Discussion Section

Week 1

Bret Stevens

August 8, 2019

University of California, Davis

Introduction

Hi.

Introduction

Hi. I'm Bret.

Introduction

Hi. I'm Bret. I'm your TA.

Introduction

Hi. I'm Bret. I'm your TA.

What are we going to do in these discussion sections?

Introduction

Hi. I'm Bret. I'm your TA.

What are we going to do in these discussion sections?

(1) Review the week's material
(2) Discuss coding/homework
(3) Answer any lingering questions you have from lecture or the homework

Introduction

Hi. I'm Bret. I'm your TA.

What are we going to do in these discussion sections?

(1) Review the week's material
(2) Discuss coding/homework
(3) Answer any lingering questions you have from lecture or the homework

Generally, we'll do things in that order. Hopefully I can answer your questions in my reviews.

Introduction

That’s what I will be doing.

Introduction

That’s what I will be doing. What should you be doing?

Introduction

That’s what I will be doing. What should you be doing?

(1) Taking/amending your notes from lecture
(2) Working through the homework problems with me
(3) Bringing and thinking about good questions to ask in discussion (don’t be selfish)

That's what I will be doing. What should you be doing?

(1) Taking/amending your notes from lecture
(2) Working through the homework problems with me
(3) Bringing and thinking about good questions to ask in discussion (don't be selfish)

If you own a portable computer, please bring it to class. It will be much easier to follow along in the coding sections. If you do not have a portable computer, then please take notes during the coding section and feel free to ask me to send you the code after class.

Lecture Review

Statistics Review

Random Variables

- Just something that is not a constant
- Most things in life are random variables
- We can't know what it will be before we observe it
- Can be discrete or continuous
- It typically has some sort of distribution
- Random variables are kind of like functions

Statistics Review

Random Variables

- Just something that is not a constant
- Most things in life are random variables
- We can't know what it will be before we observe it
- Can be discrete or continuous
- It typically has some sort of distribution
- Random variables are kind of like functions

Probability Distributions

- Describes a random variable
- For a discrete variable its called a "Probability Mass Function"
- For a continuous variable its called a "Probability Distribution Function"
- These things are actual functions, just like you would use in a math class

Statistics Review

Random Variables

- Just something that is not a constant
- Most things in life are random variables
- We can't know what it will be before we observe it
- Can be discrete or continuous
- It typically has some sort of distribution
- Random variables are kind of like functions

Probability Distributions

- Describes a random variable
- For a discrete variable its called a "Probability Mass Function"
- For a continuous variable its called a "Probability Distribution Function"
- These things are actual functions, just like you would use in a math class
- Draw picture

Statistics Review

Population vs Sample

- First define a group of interest
    - Can be “all students in the world” or “All students in ARE 106 SSII”
- Population is all of the things in that group, whereas a sample is a subset of the things in that group
- Since its often hard to get data on your population, we typically work with samples
- Statisticians like to differentiate statistics taken from these different types of groups
    - A statistic from the population is called a *population parameter*
    - A statistic from a sample is called a *sample statistic*
- Not all samples are created equal, we typically like to work with “random samples”

Population vs Sample

- First define a group of interest
  - Can be "all students in the world" or "All students in ARE 106 SSII"
- Population is all of the things in that group, whereas a sample is a subset of the things in that group
- Since its often hard to get data on your population, we typically work with samples
- Statisticians like to differentiate statistics taken from these different types of groups
  - A statistic from the population is called a *population parameter*
  - A statistic from a sample is called a *sample statistic*
- Not all samples are created equal, we typically like to work with "random samples"
- Give example

Statistics Review

Expectations

- The expected value of a random variable is the *population* average
  - This means that if we have a sample, the sample average is **not** the expectation of that variable (LLN)
- Properties
  - a is a constant (like 3)
  - X and Y are variables (like height and GPA)

$$E[a] = a \tag{1}$$
$$E[aX] = aE[X] \tag{2}$$
$$E[X + Y] = E[X] + E[Y] \tag{3}$$
$$E[XY] \neq E[X]E[Y] \tag{4}$$
$$E[g(X)] \neq g(E[X]) \tag{5}$$

- Conditional Expectations are just averages for a subset of the population that has a particular trait, write it as $E[Y|X]$

Statistics Review

Variance

- Measures how spread out a variable is
- Population variance - σ^2

$$Var(X) = E[X^2] - E[X]^2$$

- Sample variance - $\hat{\sigma}^2$

$$\hat{Var}(X) = \frac{\sum_i^N (X_i - \bar{X})^2}{N-1}$$

- Tricks

$$Var(aX + b) = a^2 Var(X)$$

- Standard Deviation - σ - the square root of the population variance
- Standard Error - $\hat{\sigma}$ or s - The square root of the sample variance

Covariance

- Describes how two variables move together
- Populaiton covariance

$$Cov(X,Y) = E[XY] - E[X]E[Y]$$

- Notice that this will be the same thing no matter the order
- What happens if we find the covariance of a variable with itself?
- Sample covariance

$$\frac{\sum_i^N (X_i - \bar{X})(Y_i - \bar{Y})}{N-1}$$

- Tricks

$$Cov(aX+b, dY+e) = adCov(X,Y)$$

Correlation

Correlation is just a normalized covariance

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

Corr(X,Y) $\in [-1, 1]$ where 0 means low relationship and 1 means strong relationship

Regression Review

What is a regression and when can I use one?

- Let’s say we have some data and want to answer a question with that data
- If our question is a long the lines of “Can I predict the outcome of something with the data I have?” or “How does this affect that?” we can use a regression!
- Regressions are very good at predicting outcomes, but can be used to understand relationships between variables as well
- Understanding these relationships can be tricky and to do so, we often must augment simple regression analysis to overcome some of the problems

Regression Review

What is a regression and when can I use one?

- Let's say we have some data and want to answer a question with that data
- If our question is a long the lines of "Can I predict the outcome of something with the data I have?" or "How does this affect that?" we can use a regression!
- Regressions are very good at predicting outcomes, but can be used to understand relationships between variables as well
- Understanding these relationships can be tricky and to do so, we often must augment simple regression analysis to overcome some of the problems
- Draw picture

Regression Review

- A regression is basically just a line that runs through the data

$$y = mX + b$$
$$\hat{y} = \beta_0 + \beta X_i$$

- The true y_i are typically not a simple line
- Thus, we can can say that each y_i is equal to the predicted value, plus some error

$$y_i = \beta_0 + \beta X_i + \epsilon_i$$

- Technically, we're not wrong

Ordinary Least Squares (OLS)

- We can't just draw any line through some data points and call it good
- To find the line the best predicts the outcome variable, based on our data, we use OLS
- OLS is simply an algorithm that minimizes the size of the error term
- Particularly, it minimizes the "sum of squared errors"

$$\min \sum_{i}^{N} \epsilon_i^2$$

$$\Rightarrow \min \sum_{i}^{N} (y_i - \beta_0 - \beta X_i)^2$$

- Do you all understand how we're doing that substitution?

Regression Review

Ordinary Least Squares (OLS)

- We can't just draw any line through some data points and call it good
- To find the line the best predicts the outcome variable, based on our data, we use OLS
- OLS is simply an algorithm that minimizes the size of the error term
- Particularly, it minimizes the "sum of squared errors"

$$\min \sum_{i}^{N} \epsilon_i^2$$

$$\Rightarrow \min \sum_{i}^{N} (y_i - \beta_0 - \beta X_i)^2$$

- Do you all understand how we're doing that substitution?
- Do minimization on board

Regression Review

So I did OLS, so what?

- From simple OLS we get two things, β_0 and β
  - This is your slope and intercept of your line
- How do we interpret these things?
  - $\hat{\beta}$ - A 1 unit change in X is correlated with a $\hat{\beta}$ unit change in y
  - $\hat{\beta}_0$ - Given that X_i is 0, the expected value of y is $\hat{\beta}_0$

Goodness of Fit

- A regression is only as good as the data you give it, and the model you tell it
- We can check how good our regression is at predicting the outcome variable by looking at its "goodness of fit"
- We call it this because we are "fitting" the line to the data
- There are many ways to see how good a regression is, the most common is R^2
- The important statistics to know for goodness of fit are:

Regression Review

R^2

- TSS - Total Sum of Squares - $\sum_i (y_i - \bar{y})^2$
- ESS (SSR) - Explained Sum of Squares (Regression Sum of Squares) - $\sum_i (\hat{y} - \bar{y})^2$
- RSS (SSE) - Residual Sum of Squares (Error Sum of Squares) - $\sum_i (y_i - \hat{y}_i)^2$

$$R^2 = 1 - \frac{RSS}{TSS}$$

$$= \frac{ESS}{TSS}$$

Python Review

Does everyone have Anaconda installed? Yes No

Does everyone have Anaconda installed? Yes No

Now that everyone has it installed, lets open a Jupyter Notebook

Nice

Back

Get it together >:(

Back