STAT 213
Statistical Modeling

Colin Reimer Dawson

Oberlin College

February 5, 2018

Outline

Outline

Outline

Handout

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? (Usually need to go back to step 1)
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 (Person i has group and number of words recalled: (X_i, Y_i))

Model 1: No predictors (CHOOSE step)

$$Y_i = c + \varepsilon_i$$

Words = "Typical" Number + Individual/Situational Influence

Each individual i is different, but not based on whether they slept or took caffeine.

FIT/ASSESS/USE

- Later, we will discuss how to estimate c (**FIT**ting the model to data), and how to **ASSESS** the model
- What about **USE**ing the model?
- Predict based on the inputs (in this case, none): estimate an individual's outcome using the "typical" number, c

Model 2: Now With A Predictor!

Population model (CHOOSE step):

$$Y_i = c_{X_i} + \varepsilon_i$$

$$c_{X_i} = c_{\text{sleep}} \text{ if } X_i = \text{sleep}$$
$$c_{X_i} = c_{\text{caffeine}} \text{ if } X_i = \text{caffeine}$$

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

How can we decide between two models?

Model 1: No predictors

$$Y_i = c + \varepsilon_i$$

Model 2: Predictor based on group

$$Y_i = c_{X_i} + \varepsilon_i$$

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.

$$H_0 : \mu_{\text{Sleep}} = \mu_{\text{Caffeine}}$$
$$H_1 : \mu_{\text{Sleep}} \neq \mu_{\text{Caffeine}}$$

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

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)

Outline

On the Web

- Course Website: `http://colindawson.net/stat213`
- Syllabus, slides, handouts, homework, labs, demos available there
- Exception: Solutions to homework, etc., on Blackboard
- Also on Blackboard: electronic submission of assignments

Course Outline

Part I: Linear Models (about 6 weeks)

- Review simple linear regression (about 2 weeks)
- Linear models w/ multiple predictors (about 4 weeks)

Part II: Logistic Regression (about 3 weeks)

- Regression models with a binary response variable

Part III: Experimental data and ANOVA (about 3 weeks)

Also:

- Computational Skills for Data Analysis (throughout)

Graded Components

- Each class, identify (post on Slack) at least one thing you found interesting/puzzling from
  - the last class
  - the readings for the upcoming class
- Weekly(ish) 10-15 minute quizzes, most Fridays (not this week)
- Weekly(ish) problem sets (graded for effort/completion only) $\longrightarrow$ mostly due Monday nights, turn in electronically via Blackboard
- 2 In-Class Exams
- 2 data analysis projects

Grading Principles

Non-traditional system based on four principles:

1. Feedback and marks should align with levels of competence with specific content
2. As few "choke points" as possible
3. Final grade should be based on where you get, not how quickly you get there
4. High standard on individual assignments (everyone has room for improvement)

Grading System

Grading mechanics with these goals in mind

1. Graded items are graded based on "Specific Learning Objectives" (SLOs)
2. Many opportunities for "reassessments" to replace individual quiz/exam questions (including optional final exam and project)
3. Grade for each objective no lower than the grade the last time it is assessed
4. Individual items are graded fairly stringently

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

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

First To-Dos

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)

Please fill out the background survey linked at the course website before next class

Review/crash course material:

- HW 0a: Intro to R (see "Homework" tab)
- HW 0b: Intro to Regression in R ("Homework" tab)
- Alternative/additional resources are on the website under "Resources"

First official HW due next Monday

- HW1: Some textbook Problems (see website)