right-skewed, i.e., when the right tail is much longer than the left tail.

#### A highly skewed distribution



We see a significant right skew in this data, meaning the mass of cases are bunched at lower values.

#### A highly skewed distribution



If you take the log transformation, the distribution looks like a normal distribution.

#### Types of log transformation

Linear-log models:

 $Y = a + b \log X$ 

Log-linear models:

 $\log Y = a + b X$ 

Log-log models:

 $\log Y = a + b \log X$ 

#### Interpreting Linear-Log Models

$$Y = a + b \log X$$

This result suggests that if X increases by 1 percent, Y will increase by  $0.01 \times b$ .

#### **Interpreting Log-Linear Models**

$$\log Y = a + bX$$

For example, when b = 0.04, it means if X increases by 1, Y will increase by 4 percent.

#### **Interpreting Log-Log Models**

$$\log Y = a + b \log X$$

When X increases by 1 percent, Y will increase by b percent.

Very commonly used in economic modeling.



Suppose that we want to investigate the effect of month on consumers' spending. Here, the independent variable is the month, e.g., January, February, March, ...

How to run this regression?

Suppose that we want to run regress Y (hotel rating) on X (purpose of trip). Here, purpose takes the following values: business, couple, family, friend, solo, and unknown. We run a fixed effects regression as follows.

```
mydata <-
read.csv("https://ximarketing.github.io/class/teaching
files/r-exercise.csv", fileEncoding = "UTF-8-BOM")
result <- lm(Votes ~ factor(Purpose), data = mydata)
summary(result)</pre>
```

Here, we take business as a benchmark and compare other purpose against it.

```
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)
                       0.680605
                                  0.008747
                                           77.810
                                                    < 2e-16
factor(Purpose)couple
                       0.120705
                                  0.011177
                                            10.800
                                                   < 2e-16
factor(Purpose)family
                       0.048585
                                  0.012091
                                            4.018 5.86e-05
factor(Purpose)friend
                                  0.016421
                                            1.033 0.30153
                       0.016965
factor(Purpose)solo
                       0.048603
                                  0.018849
                                             2.579 0.00992 **
factor(Purpose)Unknown
                      1.103785
                                  0.016950
                                            65.120
                                                    < 2e-16
```

You can also change your benchmark (e.g., use family as your benchmark):

```
mydata$Purpose <-
relevel(factor(mydata$Purpose), ref =
"family")
result <- lm(Votes ~ factor(Purpose), data =
mydata)
summary(result)</pre>
```

# Crowdfunding

Playing with Kickstarter Data

#### **Crowdfunding**

Are you familiar with these platforms?



# **Understanding Crowdfunding**



#### 4 Types of crowdfunding



#### **Equity Based Crowdfunding**

The backer receives shares of a company, usually in its early stages, in exchange for the money pledged.

Example:



#### **Debt Based Crowdfunding**

Debt-based crowdfunding is a crowdfunding model used to raise capital by taking loans from several investors (lenders) who expect to be repaid their loan with an added interest over the period that the loan was "used". The entire process takes place through a crowdfunding platform.

Example:



#### **Donation Based Crowdfunding**

Donation-based crowdfunding is when money is raised to support a good cause. As the name suggests, funding is raised through a crowd of people who decide to donate a certain amount of money to the cause, normally via online platforms specifically designed for the purpose.

Example:



#### **Rewards Based Crowdfunding**

Rewards-based, or seed, crowdfunding is a type of small-business financing in which entrepreneurs solicit financial donations from individuals in return for a product or service. There are about 19 times as many rewards campaigns as there are for its closely related counterpart, equity-based crowdfunding.

It is closely related to marketing and we focus on it in our class.

In the following slides, we introduce some simple economics of crowdfunding, which is a summary of early observations of online crowdfunding. The materials are adapted from the work of Ajay K. Agrawal (University of Toronto), Christian Catalini (MIT) and Avi Goldfarb (University of Toronto).

#1: Funding is not geographically constrained - When Sellaband offered royalty sharing to investors, more than 86% of the funds came from individuals who were more than 60 miles away from the entrepreneur, and the average distance between creators and investors was approximately 3,000 miles.

**\*2: Funding is highly skewed -** On the same platform, whereas 61% of all creators did not raise any money, 0.7% of them accounted for more than 73% of the funds raised between 2006 and 2009. Similarly, outcomes are highly skewed on Kickstarter, even conditioning the sample on successfully funded projects: 1% (10%) of projects account for 36% (63%) of funds.

lead to herding - The propensity of individual funders to invest in a project increases rapidly with accumulated capital. On Sellaband, in a given week, funders were more than twice as likely to invest in creators who reached 80% of their funding goal, relative to those who had raised only 20% of it. The acceleration is particularly strong towards the end of the fundraising campaign, similar to online lending platforms, and raises concerns of herding behavior. At the same time, projects that are eventually successful might slow down in the middle of the process because of a bystander effect - a reduction in the propensity to fund by new individuals because of the perception that the target will be reached regardless.

\*4: Friends and family funding plays a key role in the early stages of fundraising - Friends and family disproportionately invest early in the funding cycle, generating a signal for later funders through accumulated capital. The asymmetry between friends and family and others in terms of funding behavior is strongest for the first investment decision but subsequently fades as funders are able to monitor the creator's progress directly on the crowdfunding platform.

**#5: Funding follows existing agglomeration -** Despite the decoupling of funding and location, funds from crowdfunding disproportionately flow to the same regions as traditional sources of finance, perhaps due to the location of human capital, complementary assets, and access to capital for follow-on financing.

#6: Funders and creators are initially overoptimistic about outcomes - On Sellaband, after a first wave of funded artists failed to deliver a tangible return on investment, funders revised their expectations downwards. Similarly, Kickstarter recently faced pressures to adjust its guidelines after a number of high-profile projects encountered delays or failed to deliver on their initial promises. In the technology and design categories on Kickstarter, estimates suggest that more than 50% of products are delivered late.

\*7: Crowdfunding capital may substitute for traditional sources of financing - Capital from crowdfunding may substitute for alternative sources such as home equity loans. As house prices rise in a specific geographic region, making it easier for entrepreneurs to use home-equity loans as a source of financing, the number of entrepreneurs who turn to crowdfunding decreases.

"Crowdfunding has the potential to revolutionize the financing of small business, transforming millions of users of social media such as Facebook into overnight venture capitalists, and giving life to valuable business ideas that might otherwise go unfunded." - The Wall Street Journal

"Besides, isn't this the type of innovation we should be encouraging? Unlike exotic derivatives and super-fast trading algorithms, crowdfunding generates capital for job-creating small businesses." - The New York Times

"While founders raising cash from a big pool of small amounts of money are benefiting from quick access and the boost of popular interest, they are also forgoing some of the advice and experience of more traditional angel or venture-capital investors." - The Financial Times

## **Kickstarter**



## **Product Categories in Kickstarter**

Kickstarter supports almost all kinds of product categories including Art, Comics, Crafts, Dance, Design, Fashion, Film & Video, Food, Games, Journalism, Music, Photography, Publishing, Technology, and Theater.

Within each category, there are also several subcategories. For example, within the technology category, we have subcategories including gadgets, hardware, DIY electronics, flight, 3D printing, apps, camera equipment, etc.

## **Pebble Watch**

Pebble Watch was a smartwatch developed by the Pebble Technology Corporation. Funding was conducted through a Kickstarter campaign running from April 11, 2012 to May 18, 2012, which raised \$10.3 million; it was the most funded project in Kickstarter history, at the time.

Let's visit Pebble Watch's initial crowdfunding webpage to know more about here. Click <u>here</u> to go.

Recall that it is in 2012.

#### **Pebble Watch**

Che New York Eimes

GADGETWISE

A Smartwatch Gains Some Style,
but Few New Tricks



Harvard Business Review

**Pebble: Wearables Pioneer** 

## **Pebble Watch**

In 2015, Pebble launched its second generation of smartwatches: the Pebble Time and Time Steel. The devices were similarly funded through Kickstarter, raising \$20.3 million from over 75,000 backers and breaking records for the site. See the Kickstarter webpage <a href="https://example.com/here/beble/backers/">here</a>.

In 2016, Pebble shut down their subsequent Time 2 series watches and refunded Kickstarter backers, citing financial issues. It was purchased by Fitbit later.

## **Everyday Backpack**

This versatile pack was designed by photographers who felt other camera bags on the market lacked the ability for them to fit all of their other equipment. The bag comes with a number of zippered pockets and waterproof pouches to fit anything you need, as well as versatile handles and anti-theft straps. The campaign's original goal was \$500,000, but they ended up with 26,000 backers and \$6,565,782 before they packed up and went home.

See the webpage <u>here</u>.

## Kingdom Death: Monster 1.5

A board game based on the video game "Dark Souls" originally held the title for most funded board game of all time. That was until it's sequel, Kingdom Death: Monster 1.5, came to play. The game essentially involves players creating characters, crafting weapons/gear, building civilizations and fighting monsters. Kingdom Death: Monster 1.5, which raised \$1 million in the first 19 minutes it went live on Kickstarter, is expected to be available in 2020.

See the page <u>here</u>.

## Life on the Line

Cristian Barnett is a professional photographer living in Cambridge, England. Mr. Barnett was so fascinated with the Arctic Circle that in 2006 he started visiting the countries intersected by the circle. After seven years and a dozen trips to that area, he decided to create a book called *Life on the Line*, which would contain a selection of portraits he had taken over the years. See <a href="here">here</a> for more details about the project.

## "All-Or-Nothing"

Most crowdfunding platforms like Kickstarter strictly implement an "all-or-nothing" policy. That is, the creator (entrepreneur) must set up a target for the project. If the collected fund exceeds the target, the project is successful, and the creator uses the fund to run the project. Otherwise, the project fails, and all the money will be fully refunded to the investors (backers, consumers).

# LEARNING ABOUT KICKSTARTER

## **Our Mission**

We want to know more about the emerging crowdfunding industry. By doing so, we can

- (1) help entrepreneurs launch better projects and raise more funds from backers;
- (2) help crowdfunding platforms design better features and recommend better projects to backers;
- (3) help the government and public policymakers understand crowdfunding and regulate this industry.

## **Our Mission**

But how to learn about online crowdfunding? You may rely on online materials, conduct surveys, interview experts...

But these methods are really outdated...

Today, you are going to use *real data* to investigate the crowdfunding industry! This is also what data scientists are doing nowadays!

## THE DATA

I started to collect data from Kickstarter many years ago. At that time, crowdfunding was growing very fast, and many people wanted to figure out how crowdfunding really works.

I was one of the first a few people analyzing online crowdfunding using scientific methods.

## **BEFORE SEEING THE DATA...**

Please go to the Kickstarter website (<a href="https://www.kickstarter.com/">https://www.kickstarter.com/</a>), browse a few Kickstarter projects.

If you an entrepreneur trying to launch a successful crowdfunding campaign, what do you want to learn from Kickstarter?

To answer these questions, which data do you need to collect?

## The Data is Available at the Course Website

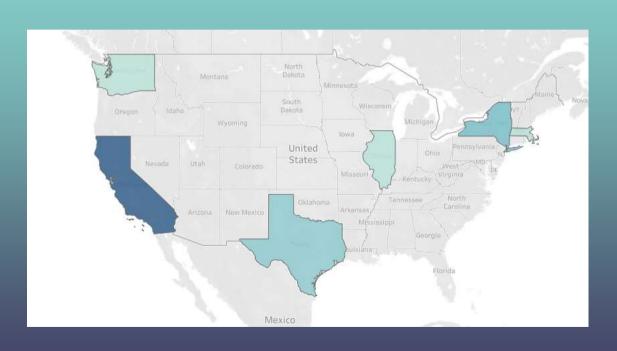
## DATA

The dataset was scraped from Kickstarter, the largest online crowdfunding website.

It covers Kickstarter's technology category.

The data is collected from the following US markets: California, Illinois, Massachusetts, New York, Texas, and Washington State.

#### Location



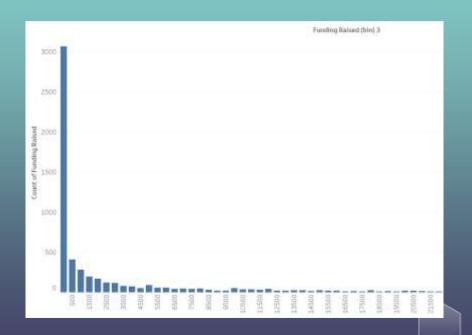
The dataset covers projects from six US states: California (CA), New York (NY), Texas (TX), Massachusetts (MA), Washington (WA), and Illinois (IL).

## Subtype

Ţ	echnology	Hardware		Web		Software		
A	pps	Gadgets	Weara	ibles	Sound	Robot	ts	30 Printing
					Camera	Space		
			DIY Electronics		Equipmen			spaces

## **Total Funding Raised**

The total funding raised by an individual project, measured by US\$. You can see from the histogram that the total funding raised is really an L-shaped distribution: Most projects received almost \$0 while some projects are very successful.



## Log Transformation

When we take the logarithm of the total funding raised, the distribution looks more like a normal distribution.

```
mydata$LogFundingRaised = log(mydata$FundingRaised
+ 1)
hist(mydata$LogFundingRaised)
```

Why we use Funding Raised + 1?

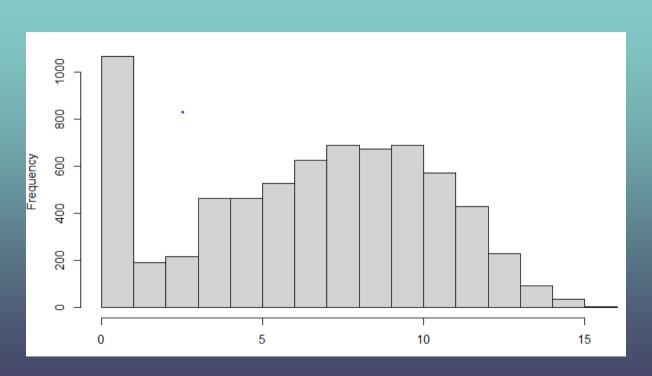
## Log Transformation

When we take the logarithm of the total funding raised, the distribution looks more like a normal distribution.

```
mydata$LogFundingRaised = log(mydata$FundingRaised
+ 1)
hist(mydata$LogFundingRaised)
```

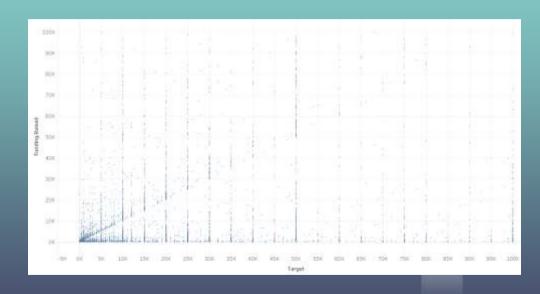
Why we use Funding Raised + 1?

# Log Transformation



## **Target**

At Kickstarter, each entrepreneur needs to specify a target for the project. The project is successful when the funds raised exceeds the target. Otherwise, the project fails and all the funds will be returned to the consumers.



## Other Measures of Project Result

Outcome: Whether or not the project succeeded. It is a binary variable (1 = success, 0 = failure).

Backers: Number of people supporting the project. If you divide funding raised by the number of backers, you will get the average fund contributed by a backer.

## **Entrepreneurs' Personal History**

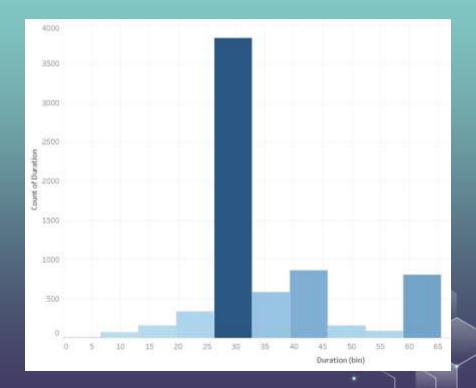
Created: Number of projects created by the same entrepreneur in the past. For example, 4 means the same entrepreneur had already created another 4 projects on Kickstarter.

Backed: Number of projects backed by the same entrepreneur in the past (i.e., the entrepreneur supporting others' projects on Kickstarter).

FbNumber: Number of Facebook friends the entrepreneur has.

## **Duration**

The duration of a project's fund raising period (in days). Most projects have a duration of around one month.



## Gender 0.40 0.35 0.30 Avg. Outcome 0.25 0.20 0.15 0.10 0.05 0.00

## Gender

In the dataset, we have three genders: males, female, and unknown. The gender is obtained by analyzing the creators' first name. Unknow refers to the case in which the name cannot be identified (e.g., a team name such as "marketing").

## **Some Other Variables**

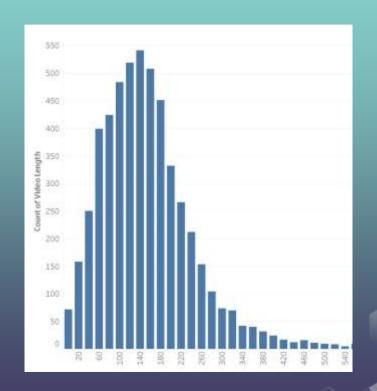
Video: Whether or not the project has a video. In the dataset, 76% of the projects have a video. Here, 1 means has video and 0 means no video.

Human: Whether or not the project's video features human-beings (usually the entrepreneurs themselves). 1 means has human-beings and 0 means no human-beings. This variable is set to 0 is the project does not have a video.

Computer: Whether or not the project's video features a computer. 1 means has computers and 0 means no computers. This variable is set to 0 is the project does not have a video.

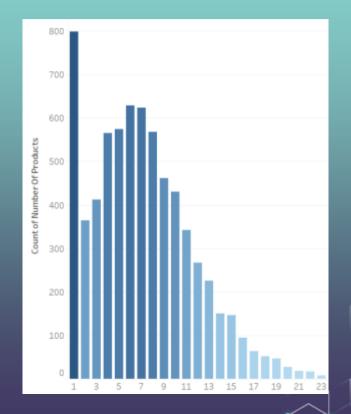
## **Some Other Variables**

Video length: the duration of the video measured in seconds. For project without a video this variable is set to 0. The following is a histogram of video length (for projects with a video).



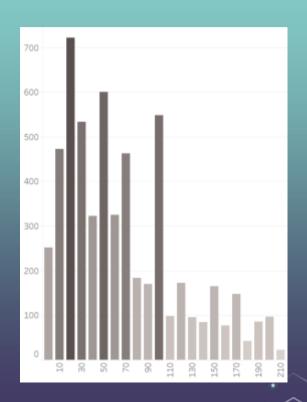
# Number of Products: In a project the entrepreneurs often offer consumers a number of products to choose from. This variable measures how many products are offered in the project.

## **Some Other Variables**



## **Some Other Variables**

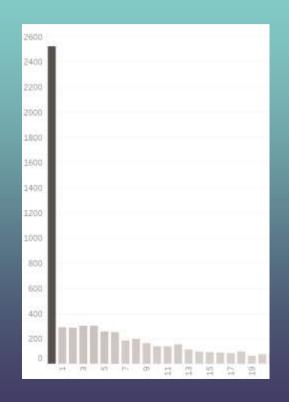
Price: A project may offer multiple products with different prices. Here, price means the median price among all product offerings.



## **Some Other Variables**

#### Photo Number:

Entrepreneurs often upload photos to their project description. This variable measures the number of photos uploaded to the project webpage.



## **Audio Measures**

For projects with a video, we analyze their audio information:

**Energy**: Whether or not the audio pitch sounds energetic. A large number of an energetic audio pitch.

Content: Whether or not the audio pitch shows signs of content.

Upset: Whether or not the audio pitch shows signs of upset.

Angry: Whether or not the audio pitch shows signs of anger.

MaxAmpVol: The max sound volume. A greater number means louder sound.

#### What should we do?

Use the data to provide recommendations for the platform or the entrepreneurs. You can focus on anything that can be helpful for the platform or the entrepreneurs:

- How do male and female entrepreneurs behave differently on Kickstarter?
   (e.g., compared to females, males may be too aggressive in setting high targets.)
- Which type of video is most productive in terms of generating funds? (e.g., is having a lengthy video always beneficial?)
- What makes a successful crowdfunding project?

## What should we do?

Each group should only ask one or two research questions in your project. Quality beats quantity. Choose the right data analysis methods and come up with a good answer to your questions, with implications for platforms or entrepreneurs.

Think about the papers we presented in class. They are your examples.

## You need to submit:

To save your time, you only need to submit a few pages of slides (no more than 20 slides) to Moodle covering your research question(s), data analysis (e.g., regression equations), findings, and implications. No reports/presentations are needed!

Deadline: Class A: Dec 26, 5:00 pm

Class B: Dec 26, 9:30 pm

Class C: Dec 27, 9:30 pm