Lesson 01 - Introduction to the class

Last Updated 06-27-2019

Course Overview

This course is designed as a primer to get the complete novice up and running with the basic knowledge of how to use the statistical programming language R in an environment that emphasizes reproducible research and literate programming for data analysis.

The target audience is anyone who wants to do their own data analysis. The course will cumulate with an peer-evaluated exploratory data analysis on either a pre-specified data set or your data set of choice.

Schedule of Topics

- Intro to the R language and it's packages
- R Studio suite of awesomeness
- Reproducible Research with R Markdown
- Functions
- Getting data into R
- Managing the spreadsheet-like data frame
- Data Visualization
- Data Manipulation and Aggregation
- Exploratory Data Analysis

Logistics

- All content is fully online on the course website: https://norcalbiostat.github.io/MATH130/
- Class time is spent expanding on ideas and concepts, working through assignments as a class or in pairs.

Grading

- Credit / No Credit. There are 100 points available in this course through attendance, assignments, and a final project.
- You must earn 70 points to receive credit for the course.
- The syllabus provides specific details on points per assignment.

Time Committment

- Fast paced, we're hitting the ground running and only have 5 weeks.
- Not designed to teach you everything about R, just enough to make you dangerous.
- Daily practice will pay off.
- But can be very rewarding
- Garbage in = garbage out
- Effort in = satisfaction

We've all been there. It will get easier

Tool choices

The term "`R`" is used to refer to both the programming language and the software that interprets the scripts written using it.

RStudio is currently a very popular way to not only write your R scripts but also to interact with the R software. To function correctly, RStudio needs R and therefore both need to be installed on your computer.

Why use R?

- Open source, cross-platform, and free
- Great for reproducibility
- Interdisciplinary and extensible
    - Examples in Data Science, Economics, Natural Sciences, Psychology.
- Tons of learning resources
- Currently R is used in Chico's upper division Applied Statistics and Data Science courses.
- Does not involve lots of pointing and clicking (that's a good thing!)
- Works on data of all shapes and sizes
- Produces high-quality graphics
- Large and welcoming community

Why use R Studio?

- Customizable workspace that docks all windows together.
- Notebook formats that allow for easy sharing of code and output, and integration with other languages (Python, C++, SQL, Stan)
- Syntax highlighting, warning errors when missing a closing parentheses.
- Cross-platform interface. Also works on Windows/iOS/Linux.
- Tab completion for functions. Forget the syntax or a variable name? Popup helpers are available.
- Free training videos available from the developers directly.
- Built in version control
- One button publishing of reproducible documents such as reports, interactive visualizations, presentations (like this one!), websites.

Programming is scary!

Learning to program has other benefits

- Improves your logical skills and critical problem solving
- Your series of commands are explictily written down and subject to review.
- Increases your attention to detail
- Increases your self reliance and empowers you to control your own research.
- Your PI will love your awesome graphics and reports.
- Some people think what you do is magic.
- Thinking graduate school? [expect to learn this on your own]
- [A few] [other lists] [of reasons]
- The key to learning the details of how things work under the hood is often by trial and error.

Why no point and click?

Because it's not reproducible.

- Which boxes did you click last time?
- New data? Gotta do it all over.
- Need to expand your model? Gotta do it all over.
- Made a mistake in the data coding? Gotta do it all over. . .

Questions?