CS&SS 508 Syllabus

Aaron Erlich
Department of Political Science
University of Washington
aserlich@uw.edu

This draft: April 2, 2015*

Office Hours

T: 2:30-3:30
W: 5:30-630
or by appt.
Padelford C-24

Course Motivation

This course is designed to make your life easier in all further stats courses. It is designed to get you to think about statistics from an algorithmic programming perspective and make you efficient in your workflow in all further CS&SS courses.

It is designed with *Software Capentry's'* methodology in mind. Scientists of all sorts spend more time data "munging" than the do in actual analysis. Moreover, they often do something and then can't replicate it. The idea of this course is to make all of you as social scientist faster of data munding and capable of creating replicable research.

On Books

Our textbook will be *The Art of R Programming* by Norman Matloff. This book will serve as a good further reference as you go deeper into R. This book will be supplemented by Peter Dalgaard's *Introductory Statistics with R*, which is designed to cover some of the more traditional statistical introduction to R. I have also referenced some free alternatives for those who do not want to purchase the Dalgaard book.

Many students do not think the need to do the readings for stats or programming classes. I will not cover everything in the readings in class, but reading material will help you greatly in being successful in the class. I would highly encourage you to do the reading. I have tried to limit the reading to fewer than 40 pages a week.

*Please let me know if you have another resource to add or find any mistakes

On versions of R

Many of you will want to you use `RStudio` when you are starting out. There is a helpful way to get started. However, it also comes with limitations. We will discuss different text editors and different ways to Run R during the first week.

Homework and Grading

Here's what you need to know to succeed in the class:

1. The entire course will be graded C/NC.
2. In order to pass the class you *required* to submit all assignments

Here is the pertinent information about the homework assignments

1. There will be five homework assignments are generally due every other week . The due dates are subject to change but will never be earlier than in the syllabus
2. In order to help you learn best, it is suggested that you work on each assignment by yourself until you get stuck. If you get stuck, you are encouraged to ask your friends for help. Please do not copy code. It helps to write your own.
3. The bulk of the assignments will require you to say in words what you have done. Data show this helps you consolidate your learning.
4. It is required to submit the assignments in **hard copy**, unless you email before class with a reason you cannot be in class. I ask you to submit hard copies, to facilitate me grading and getting your assignments back to you in a timely fashion
5. To help you get a sense of how well you are understanding the material, each assignment will be given a numeric grade.
6. You are required to obtain an average of 60 to pass the class.
7. Late assignments will only be accepted until the end of the week (Sunday) the assignment is due, except in extreme circumstances. Since I believe in incentives to encourage timely behavior, the following formula is applied to every late assignment: $\frac{\sqrt{\ln(t+.1)}}{3}$, where t is the number of days the assignment is late. What this means in practice is there a medium size penalty for one day late, a large penalty for more than one day late, but this penalty doesn't grow very much after one day. By the fourth day, you will get a passing score only if you achieve 100 percent on the assignment. You can visualize this penalty function with the following code in R `curve(sqrt(log(t+.1))/3, from=1, to=4)`
8. If you do turn in an assignment late, please put it in my Box in 101 Gowen Hall, during business hours.

Lab Sessions

Lab sessions are optional. In the first lab session, I will give students and overview of LaTeXto help them use LaTeXfor their homework assignments. Most other lab sessions will involve a small number of example problems for students to work on in an environment where they can get immediate help.If you have homework or other questions, you can bring them to the lab session.

Topics

We will move as quickly/slowly as the group needs. However, the goal is to cover the following topics. I reserve the right to make the homework slightly delayed if we do not move as fast as expected

1. Why learn R?
   (a) Replicability
   (b) Open-source benefits
   (c) Beautiful graphics
   (d) Munge and analyze all in one place
2. R as a Computer Language for Stats in your Workflow
   (a) Libraries
   (b) Learning to help yourself
   (c) Understanding directories and version control
   (d) Workflow/Integration with LaTeX
3. Managing and Understanding Data
   (a) Understanding variable assignment and control
   (b) Data types, common data classes
   (c) Matrix, vector operations
   (d) Looking at your data (a first look)
   (e) Importing/Exporting
4. Control Sequences and More (Iterating, Applying)
   (a) For loops
   (b) The apply family
   (c) Sampling/Simulation example
   (d) Creating multiple tables with one line of code example
5. Functions
   (a) When do you want a function?

   (b) Function arguments
   (c) Some examples of functions

6. Regression
   (a) Understanding the R regression model
   (b) Understanding regression objects
   (c) Visualizing regression

7. Combining it all
   (a) Use simulation to undertake randomization inference

Weeks and Readings

All of these deadlines and weeks are provisional and subject to change

Topic 1: Why R

Week 1, April 1

Mandatory Readings

1. Read sections 1-10 carefully of *Google's Style Guide*. Skim the rest. You will be responsible to ensure that your code conforms to Google's Style Guide `https://google-styleguide.googlecode.com/svn/trunk/Rguide.xml`

2. Matloff, Introduction and Chapter 1 *Do not expect you will understand everything the first time though—this is a roadmap for where we are heading and will introduce you to some key concepts that you will become more familiar with later.

Going Further

Topic 2: R As a Computer Language for Stats

Week 2, April 8

HW 1, OUT ***Mandatory Readings***

1. Browse section 1 of the *Git Guide*

2. Look at *Software Carpentry's explanation of Git*

3. Go through my LaTeXsildes if you do not attend lab

Going Further

- Software Carpentry's The Unix Shell

Topic 3: Managing and Understanding Data

Week 3, April 15

HW 1, IN
HW 2, OUT
Mandatory Readings

1. Matloff, Chapter 2,3.1,3.2,3.5,3.7. This is a heavier than average reading week.

Going Further

Week 4, April 22

Mandatory Readings

1. 5.1,5.2 (through 5.22), 6.1, 4.1,4.2,4.3

Going Further

- Learn about a faster version of the `data.frame` for big data the `data.table` package

Topic 4: Control Sequences and More

Week 5, April 29

HW 2, IN
HW 3, OUT
Mandatory Readings

1. Matloff, Chapter 3.4,4.4,5.4,6.2, 7.1
2. Read Hadley Wickham on functionals `http://adv-r.had.co.nz/Functionals.html`

Going Further

- Learn about using the `plyr` library for easy data manipulation

Topic 5: Functions

Week 6, May 6

Mandatory Readings

1. Matloff Chapter 7.2-7.4.1

Going Further

- Matloff 7.4.1-
- Professor Shalizi of CMU on Writing Functions

Topic 6: Regression and More

Week 7, May 13

HW 3, IN
HW 4, OUT
Mandatory Readings

1. Dalgaard, Ch. 6, 11
2. Browse `http://www.r-tutor.com/elementary-statistics/simple-linear-regression`

Going Further

1. Faraway's guide *Practical Regression and Anova using R*

Week 8, May 20

Mandatory Readings

1. Matloff Chapter 8

Going Further

- Dalgarard, Chapter 3

Week 9, May, 27

HW 4, IN
HW 5, OUT
Mandatory Readings

1. Dalgaard, Chapter 13

Going Further

Topic 7: Combining it All

Week 10, June, 3

Mandatory Readings

1. Matloff Chapter 13

Going Further

Finals Week, June 10

HW 5 IN to my box at Gowen 101

1 Other Reference Materials

1. UNT Materials
2. Software Carprentry's Programming with R