



# DS-UA 112

## Introduction to Data Science

Week 9: Lecture 1

Models - Working with Loss Functions





How can we fill in the  
missing pieces in a  
description of some  
population?

# DS-UA 112

## Introduction to Data Science

### Week 9: Lecture 1

### Models - Working with Loss Functions

*Adapted from Dudoit, Nolan, Gonzalez*



# Announcements

- ▶ Please check Week 9 agenda on NYU Classes
  - ▶ Lab 7
    - ▶ Due on Friday March 27 at 12PM
  - ▶ Project 1
    - ▶ Due on Monday April 6 at 12PM
  - ▶ Survey

[https://nyu.qualtrics.com/jfe/form/SV\\_3DCWUa4yc08L0wt](https://nyu.qualtrics.com/jfe/form/SV_3DCWUa4yc08L0wt)



# Announcements

- ▶ Recordings
  - ▶ Lecture
  - ▶ Section
  - ▶ Office Hours
- ▶ Media Gallery
- ▶ Zoom

Check @449 on  
Piazza for more  
information



# Review

- ▶ Text as Data
  - ▶ We studied Huckleberry Finn by Mark Twain and Little Women by Louisa May Alcott
- ▶ We tried to understand aspects of the novels through numbers

## Questions

- ▶ Can we make deductions about the characters or the relationships between characters by the occurrence of their name?

# Review

- ▶ Text as Data
  - ▶ We studied Huckleberry Finn by Mark Twain and Little Women by Louisa May Alcott
- ▶ We tried to understand aspects of the novels through numbers

## Questions

- ▶ Can we distinguish the novel by the length of sentences?
- ▶ Can we distinguish the novel by the length of words?

# Review

Zipf Rule

- ▶ Text as Data
  - ▶ We studied Huckleberry Finn by Mark Twain and Little Women by Louisa May Alcott
- ▶ We tried to understand aspects of the novels through numbers

## Questions

- ▶ What is the relationship between the frequency of words and the rank of a words ordered by number of occurrences?

# Review

- ▶ Define frequency to be the proportion of times a word appears in the text
- ▶ Define rank to be the ranking of words in descending order according to occurrence

$$Frequency = \frac{1}{Rank}$$

- ▶ Zipf Rule
  - ▶ Proposed by linguist George Kingsley Zipf to describe patterns in language
  - ▶ The rule relates **frequency** and **rank**



## Review

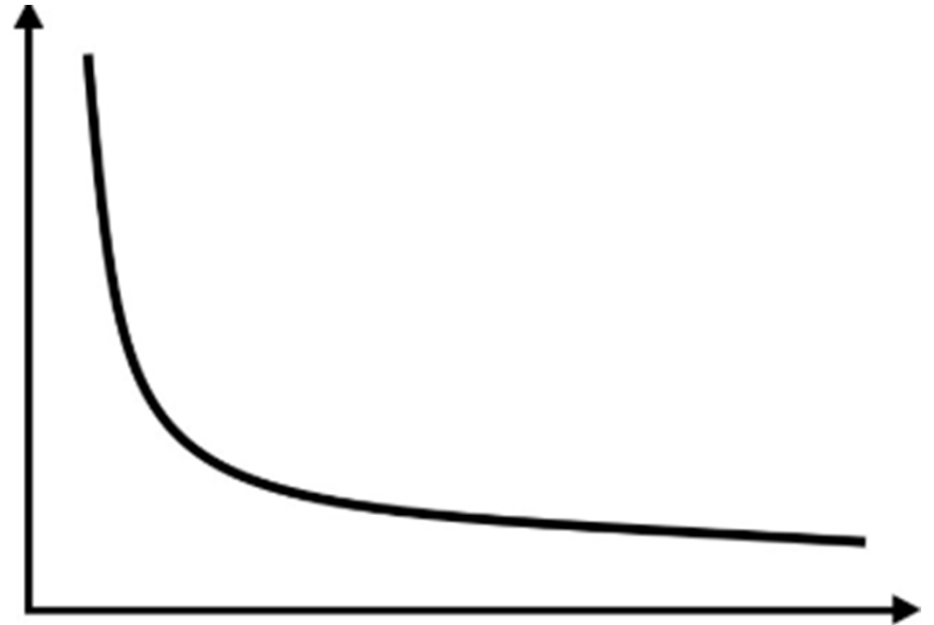
- ▶ Define frequency to be the proportion of times a word appears in the text
- ▶ Define rank to be the ranking of words in descending order according to occurrence

$$Frequency = \frac{1}{Rank}$$

- ▶ So if the most common word has frequency 0.2, then the second most common would have frequency  $0.2 / 2$ , the third most common would have frequency  $0.2 / 3$ , ...

# Review

- Zipf rule describes a **power law** meaning a relationship between two quantitative variables involves taking a power transformation



*Dependent Variable =  
(Independent Variable)<sup>power</sup>*

# Review

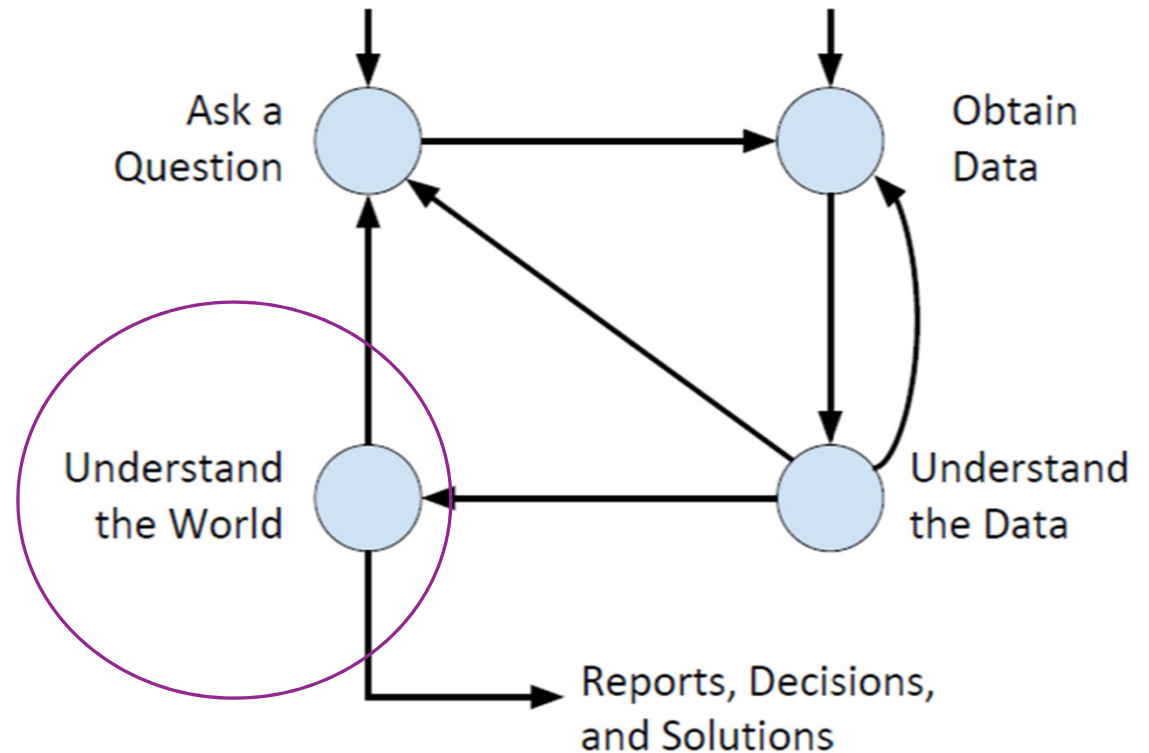
- ▶ Zipf rule describes a **power law** meaning a relationship between two quantitative variables involves taking a power transformation

*Dependent Variable =  
(Independent Variable) <sup>power</sup>*

- ▶ Power laws occur in many contexts to model
  - ▶ Population of Cities
  - ▶ Distribution of Wealth
  - ▶ Size of Companies

# Agenda

- Modelling
  - Estimation
  - Prediction
  - Inference



# Models

15% of bill is customary amount for tip in US

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
...	...	...	...	...	...	...	...
241	22.67	2.00	Male	Yes	Sat	Dinner	2
242	17.82	1.75	Male	No	Sat	Dinner	2
243	18.78	3.00	Female	No	Thur	Dinner	2

- ▶ Tips for waiters and waitresses
- ▶ Estimation: Is 15% the average amount for a tip in US?
- ▶ Prediction: Could we use information like day and time to guess the tip?
- ▶ Inference: If the average is 15% then what is the chance that the tip exceeds 25%?

## Exercise

### Completing the Square

- Show that the expression

$$x^2 + 6x = -2$$

is equivalent to the expression

$$(x+3)^2 = 7$$

- Suppose we have the expression

$$x^2 + bx = c$$

Find  $e$  and  $f$  such that

$$(x + e)^2 = f$$

# Models

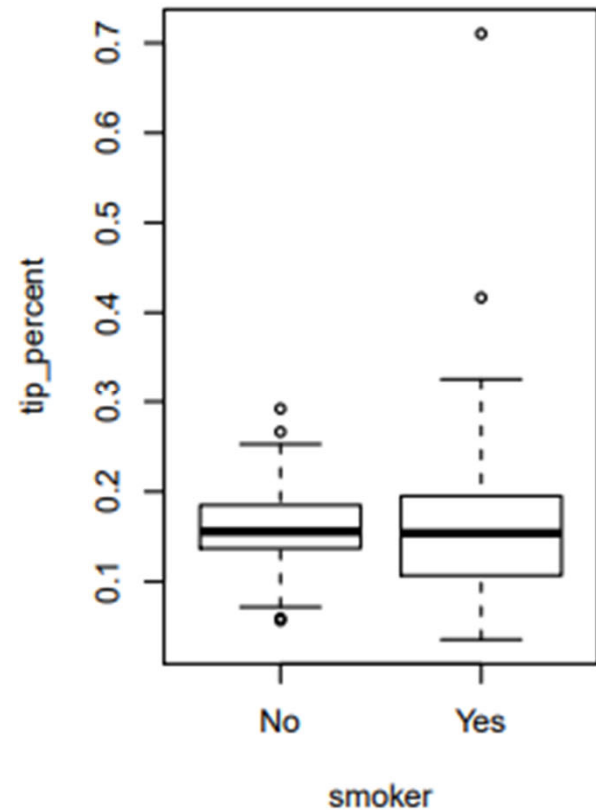
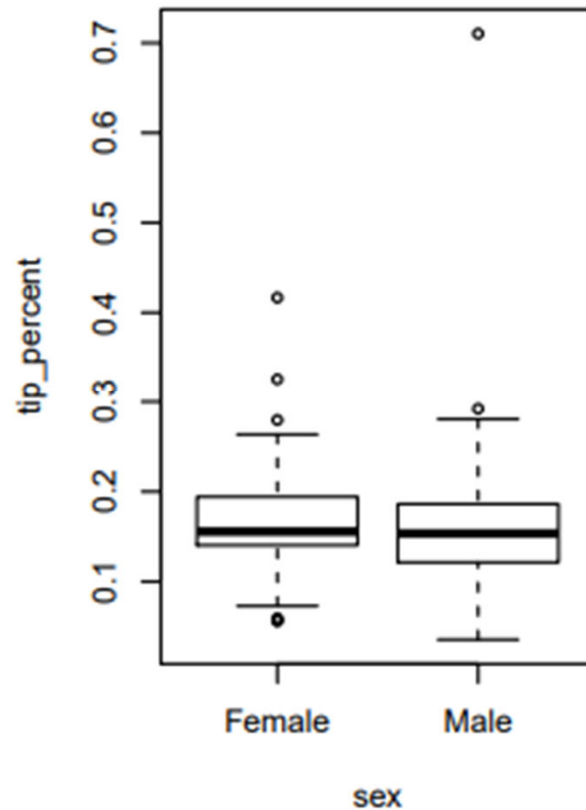
15% of bill is customary amount for tip in US

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3
...	...	...	...	...	...	...	...
241	22.67	2.00	Male	Yes	Sat	Dinner	2
242	17.82	1.75	Male	No	Sat	Dinner	2
243	18.78	3.00	Female	No	Thur	Dinner	2

- ▶ The mean tip percentage is 16.08%
- ▶ Most tips are between 10 and 20% with a few outlying large tips. Maximum of 70%.
- ▶ The tip percentage does not appear, however, to vary much with variables such as sex, smoker, day, time, and size.

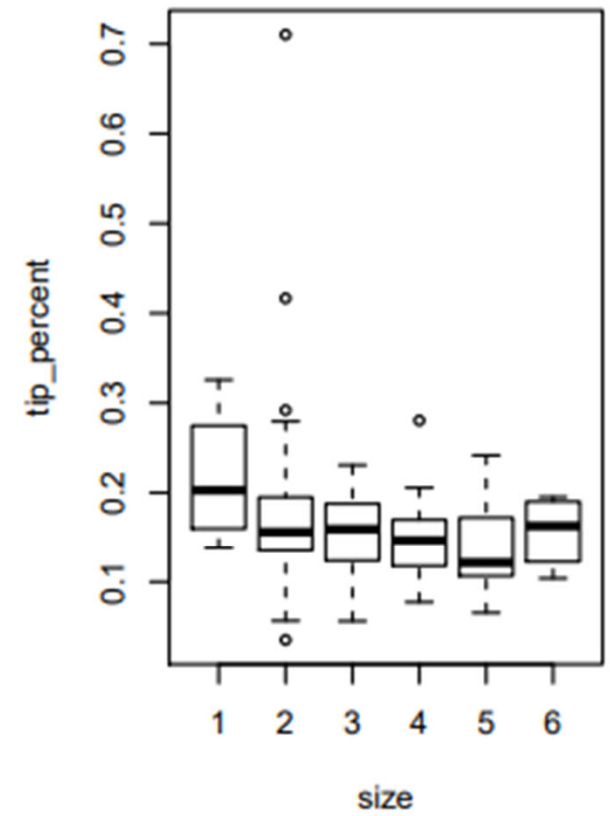
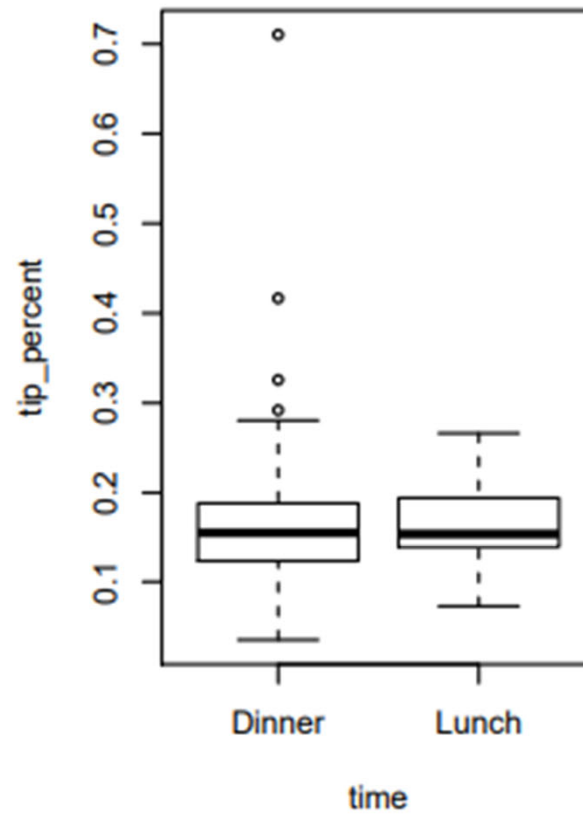
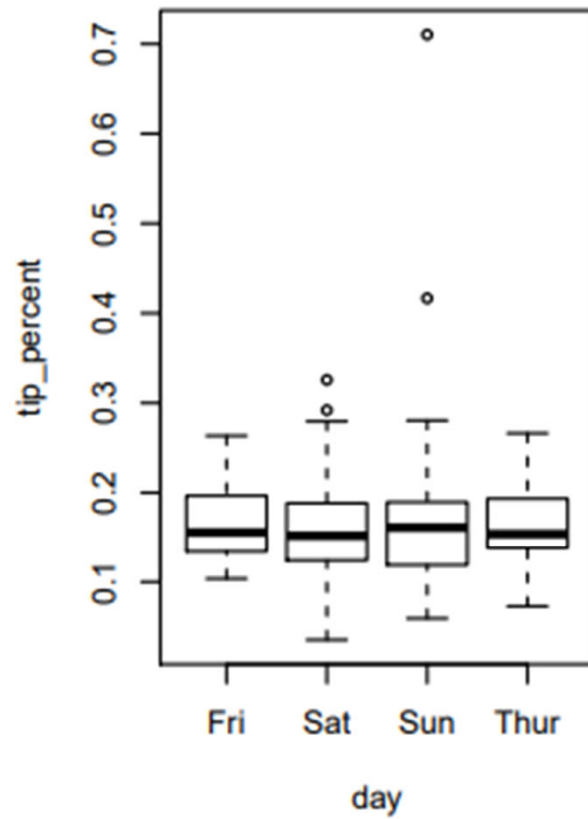
# Models

- ▶ Since sex, smoker, day, time, and size can be thought of as qualitative variables, we can produce side-by-side boxplots to investigate the relationships
- ▶ If the range of values in the tips does not significantly differ between variables, then we do not have a relationship

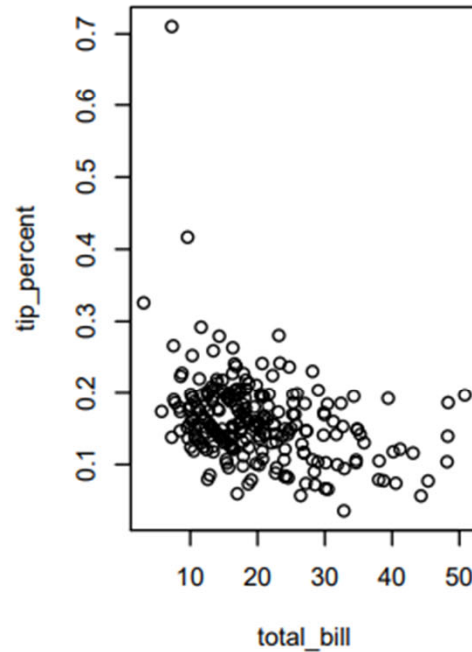




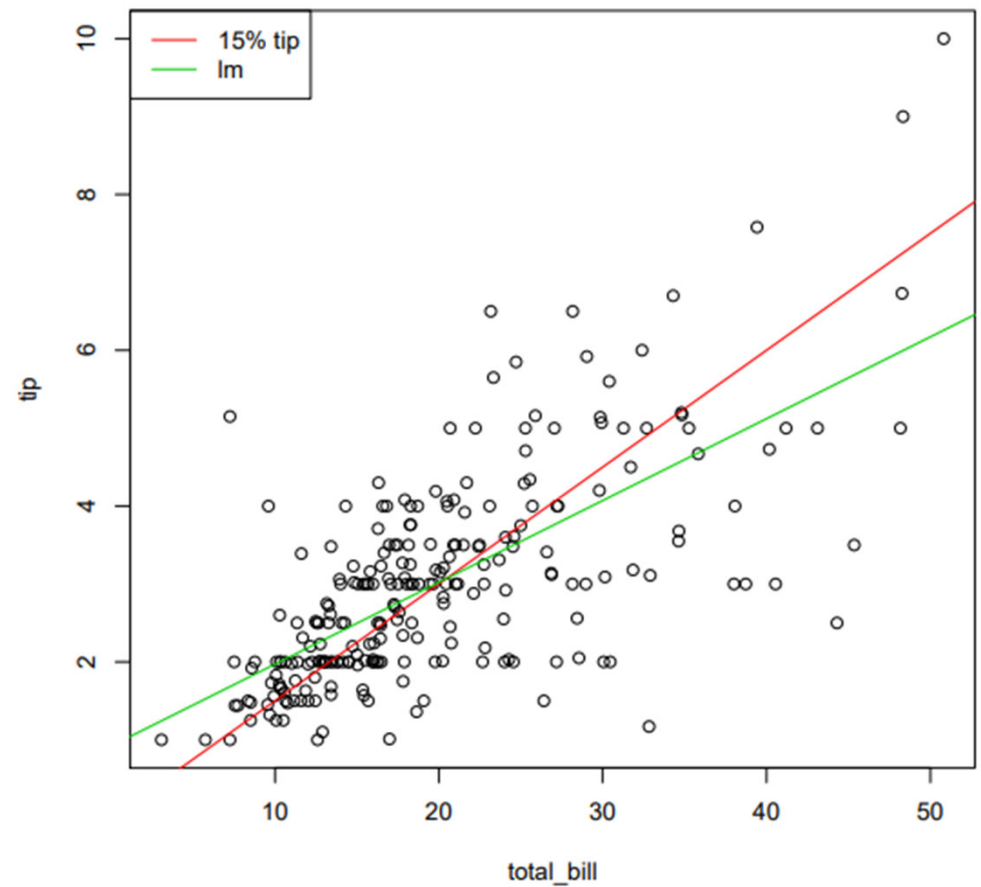
# Models



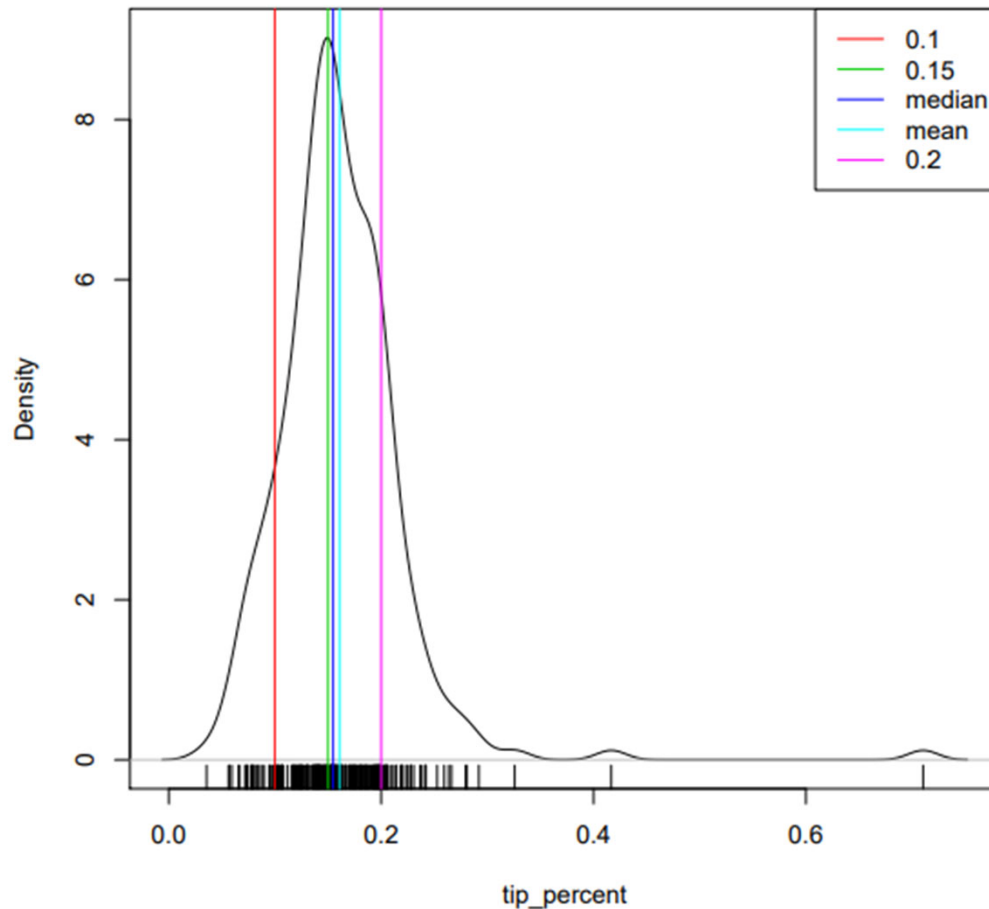
# Models



- Removing outlier leads to a linear trend
- The slope estimates the percent tip

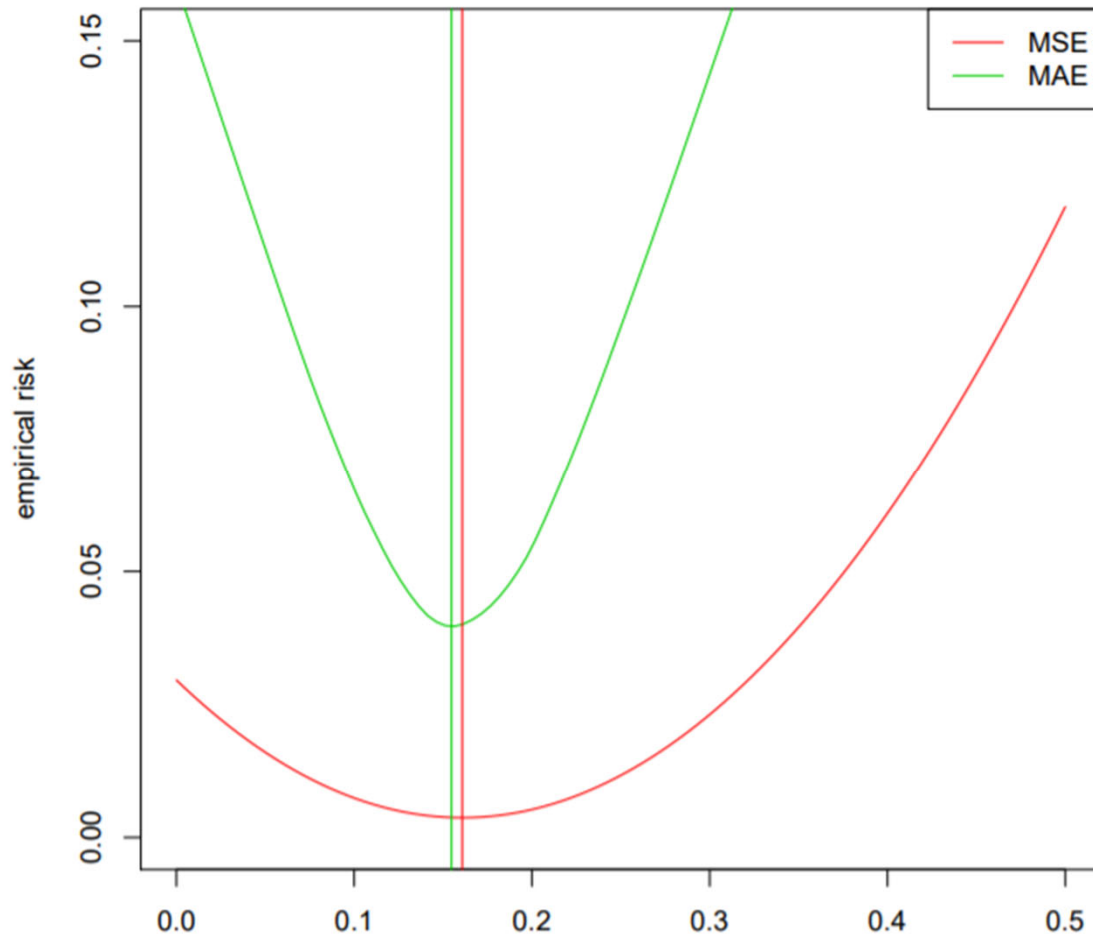


# Models



- ▶ We have different approaches to summarizing the tip as a percentage of the bill
- ▶ The mean captures the average value among the tips
- ▶ The median captures the middle value among the tips

# Models



- ▶ MSE stands for Mean Square Error
- ▶ MAE stands for Mean Absolute Error
- ▶ Note that MSE involves squaring numbers. Since we are squaring numbers between 0 and 1 in the tips dataset, MSE lies below MAE

# Summary

- ▶ Power Law
  - ▶ Zipf Rule
- ▶ Models
  - ▶ Estimation
  - ▶ Prediction
  - ▶ Inference
- ▶ Loss Functions

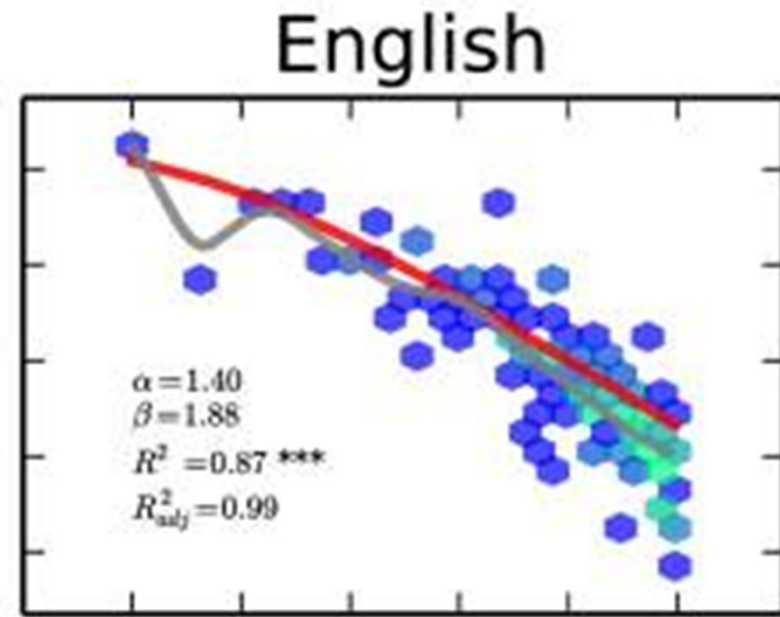
## Goals

- ▶ What is square loss and absolute loss?
- ▶ How can derivatives help use to determine the parameters in loss functions?

# Questions

- ▶ Questions on Piazza?
  - ▶ Please provide your feedback along with questions
- ▶ Question for You!

Where else can you find the Zipf Rule?



# Questions

- ▶ Questions on Piazza?
  - ▶ Please provide your feedback along with questions
- ▶ Question for You!

Where else can you find the Zipf Rule?

