Statistics 405/705 **Spring, 2017**

Introduction to Programming in R

Instructor and TA

Robert Stine	444 Huntsman Hall	stine@wharton
Ruoqi Yu		ruoqiyu@wharton

There are two sections of this course; these meet at 12-1:30 pm and 1:30-3:00 pm on Mondays and Wednesdays in 240 JMHH. Please attend the appropriate section.

Professor Stine's office hours follow classes on Monday from 3-5 pm and on Wednesday from 3-4:30 pm. Ruoqi will post her office hours on the Canvas website.

Overview

This course introduces students to the R statistical programming environment. R is the de-facto standard for writing statistical software among statisticians and has made substantial inroads in other applied disciplines, from biology to physics to sociology. R is a free, open-source implementation of the S language, originally develped at Bell Labs. Implementation of R are available for Windows, Mac OS X, and Unix/Linux systems.

This course is intended for students who are familiar with statistical methods at the level of Stat 102, Stat 112, Stat 431, or Stat 613. These courses introduce the foundations of statistical inference (concepts such as random variables, standard error, and confidence intervals) and regression analysis. This course does not presume prior programming experience.

Each class introduces concepts of R programming through combination of lecture and hands-on tasks. You will want to bring your laptop with the course software to each lecture. That said, you are expected to limit your use of the laptop to the course material and avoid browsing other content or email.

Grading

Grades for this course are determined by performance on five weekly homework assignments and a final project.

Weekly assignments	65%	(equally weighted)
Project	35%	

Assignments are due weekly; see the course outline below for the due dates. Assignments will generally be due on the Wednesday of the week following the lectures that cover the relevant material. That means you'll have the weekend to work on the assignment and can ask questions on Monday before turning in the assignment on Wednesday. The project is basically a longer assignment.

Note: *All questions about grades on assignments must be resolved within one week of posting grades.*

Collaboration

Assignments are to be completed individually. I expect you to talk with each other about what's going on in the class and help each other understand the concepts. Each assignment, however, is your own responsibility. You can get Google to help you, but not your classmates. You won't learn R unless you do these assignments yourself.

Materials

Lecture notes

These will be distributed via the course Canvas web page (in the Files section). The notes will be in the form of R-notebooks. R-notebooks are distinguished the the suffix ".Rmd" in the file name. These files combine text, R commands, and the output from the R commands. In each class, we'll work through these notebooks together.

Software

R-studio. Download the most recent, appropriate version (*i.e*., Windows, Mac or Linux) of this free software from the website

https://www.rstudio.com/products/rstudio/download/

Textbooks

This course does not have a textbook. To see further discussion of the material beyond that included in the lecture notebooks, you have many choices. R itself includes a built-in help facility that provides a summary of what each function in R does, with examples. Many will find these descriptions "terse" and benefit from further description and examples.

In general, once you know the topic and the name of the relevant R function, it's hard to beat Google as a source for more information. The R-Studio site itself includes a substantial amount of on-line help and documentation. Google can also help you find more information and "courses" on line (such as the course at https://www.datacamp.com/courses/free-introduction-to-r).

For those who want to read about the foundations of the R language, several long manuals that describe the language are available from the CRAN (short for Comprehensive R Archive Network) website at

https://cran.r-project.org/manuals.html

For those who like old-fashioned textbooks, I recommend the text *An R Companion to Applied Regression* by J. Fox and S. Weisberg (Sage 2011). This book introduces R in the context of regression modeling. J. Macdonald and J Braun, *Data Analysis and Graphics Using R* covers a wider range of examples with less social science flavor. W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S* covers more advanced methods and programming.

Planned schedule of lectures and assignments

The first four lectures use R "out of the box", exploring the built-in representations of data and functions to manipulate them. The next five lectures introduce programming in R, extending the language with customized functions. The next two classes get deeper into graphics in R, including the popular package ggplot that can be used to construct impressive plots. The following classes dive into manipulating messy data, introducing another population package named plyr. The last class is an opportunity to fill in gaps and survey further, more advanced topics.

Date	Topic	Assignments
Jan 11	Introduction to R and R-Studio	
Jan 18	Data structures in R	
Jan 23	Data frames and files	
Jan 25	Regression models in R	Assign 1 due
Jan 30	Functions in R	
Feb 1	Iterative algorithms and control statements	Assign 2 due
Feb 6	Tracing and debugging functions in R	
Feb 8	Simulation techniques in R	Assign 3 due
Feb 13	Application: bootstrap resampling	
Feb 15	Graphics in R	Assign 4 due
Feb 20	Graphics using ggplot	
Feb 22	Manipulating data with plyr	Assign 5 due
Feb 27	Application: Dealing with transactional data	
Mar 1	Going further: Object-oriented programming and generic functions	
Mar 3		Project due