STAT 311: Introduction

Y. Samuel Wang

Summer 2016

Course Goals

- Build statistical intuition and quantitative literacy
- Assess and interpret statistical studies and methods
- Carry out basic data analysis using statistical software
- Understand basic probability concepts

Course Overview

Course Schedule:

- Lecture: MWF 8:30 - 9:20
- Quiz Sections (Lab): TR 8:30 - 9:20 or 9:30- 10:20

Course Overview

Course Schedule:

- Lecture: MWF 8:30 - 9:20
- Quiz Sections (Lab): TR 8:30 - 9:20 or 9:30- 10:20

Grading:

Weekly Homework	35%
Labs	15%
Midterm	20%
Final Exam	30%

Ways to get Help

- Office Hours
  - Sam: Wednesday 9:30 - 11:30 and by appointment
  - Shiqing: TBD
- Catalyst discussion board
- Classmates

Personal Introductions

- Name
- Major (or intended major)
- What you hope to get out of this class
- Interesting fact

Why study statistics?

Hal Varian, Chief Economist at Google:

I keep saying the sexy job in the next ten years will be statisticians. People think Im joking, but who wouldve guessed that computer engineers wouldve been the sexy job of the 1990s?

Why study statistics?

Hal Varian, Chief Economist at Google:

I keep saying the sexy job in the next ten years will be statisticians. People think Im joking, but who wouldve guessed that computer engineers wouldve been the sexy job of the 1990s?

The ability to take data, to be able to understand it, to process it, to extract value from it, to visualize it, to communicate its going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it.

Probability vs Statistics

Probability

- Purely mathematical framework
- Describes long run behavior
- Convenient way to model the real world
- Examples: Rolling dice, flipping coins, Normal distribution...

Statistics

- Modeling and analyzing real data
- Utilizes probability models
- May not have one correct answer

Regression: Predicting Home Prices

$$\text{Price} \approx \beta_1 X_1 + \beta_2 X_2 + \dots \tag{1}$$

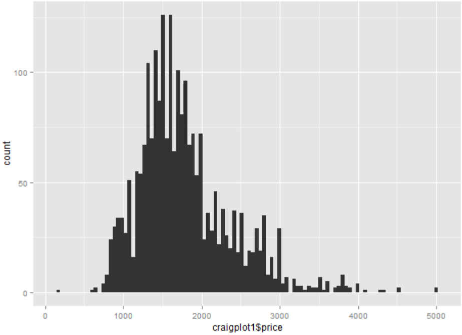
Regression: Predicting Home Prices

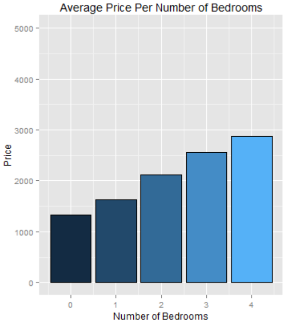
$$\text{Price} \approx \beta_1 X_1 + \beta_2 X_2 + \dots \tag{1}$$

- What factors should I include?
- What effect do we think those factors will have?
- How might I gather the data
- What should I be careful about?

Regression: Predicting Home Prices

Data Analysis taken from: http://www.racketracer.com/ Dec 2014




Average Price Per Number of Bedrooms

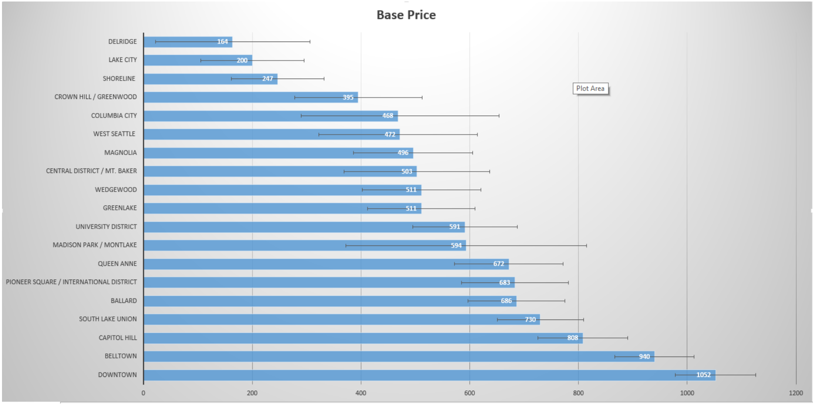
Regression: Predicting Home Prices

Data Analysis taken from: http://www.racketracer.com/ Dec 2014

$$\begin{aligned} \text{Price} \approx \text{Neighborhood Intercept} \\ +69.68 \times \text{FT}^2 \\ +322 \times \text{BTHRM} \\ +107 \times \text{BDRM} \end{aligned} \tag{2}$$

Regression: Predicting Home Prices

Data Analysis taken from: http://www.racketracer.com/ Dec 2014



Regression: Predicting Home Prices

Data Analysis taken from: http://www.racketracer.com/ Dec 2014

Each picture posted increased the price by $9 and each additional line in the add increased the price by $2.83.

Regression: Predicting Home Prices

Data Analysis taken from: http://www.racketracer.com/ Dec 2014

Each picture posted increased the price by $9 and each additional line in the add increased the price by $2.83.

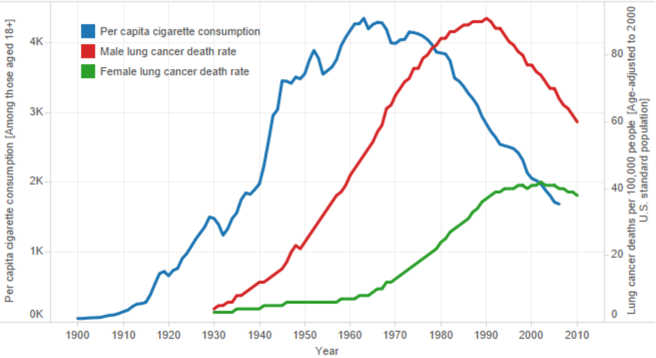
What other things would we want to predict?

Hypothesis Testing: Smoking vs Cancer

Does smoking cause cancer?

Scroll over trend lines to see data. Use the controls in the top left to zoom in or out and to reset the graph.

Trends in Tobacco Use and Lung Cancer Death Rates in the U.S.



Death rates source: US Mortality Data, 1960-2010, US Mortality Volumes, 1930-1959, National Center for Health Statistics, Centers for Disease Control and Prevention.
Cigarette consumption source: US Department of Agriculture, 1900-2007.

Hypothesis Testing: Smoking vs Cancer

- A few small studies released in 1920s-1940s
- Asked for self reported habits
- Article published in 1954 in JAMA by Hammond and Horn
- 188,000 men 50-69 years old in 10 states

Hypothesis Testing: Smoking vs Cancer

- A few small studies released in 1920s-1940s
- Asked for self reported habits
- Article published in 1954 in JAMA by Hammond and Horn
- 188,000 men 50-69 years old in 10 states

How might we have addressed the question?

- How can we make a decision?
- What type of data should I gather?
- How can we gather the data?
- What types of issues might make this more complicated?
- What kind of conclusions can I draw?

Hypothesis Testing: Smoking vs Cancer

It was found that men with a history of regular cigarette smoking have a considerably higher death rate than men who have never smoked or men who have smoked only cigars or pipes

– Hammond and Horn 1954

What other questions are similar but unanswered?

Probability: How to lose less money

You're playing poker and your opponent has just bet a large sum of money. Should you call?

Probability: How to lose less money

You're playing poker and your opponent has just bet a large sum of money. Should you call?

- What pieces of information are important?
- How can we use those pieces of information to make a decision?
- How can I decide whether the decision is good or bad?

Probability: How to lose less money

What are other areas where we can use data to drive our decisions?

Reminders

- Bring laptop to class if you have one