Class Introduction

MSCR 596: Advanced Data Management in R

Beau B. Bruce, MD, PhD

(use right arrow key to advance slides)

Pros and Cons of R

Pros

- It is free (but not cheap!)
- Interactive and flexible
- Has all the features needed to start with raw data, analyze it, and transform it into a beautiful, reproducible report
- Easy to modify and expand
- Exceptional and easy to use graphics
- Huge package repository with cutting edge methods available (usually even before commercial software)
- Interfaces well with other software (databases, BUGS, etc.)

Cons

- Steep initial learning curve
- Because it is free, no commercial customer support
- Memory limits working with **gigantic** datasets
- Because it relies on community development, some areas may lag behind occasionally

Teaching Philosophy

Principle-oriented approach

> *It is better to obtain a solid understanding of the core principles. . . than a hazy understanding of a long laundry list of ideas. If you've understood the core ideas well, you can rapidly understand other material.*
>
> *Technologies come and technologies go, but insight is forever.*

Michael Nielson, "Neural Networks and Deep Learning"

Qualities of a Hacker

- Always a “newbie”
    - Taking on new challenges
    - Often feeling “in over their head” (cf. Imposter Syndrome)
    - Unafraid to say “I don’t know”
- Willing try to find answers on their own
- Know where to look for answers
- Curiosity
- Not afraid to “break” something
- Like to tinker
- Persistent

Not a statistics/math course

- We will not spend much time on how to perform a given statistical analysis in R nor will we cover the underlying statistical theory of analyses.
- Anyone who successfully completes the course
  - will have acquired the skills to run any common statistical analysis available in R
  - will have gained the much more valuable skills of creatively solving challenging data management problems
  - and will able to learn new techniques with relative ease

Data Science Venn Diagram

Drew Conway

Inquiry-based learning

- Impossible to learn programming by
  - Reading a book
  - Watching a lecture
- Only way to learn to program is to program
  - But just parroting/cutting 'n' pasting will not get you far
  - Need *creativity* for each problem
- Each module contains experimentation to develop your "hacker" mentality and force you to be creative
- Once you learn to think correctly the grammar will simply become the physics of your programming universe
  - "You just have to master, *once*, a particular way of thinking, and you will no longer need all those rules" Keith Devlin

Combine *Homo faber* with *Homo sapiens*

- i.e., the “human who makes” with “human who knows”
- recommending readings:
  “Becomming a 21st Century Tinkerer”, WSJ
  “Cultivating the Entrepreneurial Learner in the 21st Century”, John Seely Brown

Gecko&Fly

Quoteswave

Where to Get Help

Books

- No textbook for the course, but some people find books helpful
- Free books:
  - Analysis of Epidemiological Data using R and Epicalc
  - A Little Book of R for Biomedical Statistics
- Paid books:
  - R for Dummies
  - A Beginner's Guide to R

Online

- Try Google, but Rseek (rseek.org) is often better
- StackExchange (stackexchange.com) is a question and answer site

Within R

Get help on a specific function/command/etc. (only for loaded packages:

```
?<text>
```

Search the help files of all packages you have installed:

```
help.search("<text>")
```

Search all of the R documentation for a solution (even in packages you don't have installed):

```
RSiteSearch("<text>")
```

People

- Your classmates
  - Only ask that for the homework and other evaluative elements of the class that you always write your code independently of each other.
  - It is ok, however, to discuss general approaches to problems before you then work on your own.
  - Important though to make sure you can really do both the creative and technical components so don't over do talking to others.
- Me (but please try the others first)

Asking good questions

- Describe exactly the steps you took to reproduce the problem
  - Provide a minimal, complete, and working example
- The output you got
- The output you expected
- Details about the version of R, packages, and operating system
- What you did to try and fix it (spend at least 1 hour trying to solve the problem before reaching out for help)

Course Structure

How your grade is determined

Interacting in class and staying on top of the class is important because the material builds upon itself. Thus, the grade reflects this fact with 80% coming from participation and homework.

The remainder will come from a final project.

Class participation (40%)

- Unexcused absences 3% off final grade after first
- Excused absences will be almost always be allowed if you contact the instructor BEFORE class with explanation,
- If you contact the instructor DURING/AFTER class your absense will only rarely, be excused and only in the case of extraordinary circumstances

Homework assignments (40%)

- Each module has some programming problems to work on at home
- These must be submitted by e-mail before next week's class begins
- I will provide personalized feedback as soon as I can but always before the following class
- Each will be a small part of your overall grade (usually ∼2-4 modules a week; 10-20 very short problems total or fewer longer problems), but will be critical for self-assessment

Final project (20%)

- A statistical analysis that transforms a raw data file into a well-formatted, programmatically created report using the principles learned throughout the class
- Requires one substantial figure
- Requires one substantial table
- You can choose one of your own datasets or we can work to find one on the web
- *Brief* proposal (dataset, description of figure & table) due on Oct 28 so that I can approve by Nov 10, providing you adequate time to do and turn in by Dec 8

Closing Thoughts

"Using R is a bit akin to smoking. Beginnings are difficult, one may get headaches, and even gag on the first experiences. But in the long run, it becomes pleasurable, and even addictive. Yet, deep down, for those willing to be honest, there is something not fully healthy in it."
François Pinard; R-help; 20 Aug 2007