Statistics 625a: Case Studies
Jay Emerson

Days and Times TBA
@ Organizational Meeting, Sept 2, 4 PM

24 Hillhouse Avenue, Computer Lab

Practical Information

- Please email me with any questions: john.emerson@yale.edu.
- Lectures/labs: the course will meet twice each week, days/times TBA.
- There is no textbook, only notes and datasets. If you have a particular problem or data set of interest, please let us know at the beginning of the semester.
- We will use R almost exclusively (http://www.r-project.org).
- The course web page is on the new classes (version 2) server: http://classesv2.yale.edu. Please log in and join the class.

Objectives and Prerequisites

Statistical Case Studies involves the statistical analysis of a variety of problems including death penalty cases from California, nationalistic biases in the judging of Olympic diving, a job hiring discrimination lawsuit, the prediction of college basketball games, and airline on-time performance. We will emphasize methods of choosing data, acquiring data, and assessing data quality.

Graduate, professional, and undergraduate students from any department are welcome, but must seek permission (discussing their background in statistics and goals for the semester) at or before the first class meeting. At least one prior course in statistics is required, but the most important prerequisite is a willingness to get your hands dirty working with real data sets. This will entail a certain amount of "programming," which we believe can be best taught by example, trial and error.

Supplementary Material

Some useful references, in no particular order:

- Alan Agresti: Categorical Data Analysis or An Introduction to Categorical Data Analysis.
- Sanford Weisberg: Applied Linear Regression.
- Maindonald and Broun: Data Analysis and Graphics Using R: An Example-based Approach.
- James Gentle: Elements of Computational Statistics.
- An Introduction to R, (FREE!) by Venables and Smith and the R Development Core Team. You can download the newest version from http://www.r-project.org. For $10-$15 you'll have quite a nice R reference book, at least covering the basics.
- Chambers and Hastie: Statistical Models in S.
- Venables and Ripley: Modern Applied Statistics with S.
- Everitt and Hothorn: A Handbook of Statistical Analyses Using R.
- Heiberger and Holland: Statistical Analysis and Data Display: An Intermediate Course with Examples in S-Plus, R, and SAS.
- Deepayan Sarkar: Lattice: Multivariate Data Visualization with R.
- Unwin, Theus, and Hofman: Graphics of Large Datasets: Visualizing a Million.
- Cook and Swayne: Interactive and Dynamic Graphics for Data Analysis.

Homework

The homework, class presentations and participation, and a short project constitute the entire grade for this class. There are no exams. A certain amount of emphasis will be placed on writing short reports, to help students prepare for the qualifying practical exam. Students not required to take the practical exam may (optionally) take the exam instead of doing the class project; please speak with me if this option interests you. Students will be expected to take turns giving short presentations or demonstrations of their analyses throughout the semester.

Computing

Most students have prior experience with R (a free implementation of the S language, which works in Windows, Linux, or Mac OS-X). It can be downloaded from http://www.r-project.org, along with "An Introduction to R" by Venables, Smith, and the R Development Core Team. If you have never used R before and need help getting started, please speak with Jay directly. The computers in the Statlab (or the cluster computers in the basement of 24 Hillhouse) should have everything you need for this class.

Week 1 (one class only): Getting Your Hands Dirty

A crash course in HTML and web page basics (using student accounts on Pantheon). Believe it or not, this will have obvious practical use (for sharing class material between students/faculty and for submission of homework) but will also serve us well when we confront our first web data scrape.

Weeks 2-3: The Death Penalty in California

What factors account for who is sentenced to death in death-eligible first-degree murder cases? We will consider the nature of the crime, the prior felony record of the defendant, the race/ethnicity of the defendant and victim(s), the gender of the victim(s), the location of the crime, whether the case was tried or not, as well as other available information.

Weeks 4-5: College Basketball Point Spreads (basic data processing and analysis)

The web contains a wealth of information. Some data sets are easily downloadable (perhaps as Excel files). Others (and often the more interesting and obscure ones) require a little more work. This introduction to scraping data from the web will explore the accuracy of bookies' point spreads in predicting the outcome of Division I college basketball games.

Weeks 6-9: Nationalistic Biases in Judging of Olympic Diving (advanced data processing and analysis)

Hopefully data from the 2008 Olympic Games in Beijing will be available (we might consider other competitions, as well). Otherwise, we'll use older data. What can be said about the nationalistic preferences of the judges?

Week 10: A Job Hiring Discrimination Lawsuit

A final exercise.

Conclusion: Student Presentations