Homework 3

Due September 30, 2020

2020-08-27

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.

The last couple of weeks, we spoke about R, version control and Reproducible Research, munging and 'tidying' data, good programming practice, some basic programming building blocs, and finally matrix/vector operations. In this homework, we will put this all together and actually analyze some data. Remember to adhere to both Reproducible Research and Good Programming Practices, ie describe what you are doing and comment/indent code where necessary.

Problem 1

In the "Getting and Cleaning Data" lesson set, you should be comfortable with lessons 1-3. Work through the "R Programming E" lesson as you see fit. Lessons 1-9 and 15 are ones you should consider paying special attention to. If you prefer the Rstudio.cloud Primers, the Primer on "Write Functions" is well done.

From the R command prompt:

```
library(swirl)
install_course("R_Programming_E")
install_course("Getting_and_Cleaning_Data")
install_course("Exploratory_Data_Analysis")
swirl()
```

Problem 2

Create a new R Markdown file (file–>new–>R Markdown–>save as.

The filename should be: HW3_pid, i.e. for me it would be HW3_rsettlag

You will use this new R Markdown file to solve the following problems:

Problem 3

In the lecture, there were two links to programming style guides. What is your takeaway from this and what specifically are *you* going to do to improve your coding style?

Problem 4

Good programming practices start with this homework. In the last homework, you imported, munged, cleaned and summarized datasets from Wu and Hamada's *Experiments: Planning, Design and Analysis*.

Problem 5

A situation you may encounter is a data set where you need to create a summary statistic for each observation type. Sometimes, this type of redundancy is perfect for a function. Here, we need to create a single function which takes as input a two column dataframe and returns a vector containing

1. mean of column 1
2. mean of column 2
3. standard dev of column 1
4. standard dev of column 2
5. correlation between column 1 and 2

I will look at the code and comment on it, so make it NICE!!

We will use this function to summarize a dataset which has multiple repeated measurements from two devices (dev1 and dev2) by thirteen Observers. This file is preformatted as an R object, so it will read in nicely. "url <- https://github.com/rsettlage/STAT_5014_Fall_2020/blob/master/homework/HW3_data.rds". Please load the file (?readRDS – really nice format for storing data objects), loop through the Observers collecting the summary statistics via your function for each Observer separately and store in a single dataframe.

The output of this problem should be:

a. A single table of the means, sd, and correlation for each of the 13 Observers (*?kable*). From this table, what would you conclude? You can easily check your result using dplyr's group_by and summarize.

b. A box plot of dev, by Observer (*?boxplot*). From these plots, what would you conclude?

c. A violin plot of dev by Observer (*??violin* two "?" will search through installed packages). From these plots, what would you conclude? Compared to the boxplot and summary statistics, thoughts?

Now that you have made some conclusions and decided what your analysis may look like, you decide to make one more plot:

d. a scatter plot of the data using ggplot, geom_points, and add facet_wrap on Observer. For instance: `ggplot(df, aes(x=dev1,y=dev2)) + geom_point() + facet_wrap(Observer~.)`

What do you see? Combining the scatter plot with the summary statistics, what is the lesson here? As you approach data analysis, what things should you do in the "Exploratory Data Analysis" portion of a project to avoid embarrassment from making erroneos conclusions?

Problem 6

Some numerical methods are perfect candidates for funtions. Create a function that uses Reimann sums to approximate the integral:

$$f(x) = \int_0^1 e^{-\frac{x^2}{2}}$$

The function should include as an argument the width of the slices used. Now use a looping construct (for or while) to loop through possible slice widths. Report the various slice widths used, the sum calculated, and the slice width necessary to obtain an answer within $1e^{-6}$ of the analytical solution.

Note: use good programming practices. For help on Reimann sums:
https://www.khanacademy.org/math/ap-calculus-ab/ab-integration-new/ab-6-2/a/left-and-right-riemann-sums

Problem 7

Create a function to find solutions to (1) using Newton's method. The answer should include the solutions with tolerance used to terminate the loop, the interval used, and a plot showing the iterations on the path to the solution.

$$f(x) = 3^x - sin(x) + cos(5x) \tag{1}$$

For a refresher on Newton's method:
https://en.wikibooks.org/wiki/Calculus/Newton%27s_Method

Problem 8

In most of your classes, you will be concerned with "sums of squares" of various flavors. SST = SSR + SSE for instance. Sums of square total (SST) = sums of square regression (SSR) + sums of square error (SSE). In this problem, we want to compare use of a for loop to using matrix operations in calculating these sums of squares. We will use data simulated using:

```
X <- cbind(rep(1,100),rep.int(1:10,time=10))
beta <- c(4,5)
y <- X%*%beta + rnorm(100)
```

Without going too far into the Linear Regression material, we want to calculate SST =

$$\Sigma_{i=1}^{100}(y_i - \bar{y})^2$$

Please calculate this using:

a. accumulating values in a for loop

b. matrix operations only

Note, you can precalculate mean(y) and create any vectors you need, ie perhaps a vector of 1 (ones). In both cases, wrap the calculation in the microbenchmark function. Report the final number and timings.

Problem 9

Finish this homework by pushing your changes to your repo.

Only submit the .Rmd and .pdf solution files. Names should be formatted HW3_pid.Rmd and HW3_pid.pdf