Lecture 0
Introduction and Syllabus
STAT 432, Spring 2021

Course Topics

Linear Regression: modeling the relationship between a single continuous response variable Y, and one or more explanatory variables $X_1, X_2, \cdots, X_p$.

Logistic Regression: modeling the relationship between a single binary response variable Y, which takes on values 0 or 1, and one or more explanatory variables $X_1, X_2, \cdots, X_p$.

This course will focus on the application and computation of regression models using R and RStudio. Various real data sets will be investigated throughout the semester. We will also discuss some mathematical theory, although this will not be emphasized that heavily.

Motivation

Regression modeling has several objectives:

- Making predictions for future or unknown values of the response variable, and evaluating the uncertainty in those predictions.
- Assessing the relationship between the response and explanatory variables.
- Providing insight into the data structure (e.g., checking for unusual or influential observations)

Grading

- 60% Two Exams (take-home)
- 20% Homework Assignments
- 20% Project

Exams: There will be two exams, each worth 30% of your grade. There will be no final exam.

Homework: There will be weekly or biweekly homework assignments. You should receive full credit, or close to full credit, if you put in a reasonable effort, and turn in your work on time. Also, I may not grade every problem, but I will post solutions on Blackboard.

Project and Presentation: For the project you will find a data set of interest, and then conduct a regression analysis using that data set. It will be due the last week of class.

Policy on Late Assignments: Late homework will generally not be accepted. However, your lowest scoring homework assignment will be dropped. I may agree to extensions on due dates if you are experiencing an emergency or illness.

Textbooks

The class will mostly focus on slides and notes. I will use the textbooks listed below as references. You can access electronic versions of these textbooks on the internet, so no purchase is necessary. You will only be expected to know material covered in lecture slides (which will be posted on Blackboard) and notes.

James, G., Witten, D., Hastie, T., and Tibshirani, R. *An Introduction to Statistical Learning with Applications in R*. Springer, 2013.

PDF version posted on Blackboard in the "Resources" folder.
Book website: `https://statlearning.com/`

This book is written at the undergraduate level, and provides an accessible introduction to linear and logistic regression, cross-validation, and statistical learning algorithms.

Textbooks

Sanford Weisberg. *Applied linear regression*, John Wiley & Sons, Fourth Edition, 2014.

Free electronic version: `http://library.csueastbay.edu/home`
Book website: `http://users.stat.umn.edu/~sandy/alr4ed/`

This book is written at an advanced undergraduate or first year graduate level. Some background in statistical methods, calculus, matrix algebra, and programming is assumed.

Diez, D.M., Barr, C.D. and Cetinkaya-Rundel M. *OpenIntro: Statistics*, Fourth edition, 2019.

PDF version posted on Blackboard in the "Resources" folder.
Book website: `https://www.openintro.org/`

This book assumes no prerequisites, and is written for college students taking introductory statistics.

Software

We will use R and RStudio for data analysis and linear regression modeling. This course does not assume background in R programming. Some R topics we will cover:

- Vectors, matrices, and data frames
- Loading data files into R
- Data visualization and summary statistics (base R and `ggplot2`)
- Linear regression modeling with `lm()`
- Logistic regression modeling with `glm()`
- Report writing and reproducible research (R Markdown)

Why learn R?

- It is a **free** and open-source software that runs on most operating systems (Windows, Mac, Linux).
- It is one of the most **popular** programming languages used for statistics and data science.
- It is a desired and necessary skill for many statistics and data science **jobs** in academia, industry, and government.
- It is a legitimate programming language that allows more control and **reproducibility** than a point-and-click interface.