Lecture 00
Introduction and Syllabus
STAT 632, Spring 2022

Course Topics

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

The course will cover practical applications to a variety of real data sets, the mathematical theory of the linear and logistic regression model, and computation in R.

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)

Additional Topics

We may also cover several modern statistical modeling techniques related to linear regression:

- Regularization methods such as ridge regression and LASSO that are useful when there are many predictor variables ($p \approx n$ or $p > n$).
- Decision trees and random forest models that are useful when there are many predictor variables with nonlinear relationships.
- Generalized least squares estimation, which is a method that can be used when the data are autocorrelated (e.g., time series data)
- Matrix algebra as it relates to estimation

Grading

- 40% Two Midterm Exams (20% each)
- 30% Homework
- 20% Project Paper
- 10% Presentation

Exams: There will be two midterm exams, each worth 20% of your grade. There will be no final exam.

Homework: There will be 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. You are encouraged to use R Markdown for data analysis and coding exercises. I will drop your lowest scoring homework assignment.

Project and Presentation: For the project you will need to find a data set of interest, and then conduct a regression analysis using that data set. You will be required to give a presentation on your project during the last two weeks of class. For the final project you are also encouraged to use a modern method (e.g., random forests, LASSO).

Textbooks

Simon Sheather. *A Modern Approach to Regression with R*, Springer, 2009.

Free electronic version: `http://library.csueastbay.edu/home`
Data sets and R code: `http://gattonweb.uky.edu/sheather/book/`

This will be the main textbook for the course. If you wish, you may buy a hard copy through the CSUEB student store, Amazon, or Springer.

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

Free PDF version: `http://www-bcf.usc.edu/~gareth/ISL/`

We will reference this textbook when covering statistical learning topics such as cross-validation, LASSO, and random forests.

Software

We will use R and RStudio for data analysis and statistical modeling. This course assumes some familiarity with computer programming. We will try to cover all of the following R topics:

- Data visualization and summary statistics (base R and `ggplot2`)
- Linear regression modeling with the `lm()` function
- Logistic regression modeling with the `glm()` function
- Random forest modeling with the `randomForest` package
- LASSO and ridge regression with the `glmnet` package
- Report writing and reproducible research (R Markdown, knitr)

Software

You are also encouraged to learn LaTeX for mathematical typesetting. LaTeX can be combined with R using Markdown or knitr.

- Download LaTeX: `https://www.latex-project.org/`
- Resource for learning LaTeX: `https://www.overleaf.com/learn/latex/Main_Page`

Analysis of Variance versus Regression

Difference between ANOVA and regression? [Categorical vs. Quantitative Predictors]

Sir Francis Galton (regression towards mediocrity, `https://en.wikipedia.org/wiki/Francis_Galton`)
He was the half-cousin of Charles Darwin, founded correlation in the statistical sense and is credited with popularizing regression.