STAT 340: Applied Regression Methods

About the Course

Instructor Evan Ray

Email: `eray@mtholyoke.edu`

Office: Clapp 404C

Office Hours: I will hold regularly scheduled office hours each week at times to be selected by you. These times will be posted on the course web site. Please do not hesitate to contact me to set up appointments for additional office hours at any time!

Classes

Mon, Wed, Fri 9:30 - 10:45 am in Clapp 402.

Please plan to bring your laptop to class (we'll need to have sufficient numbers to allow you to work in pairs and take advantage of the technology available to us). Please contact me if you do not have access to a working laptop or if you forget yours; **the department has laptops available for you to use in class**.

Course Website The course website is at `http://www.evanlray.com/stat340_f2019/`. I will update it regularly with lecture notes and materials used in class. Lab assignments we will work through in class and homework assignments will be distributed on GitHub.

Description In this course, we will build on the ideas you have developed in introductory and intermediate statistics courses to develop a set of statistical methods and computational tools that can be used for common data analysis tasks in the frequentist framework. We will take the "applied" part of the course title seriously, focusing on working with real data sets to answer real questions. We will interpret "regression" broadly, discussing approaches for regression and classification.

In terms of statistical modeling, we will cover the following material:

- Review of simple and multiple linear regression; matrix formulation of regression; polynomial regression
- Evaluating model performance via train/test splits; cross-validation for model selection
- Regression with splines
- Non-parametric regression with K nearest neighbors
- Classification with two classes using logistic regression and K nearest neighbors
- Approaches to multiple regression with many candidate explanatory variables: variable selection, dimension reduction, and penalized estimation
- Tree-based approaches to regression and classification, including discussion of bagging (random forests) and boosting

- Generalized Additive Models (GAMs)

In terms of statistical computation, we will use R and cover the following material:

- Data management in R's tidyverse
- Data visualization with R's ggplot2
- Version control and collaboration with git and GitHub

Textbook We will be using "An Introduction to Statistical Learning with Applications in R" (ISLR) by James, Witten, Hastie, and Tibshirani as the primary text for the class. This is available as a hardcover (ISBN 978-1-4614-7137-0), and a pdf of the book can be downloaded for free from the author's website at `http://www-bcf.usc.edu/~gareth/ISL/`. A copy is on reserve at the library.

ISLR is an undergraduate-level version of "Elements of Statistical Learning" (ESL) by Hastie, Tibshirani, and Friedman. If you ever want more detail and mathematical depth, ESL is a great place to look. ESL is also available for free at `https://web.stanford.edu/~hastie/ElemStatLearn/`.

In addition, I will occasionally assign readings from "R for Data Science" (R4DS) by Wickham and Grolemund, which is available for free in website format at `http://r4ds.had.co.nz/`. If you wish, it is also possible to purchase a physical copy of this book.

Time commitment While the exact time commitment for the class will vary individually and over the course of the semester, I recommend that you budget approximately three out-of-class hours for every class hour to complete the reading, assignments, and homework. I have designed the class so that it should be feasible to satisfactorily complete the requirements with approximately twelve hours per week of time commitment. If you are spending more time than this on a regular basis I would encourage you to check in with me.

Policies

Attendance Your attendance in class is crucial, unless you are sick. If you are sick, please let me know and stay home and rest; I hope you feel better!

Collaboration Much of this course will operate on a collaborative basis, and you are expected and encouraged to work together with a partner or in small groups to study, complete homework assignments, and prepare for exams. However, every word that you write must be your own. Copying and pasting sentences, paragraphs, or large blocks of R code from another student is not acceptable and will receive no credit or a penalty. No interaction with anyone but the instructor is allowed on any exams or quizzes. All students, staff and faculty are bound by the Mount Holyoke College Honor Code.

To sum up: On homeworks and labs, **I want you to work together**. *But,* **you must write up your answers yourself.**

Cases of dishonesty, plagiarism, etc., will be reported.

Technology

Computing with R Modern statistics can't be done without computation. We will use the R statistical programming language in this course. R is one of the most commonly used programming languages in academic statistics, and I use it daily; it's also very commonly used in statistics and data science positions in industry. Knowing R is a marketable skill. In this class, you will use R nearly every day, and for many homework problems. I expect that you are familiar with R from previous classes, but I do not expect that you are an expert at R yet. That said, it is imperative that you let me know if you are confused about anything we are doing in R.

We will use R via RStudio; Mount Holyoke's version of RStudio Server can be accessed at `https://rstudio.mtholyoke.edu/`. You are also welcome to work locally on your own computer if you have RStudio set up; however, please make sure you have installed at least version 3.5.0 of R and the latest versions of any R libraries we use.

It will be important to **bring your laptop to class**; we will be using R nearly every day. Much of this work will be done in pairs, but we need to ensure that there is a sufficient number of computers. Please let me know if this presents any issues, as there are laptops available for you to borrow.

Version Control with Git and GitHub Git is a version control system that facilitates working on coding and writing projects collaboratively, and allows you to revert your code to a previous version if you realize that you made a mistake. Version control systems such as git are used in most modern data science and statistics positions in industry. Part of my goal as an educator in the statistics program is to ensure that you are prepared to enter the work force, and for that reason the basic use of git is a learning objective for this course. This means that all labs and the computational portion of homework assignments will be distributed to you in git repositories and submitted by committing and pushing the completed assignment to GitHub. I will provide further details and walk through this process, as well as basic interaction with git, in class. Note that we will use the graphical interface to git that is built into RStudio rather than the command line interface to git.

Assignments

Your grade for this course will be a weighted average of scores from several components:

Item	Weight
Participation and Labs	5%
Homework	15%
Quizzes and Midterms	60%
Final Project	20%

Participation and Labs The best way to learn statistics is to do it. This class will be built around a series of labs that we will do in class. Although I will not grade these labs for correctness, I expect you to complete them and push your work to GitHub. I will occasionally look at submitted

labs to see how everyone is doing and whether there are any points I need to address in class. I am always happy to answer any questions you have about these labs.

Homework We will have regular homework assignments to be completed outside of class. Occasionally, questions may relate to material in the reading that will not be covered in class.

Exams There will be two or three "midterm" exams and occasional in-class quizzes. Midterms will definitely have a take-home component, and may also have an in-class written component. Quizzes will always be announced at least once class session in advance, and midterms at least one week in advance. We will not have a cumulative final, but we may have a late-in-the-semester "midterm". No communication with anyone besides the instructor is allowed on these assessments.

Project A large component of the course will be a project which will be presented to your classmates. Briefly, this project will entail application of a statistical model beyond multiple regression to a data set of your choosing and/or a simulation study to compare performance of two chosen statistical models. A separate handout will provide additional details.

Extra Credit Extra credit is available in several ways: attending an out-of-class lecture (as will be announced) and writing a short review of it; pointing out a substantial mistake in the book, a homework exercise or exam solution; drawing my attention to an interesting data set or news article; etc. The extra credit is applied when a student is near the boundary of a letter grade.

Grading When grading your written work, I am looking for solutions that are technically correct and reasoning that is clearly explained. *Numerically correct answers, alone, are not sufficient* on homework, tests or quizzes. Neatness and organization are valued, with brief, clear answers that explain your thinking. If I cannot read or follow your work, I cannot give you full credit for it.