EDP 380C-20 (10875): Data Analysis, Simulation, and Programming in R
Spring 2017, Tues/Thur, 11:00 am - 12:30 pm
SZB 432

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

For a long time I have thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have had cause to wonder and to doubt…. All in all, I have come to feel that my central interest is in data analysis*, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data.*

- Tukey, J. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, *33*(1), 1–67.

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 accessing, cleaning, and manipulating data. It also introduces the concept of reproducibility as a fundamental tenet of high-quality data analysis. The third section of the course introduces data visualization techniques and a selection of statistical models that are useful for data-analysis, including linear regression models and generalized linear models. Each section will also include content relevant to designing and implementing Monte Carlo simulation studies, which are an important tool for evaluating the performance of statistical estimation and inference procedures.

Learning Goals

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

- Access, 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: Grolemund, G., & Wickham, H. (2016). *R for Data Science*. Sebastopol, CA: O'Reilly Media, Inc. Note: This book is freely available online at http://r4ds.had.co.nz/. You can also purchase a hard copy for about $30 if you prefer.
- Further readings posted on Canvas.
- Recommended further resources:
    - Lander, Jared P. (2014). *R for Everyone: Advanced Analytics and Graphics*. Upper Saddle River, NJ: Addison-Wesley. This book covers similar content to Grolemund & Wickham (2016), but might be more accessible to students who are entirely new to R.
    - 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.

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. 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.

Package demonstration. Students will work in pairs to give a short (~10 minute) presentation about an R package of their choosing. The presentation should explain ***why*** and ***how*** the package is useful, provide an overview of the package's capabilities, and

give examples of how to use the central functions. Presentations will be scheduled to begin the second week of April.

Course Project. Students may choose between one of the following two options for an end-of-course project. Final drafts of the course project will be due during the final exam period at the conclusion of the course. ***Students in Quantitative Methods are strongly encouraged to do the Monte Carlo Simulation project.***

1. ***Data analysis project.*** This project option 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).

2. ***Monte Carlo simulation project.*** This project option 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.

Evaluation

- Homeworks (60%).
- Package demonstration (10%).
- Final project (30%).

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.

Carrying of Handguns

Students in this class should be aware of the following university policies:

- Individuals who hold a license to carry are eligible to carry a concealed handgun on campus, including in most outdoor areas, buildings and spaces that are accessible to the public, and in classrooms.
- It is the responsibility of concealed-carry license holders to carry their handguns on or about their person at all times while on campus. Open carry is NOT permitted,

meaning that a license holder may not carry a partially or wholly visible handgun on campus premises or on any university driveway, street, sidewalk or walkway, parking lot, parking garage, or other parking area.

Emergency Evacuation Policy

Occupants of buildings on the UT Austin campus are required to evacuate and assemble outside when a fire alarm is activated or an announcement is made. Please be aware of the following policies regarding evacuation:

- Familiarize yourself with all exit doors of the classroom and the building. Remember that the nearest exit door may not be the one you used when you entered the building.
- If you require assistance to evacuate, inform the instructor in writing during the first week of class.
- In the event of an evacuation, follow the instructions of the instructor.

Do not re-enter a building unless you're given instructions by the Austin Fire Department, the UT Austin Police Department, or the Fire Prevention Services office.

Tentative schedule

R fundamentals

1/17 – Course introduction
1/19 – A first look at R and the R environment
1/24 – Reading in data
1/26 – Working with data frames
1/31 – Working with vectors
2/2 – Strings, factors, and dates
2/7 – Lists, matrices and arrays
2/9 – Writing helper functions

Data wrangling

2/14 – Cleaning data sets
2/16 – Merging
2/21 – Reshaping
2/23 – Group-wise data manipulation
2/28 – Working with databases

Data visualization and analysis

3/2 – Project organization and reproducibility with knitr and RMarkdown
3/7 – Visualizing data - basic graph types
3/9 – Visualizing data - fancy stuff
3/14 – No class (Spring break)
3/16 – No class (Spring break)
3/21 – Smoothing techniques
3/23 – Linear regression models
3/28 – Linear regression diagnostics
3/30 – Robust standard errors
4/4 – Generalized linear models
4/6 – Hierarchical linear models

Simulation Studies and Package Demonstrations

4/11 – Monte Carlo simulation
4/13 – Data-generating models and estimation procedures
4/18 – Performance criteria, experimental designs
4/20 – Experimental design and implementation
4/25 – Parallelization
4/27 – TBD
5/2 – TBD
5/4 – TBD