Statistics 5014: Homework 1

Due Wednesday September 6

2017-08-29

For each assignment, turn in by the due date/time. Late assignments must be arranged prior to submission. In every case, assignments are to be typed neatly using proper English in Markdown.

This week, we spoke about Reproducible Research, R, Rstudio, Rmarkdown, and LaTeX. To summarize the ideas behind Reproducible Research, we are focusing on Reproducible Analysis. For us, Reproducible Analysis is accomplished by mixing code, figures and text into a cohesive document that fully describes both the process we took to go from data to results and the rational behind our data driven conclusions. Our goal should be to enable a moderately informed reader to follow our document and reproduce the steps we took to reach the results and hopefully conclusions we obtained.

Problem 1

Set up your computing platform. For this class, we will be using Git, R, and Python as the main software choices. If you install or get accounts as shown in the following table in the order listed, you should end with a usable install of all softwares used in the class.

Package	Source
Git:	https://git-scm.com/
Github:	https://github.com (account)
Latex	https://miktex.org/
R:	https://cran.r-project.org/
Rstudio:	https://rstudio.com/
Python:	https://www.python.org/
ARC account:	arc.vt.edu (user requests, account request)
Command line:	native Terminal (Mac)
	Putty (or equivelant WinSCP, etc; Windows)
	https://secure.hosting.vt.edu/www.arc.vt.edu/accessing-unix-system/#sshClients

To make sure your Rstudio is set up correctly, choose File, New File, R Markdown, and click pdf. Now use the Knit pulldown to knit to pdf. If the document renders correctly, you are good.

Problem 2

We will be using GitHub in all future assignments. To prepare for this, you need to do two things.

Part A

Get a GitHub account if you didn't in Problem 1. This sometimes works best if your user name is the same as your local computer login user name.

Part B

Set up ssh keys. In R, go to Tools –> Global Options. Click on the Git/SVN icon in the window that pops up. If you havn't messed with SSH keys, go ahead and click generate RSA key. Once done, click on view public key and then copy the key (everything in the window).

You now need to add this to your GitHub profile. Click on the pulldown to view your profile and choose settings then SSH and GPG keys. Click on New SSH key and paste you RSA public key there.

Please include a link to your github page.

Problem 3

R is an open source, community built, programming platform. Not only is there a plethra of useful web based resources, there also exist in-R tutorials. To speed our learning, we will use one such tutorial *swirl*. Please install the *swirl* package, install the "R_Programming_E" lesson set, and complete the following lessons 1-3 and 15. Each lesson takes about 10 min.

From the R command prompt:
install.packages("swirl")
library(swirl)
install_course("R_Programming_E")
swirl()

Problem 4

Now that we have the R environment setup and have a basic understanding of R, let's add Markdown (choose File, New File, R Markdown, pdf).
a. In this new Rmarkdown file, please type a paragraph about what you are hoping to get out of this class. Include at least 3 specific desired learning objectives in list format.
b. To this, add 3 density functions (Appendix Cassella & Berger) in centered format with equation number, i.e. format this as you would find in a journal.

Problem 5

A quote from Donoho (1995): "an article about computational results is advertising, not scholarship. The actual scholarship is the full software environment, code and data, that produced the result." To the document created in Problem 3, add a summary of the steps in performing Reproducible Research in numbered list format as detailed in:
http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003285.

Next to each item, comment on any challenges you see in performing the step. If you are interested in learning more, a good summary of why this is important can be found in
- https://www.informs.org/ORMS-Today/Public-Articles/October-Volume-38-Number-5/Reproducible-Operations-Research
- https://doi.org/10.1093/biostatistics/kxq028
- http://statweb.stanford.edu/~wavelab/Wavelab_850/wavelab.pdf

Problem 6

Please create and include a basic scatter plot and histogram of an internal R dataset. To get a list of the datasets available use library(help="datasets").

This document containing solutions to Problems 3-5 should be typed in RMarkdown and Knit'd to create a pdf document turned in at the beginning of the next class.

Optional preperation for next class:

We will be using Git for future assignments, if you have time, please follow the git tutorial at:
- http://www.molecularecologist.com/2013/11/using-github-with-r-and-rstudio/
- https://try.github.io/levels/1/challenges/1