STAT 215
Statistical Modeling

Colin Reimer Dawson

Oberlin College

August 29, 2017

Outline

Statistics: An Alternative to "Alternative Facts"

"You're saying it's a falsehood. ... Sean Spicer, our press secretary, gave alternative facts to that." – Kellyanne Conway, White House counselor, on *Meet the Press*, 1/22/2017

"You know, the very powerful and the very stupid have one thing in common," the Doctor said. "They don't alter their views to fit the facts. They alter the facts to fit their views." – Doctor Who, *The Face of Evil, Part 4, aired 1977*

"In God we trust, all others must bring data." – attributed to statistician W. Edwards Deming, 1900-1993.

Statistics: A Highly Employable Skill

"[A] new analysis of help-wanted postings for entry-level jobs suggests that [liberal arts] graduates can improve their job prospects markedly by acquiring a small level of proficiency in one of eight specific skill sets, such as social media or data analysis. For example, the analysis found an additional 137,000 entry-level jobs for liberal-arts graduates who had data-analysis or management skills. It also found that such data-analysis jobs paid an average of $12,700 above the average salary for jobs traditionally open to liberal-arts graduates without such skills." – Chronicle of Higher Ed., 6/9/16

Defensive Statistical Literacy

"There are three kinds of lies: lies, damned lies, and statistics." – Unknown (questionably attributed by Mark Twain to Benjamin D'Israeli)

- An overarching course goal: become a literate *consumer* of statistics

On the Web

- Course Website: `http://colindawson.net/stat215`
- Syllabus, slides, handouts, homework, labs, demos available there
- Exception: HW/Lab Solutions (on Blackboard)
- Also on Blackboard: electronic submission of assignments

Course Outline

Part I: Intro Material (5 weeks)

- Data Collection, Structure of Data ($\sim$1.5 weeks)
- Data Description and Visualization ($\sim$0.5 weeks)
- Statistical Inference: Conceptual Foundations ($\sim$3 weeks)

Part II: Statistical Modeling (8 weeks)

- Linear Regression ($\sim$ 3 weeks)
- Logistic Regression ($\sim$ 2 weeks)
- Design of Experiments/Analysis of Variance ($\sim$ 2 weeks)
- Other topics / catch up ($\sim$ 1 week)

Also:

- Computational Skills for Data Analysis (throughout)

Graded Components

- (20%) Weekly(ish) Quizzes $\longrightarrow$ (see replacement quiz policy)
- (10%) Homework + lab problems (graded for effort/completion only)
- (30%) 2 In-Class Exams $\longrightarrow$ (also see replacement quiz policy)
- (20%) 2 Individual mini-projects
- (20%) Group Final Project

Models are...

- simplifications
- approximations
- not perfectly correct
- useful for a particular purpose

All models are wrong but some models are useful.

Now it would be very remarkable if any system existing in the real world could be exactly represented by any simple model. However, cunningly chosen parsimonious models often do provide remarkably useful approximations. For example, the law $PV = RT$ relating pressure P, volume V and temperature T of an "ideal" gas via a constant R is not exactly true for any real gas, but it frequently provides a useful approximation and furthermore its structure is informative since it springs from a physical view of the behavior of gas molecules.

For such a model there is no need to ask the question "Is the model true?". If "truth" is to be the "whole truth" the answer must be "No". The only question of interest is "Is the model illuminating and useful?". — George Box, 1978

Data: Numbers With a "Story"

$$\text{DATA} = \text{PATTERN} + \text{IDIOSYNCRACIES}$$

How do we decide what "the pattern" is? This, in a nutshell, is the project of modeling

Purposes of Statistical Models

1. Making predictions
2. Understanding relationships

2*. Assessing differences

The Project, More Formally

Find a relationship between a **response variable** (Y) and one or more **predictor/explanatory variables**, $X_1, \ldots, X_k$.

$$Y = f(X) + \varepsilon$$

$$\text{DATA} = \text{PATTERN} + \text{IDIOSYNCRACIES}$$

The Process of Statistical Modeling

1. **Choose** — Pick a form (or forms) for the model (or models)
2. **Fit** — Estimate parameters (if any)
3. **Assess** — Is the model adequate? Could it be simpler? Are conditions met?
4. **Use** — Answer the question of interest (e.g., make predictions)

Example: Sleep and Caffeine

A sample of 24 adults are randomly divided equally into two groups and given a list of 24 words to memorize. During a break, one group takes a 90-minute nap while another group is given a caffeine pill. The response variable of interest is the number of words participants are able to recall following the break. We are testing to see if there is a difference in the average number of words a person can recall depending on whether the person slept or ingested caffeine.

Prediction and Testing: Sleep vs. Caffeine

How can we predict how many words someone will remember?

Data: Results of a recall experiment

Model 1: No predictors (CHOOSE step)

$$Y = c + \varepsilon$$

Words = "Typical" Number + Individual/Situational Influence

Individuals differ, but not based on whether they slept or took caffeine.

Prediction Error: the Residual

The model (population level):

$$Y = c + \varepsilon$$

The prediction (based on sample data):

$$\hat{Y} = \hat{c}$$

The prediction error: Actual Minus Predicted

$$Y - \hat{Y}$$

FIT/ASSESS/USE

- Later, we will discuss how to pick $\hat{c}$ (**FIT**ting the model to data), and how to **ASSESS** the model
- What about **USE**ing the model?
- Calculate $\hat{Y}$ the *predicted* number of words recalled

Model 2: Now With A Predictor!

Population model (CHOOSE step):

$$Y = c_i + \varepsilon$$

$$i = 1 \text{ if } X = \text{sleep}$$
$$i = 2 \text{ if } X = \text{caffeine}$$

$$f(X) = \begin{cases} c_1 & \text{if } X = \text{sleep} \\ c_2 & \text{if } X = \text{caffeine} \end{cases}$$

How can we decide between two models?

Pairs: How would you decide which model is better? (ASSESS step)

Simplicity vs. Fit

- The more complex model is guaranteed fit the data better (or at least no worse). (Why?)
- Need to balance fit by simplicity.
- "All else equal", prefer the simpler model.
- But what counts as "equal"? Exactly equal only?

Hypothesis Testing as Model Selection

Can adopt the simpler model by default, and see if there's enough evidence to reject.

$$H_0 : \mu_{\text{Athletes}} = \mu_{\text{Non-athletes}}$$
$$H_1 : \mu_{\text{Athletes}} \neq \mu_{\text{Non-athletes}}$$

$$H_0 \Leftrightarrow \text{Model 1}$$
$$H_1 \Leftrightarrow \text{Model 2}$$

USE and Interpretation

- Suppose we reject H_0 and favor the more complex model. Now we can make predictions. What can we conclude?
- In using the model to draw conclusions, we need to be sensitive to how the data was collected. (Really, should keep this in mind at every step)

A Note on Software

- We will use R (the "engine") via RStudio (the "control panel")
- Two options: Access via a log-in on your browser (`rstudio.oberlin.edu`), or install on your own computer (see below)
- Browser version a bit less smooth at times, getting data and work in and out is a bit clunkier at times, but less you need to manage
- If you want to use your own computer in lab, please install both R/RStudio on your computer before you get there.

R: http://www.r-project.org
RStudio: http://www.rstudio.com

HW 0(a) Everyone enrolled should come to my office hours sometime in the first two weeks, just for an intro. Book a 10-minute slot via my Google calendar (link on the course website)

First graded HW due electronically on Tuesday

- Lab 1 (we will start this in lab tomorrow)
- Textbook Problems (see website)