EDP 384: Data Analysis, Simulation, and Programming in R
Spring 2015, Tues/Thur, 2:00 - 3:30 pm
SZB 524

Instructor: James E. Pustejovsky
Email: pusto@austin.utexas.edu
Phone: 512-471-0683
Office hours: Mondays, 1-3 pm or by appointment
Office: SZB 538 D

TA: Daniel M. Swan
Email: dswan@utexas.edu

Course Description

This course provides training in using the open-source statistical programming environment called R to accomplish 1) real-world, reproducible data analysis and 2) design and implementation of statistical simulations. The first part of the course introduces the logic of R's primary data structures and how to work with functions. The second part of the course covers tools and best practices for cleaning, manipulating, and visualizing data. It also introduces the concept of reproducibility as a fundamental tenet of high-quality data analysis. The third section of the course focuses on how to run and work with linear regression models and generalized linear models. The final section of the course covers how to design and run Monte Carlo simulations in order to evaluate the performance of statistical estimation and inference procedures.

Learning Goals

By the end of this course, you should be able to use R to…

- Clean, manipulate, and organize data;
- Create clear, information-rich graphical displays of data;
- Run data analysis projects that are well organized and fully reproducible; and
- Program basic Monte Carlo simulations

Prerequisites

There are no formal prerequisites for this course, though basic knowledge of linear regression analysis will be useful. The course does NOT require previous experience with R programming.

Readings

- Required text: Lander, Jared P. (2014). *R for Everyone: Advanced Analytics and Graphics*. Upper Saddle River, NJ: Addison-Wesley.
- Readings posted on Canvas.

- Recommended further reading: Matloff, Norman (2011). *The Art of R Programming: A Tour of Statistical Software Design*. San Francisco, CA: No Starch Press. Students with prior programming experience who wish to understand the guts of how R works at a more advance level will find this book to be very useful (though it is not required for the course).

Course components

Class meetings. Class meetings will consist of a mixture of lecture, demonstration, and group exercises designed to help you master the relevant material. Your active engagement in the group exercises is *essential* for success in the course.

Homeworks. Homework will be assigned every 1-2 weeks, with more assignments towards the beginning of the semester. Each of the assignments will involve writing R code to analyze data, providing further opportunity for you to apply and test your understanding of the material discussed in class.

Data analysis project. This project involves completing a small data analysis project based on a topic and dataset of your own choosing. The goal of the project is to demonstrate your ability to organize and complete a full data analysis, from reading in the data to communicating the results. Completed projects will be assessed based on the extent to which they demonstrate competence in cleaning, manipulating, visualizing, and analyzing a real-life dataset. You may work in pairs for this project (though you are not required to do so). The first draft of this project will be due after Spring break.

Monte Carlo simulation project. The final project for this course involves programming a small Monte Carlo simulation using R. The simulation design can be:

- an exact replication of a published simulation study (chosen by the student);
- an extension (or conceptual replication) of a published simulation study; or
- an original study (for advanced students only).

Completed projects will be assessed based on the extent to which they demonstrate competence in organizing and implementing the simulation design. You must complete the simulation project individually. The final project will be due during the final exam period.

Evaluation

- Homeworks (60%).
- Data analysis project (20%).
- Monte Carlo simulation project (20%).

A tentative rubric for assignment of final grades is listed below. ***The instructor reserves the right to modify this rubric.*** Square brackets correspond to ≤ or ≥; rounded parentheses correspond to < or >.

A	[90, 100]	C+	[74, 77)
A-	[87, 90)	C	[70, 74)
B+	[84, 87)	C-	[67, 70)
B	[80, 84)	D	[60, 67)
B-	[77, 80)	F	[0, 60)

Collaboration

You are encouraged to discuss the concepts, material, and homework with other students in order to better understand them. However, you must write your own solutions to the homework problems and projects (with the exception of the data analysis project, where you may work with a partner). For example, you must write your own code, run your own data analyses, and communicate and explain the results in your own words and with your own visualizations. You may NOT submit work that you have submitted or will submit to another course.

Academic Integrity

Following the University's honor code, you are expected to maintain absolute integrity and a high standard of individual honor in scholastic work. You must complete all assignments (including homeworks and projects) with the utmost honesty, which includes acknowledging the contributions of other sources to your scholastic efforts and avoiding plagiarism. ***Assignments containing any plagiarized material will not be accepted.***

Attendance

You are responsible for all of the material discussed during class meetings. If you must miss a class, it is your responsibility to obtain and thoroughly review notes or summaries of the material that they missed.

ADA Accommodations

The University of Texas at Austin provides upon request appropriate accommodations for qualified students with disabilities. For more information, please contact the Office of the Dean of Students at 471-6259, 471-4671 TTY.

Religious Holidays

By UT Austin policy, students must notify the instructor of a pending absence due to religious observance at least fourteen days in advance. If the student must miss a class, an examination, a work assignment, or a project in order to observe a religious holy day, the student will be given an opportunity to complete the missed work within a reasonable time after the absence, with no penalty.

Tentative schedule

R fundamentals

1/20 - Course introduction
1/22 - A first look at R and the R environment
1/27 - Reading in data
1/29 - Working with data frames
2/3 - Working with vectors
2/5 - Strings, factors, and dates
2/10 - Lists, matrices and arrays
2/12 - Writing helper functions

Data wrangling and visualization

2/17 - Cleaning data sets
2/19 - Reshaping and merging
2/24 - Group-wise data manipulation
2/26 - Project organization and reproducibility with knitr and RMarkdown
3/3 - Visualizing data - basic graph types
3/5 - Visualizing data - fancy stuff

Data analysis

3/10 - Smoothing techniques
3/12 - Linear regression models
3/17 - No class (Spring break)
3/19 - No class (Spring break)
3/24 - Linear regression models
3/26 - Linear regression diagnostics
3/31 - Robust standard errors
4/2 - Generalized linear models

Simulation

4/7 - Monte Carlo simulation
4/9 - Probability distributions
4/14 - Designing simulation studies
4/16 - TBD
4/21 - Data-generating models
4/23 - Estimation procedures
4/28 - Performance criteria
4/30 - Experimental design and implementation
5/5 - Describing simulation results
5/7 - Parallelization