Stat243: Problem Set 2, Due Friday Oct. 4

September 27, 2013

This covers material in Units 3 and 4 on R.
It's due at the start of class on 10/4.
Some guidelines on how to present your solutions:

1. Please use your knitr/Sweave/R Markdown solution from PS1, problem 5 as your template for how to format your solutions (only non-Statistics students are allowed to use R Markdown).
2. As with PS1, your solution should not just be code - you should have text describing how you approached the problem and what the various steps were.
3. In addition to the paper submission, which is the primary thing we will grade, please turn in your raw Latex/Markdown with embedded code electronically on bSpace (a file just containing the code with indication of what pieces of code are for what problems is fine too).
4. All your code should have comments indicating what each function or block of code does, and for any lines of code or code constructs that may be hard to understand, a comment indicating what that code does. You do not need to show exhaustive output but in general you should show short examples of what your code does to demonstrate its functionality. Depending on how long different pieces of your code are, you may want to have some of it be an appendix rather than part of the main solution.
5. Use functions as much as possible, in particular for any repeated tasks. We will grade in part based on the modularity of your code and your use of functions.

Please note my comments in the syllabus about when to ask for help and about working together.

Problems

1. Some questions about scoping.

   (a) Explain why the result of this is a vector of length 8.

```
a <- 8

genFun1 <- function() rnorm(a)

wrapFun1 <- function(data) {

    a <- length(data)
```

```
    return(genFun1())

}

wrapFun1(1:3)

## [1]  0.97037 -0.75049 -0.57684 -0.41642  0.07364  1.62472  0.34207 -1.59
```

(b) Now explain why the result of this is a vector of length 3.

```
a <- 8

genFun2 <- function(x) rnorm(x)

wrapFun2 <- function(data) {

    a <- length(data)

    return(genFun2(x = a))

}

wrapFun2(1:3)

## [1]  0.81928 -1.09989 -0.09599
```

(c) And why this is a vector of length 16.

```
a <- 8

genFun3 <- function(x = a * 2) rnorm(x)

wrapFun3 <- function(data) {

    a <- length(data)

    return(genFun3())

}

wrapFun3(1:3)

##  [1]  0.3291  0.2093 -0.4116  0.9334  0.7304 -1.3223  0.4650  0.6871
##  [9]  0.9921 -0.4312 -0.6980  0.6175 -1.0481 -1.1773  0.2598  2.0397
```

2. Consider the result of running the following code in R.

```
sapply(0:3, function(x) {

ls(envir = sys.frame(x))

}

)
```

Interpret the output in light of our discussion of frames. Discuss what functions are being called, in what order, and what objects exist in the frame of each function. What are the numbers and names (the environment name in hexadecimal (base 16)) of each frame (note the names will change each time you run the code)? You may want to use *debug()* , *trace()* or *browser()* or to step through the code and see what is being executed when you call *sapply()*.
Your answer should include the code you wrote to figure out the environment names and the result of running that code.

3. The file */scratch/users/paciorek/PUMS5_06.TXT.bz2* on the SCF filesystem is a zipped file containing household and person-specific records for California from the 2000 US Census, in particular a sample of 5% of the population. The meta data describing the file format is at http://www.census.gov/prod/cen2000/doc/pums.pdf - look for the data dictionary section to interpret the format of the rows in the data file. Your job is to take a random sample of n observations just from the household-level records (where the user can specify n), saving the result as an R data frame with the following columns extraced from the Census data file: BEDRMS, FINC, NPF, ROOMS, HHT, P18, P65. Format the fields of your data frame in a meaningful way - e.g., use factors for categorical variables as appropriate and for a variable such as FINC, you will probably want to have two columns, since FINC is a combination of a categorical and a numeric variable. Give your columns meaningful but concise names.
Some rules for your solution:

   (a) You must not unzip the file when you read the data in. This is to mimic the situation where the file is too big when unzipped to fit on the hard drive. For purposes of writing your code, you can unzip the file to look at it, but note that "less" will probably show you the contents without unzipping it. If you would like to deal with the subsetting to the household records with UNIX command line tools, you can do this (and it may make your life easier, but you should figure out how to do this via piping without ever creating either the original or the subsetted file in unzipped form. bunzip2 and bzip2 can deal with compressed files in the bzip (.bz2) format.

   (b) You must only read the n observations into R. You are not allowed to read the entire file in, nor are you allowed to read it in in blocks that contain observations other than those that you will retain.

   (c) Please make sure to write your code modularly. E.g., you might have an R function, say, 'sampleFile()' that does the sampling of rows from the file, given an input of the number of observations to sample and the name of the file. Then you could have a separate step of processing that processes a row to extract the elements of the row. Try to think of what the distinct tasks are and do each task as a separate function.

   (d) Make sure to set the random number seed using *set.seed()* so your sample is reproducible.

4. DRAMATIS PERSONAE: INSTRUCTOR, STUDENTS, GSI
   ACT 1. SCENE 1
   UC Berkeley
   INSTRUCTOR. [To STUDENTS] I set before thee the task of processing the plays of Shakespeare, to scrape as it were, in the manner of combing wool from a sheep. [To GSI] This should keep them busy.
   STUDENTS. [With gnashing of teeth] Nay, nay, thou shall not set us to this mighty task!
   ...
   ACT Post-Berkeley
   STUDENTS. [To INSTRUCTOR] We thank thee - you have set us upon a golden path, where beautiful data lie open to us in every nook of this World Wide Web.

   Translation:
   Your job is to extract data from the complete works of William Shakespeare, available as plain text at http://www.gutenberg.org/cache/epub/100/pg100.txt. You'll want to use functions from the *apply()* family as well as write your own functions to particular tasks.
   Note that for part (c), I relied on indentation to indicate the beginnings and continuations of spoken chunks. The fourth play (The Comedy of Errors) has different indentation than the others, so in my solution I excluded it. You are allowed to exclude it or alternatively to exclude a small number of other plays that make your processing particularly difficult.

   (a) Extract the plays (skip the information at the start of the file, the sonnets, and the last piece (Lover's Complaint)) into a character vector or a list, with one play per element.

   (b) Extract meta data about each play and the body of the play. The result should be in the form of an R object, with one play per element. By meta data I mean the year of the play, the title, the number of acts, and the number of scenes.

   (c) Extract the actual text spoken by the characters into your object. You should store each chunk of spoken text and its speaker. Discard stage directions, Dramatis personae information, and scene information.

   (d) Now use the constructed data object to calculate summary statistics about each play. These should include the following. When extracting words, make sure you remove punctuation such as commas and periods but not apostrophes.

      i. The number of unique speakers.
      ii. The number of spoken chunks.
      iii. The number of sentences and words spoken and average and standard deviation of number of words per chunk.
      iv. The number of unique words, as well as the most common 'meaningful' words, ignoring non-meaningful words, as specified here: http://www.textfixer.com/resources/common-english-words.txt.
      v. The average and standard deviation of the length of the words spoken in the play.

   (e) Plot some of your summary statistics as a function of time to see if there are trends in Shakespeare's plays over the course of his writing career.

   (f) Extra credit: Do some additional research and/or additional thinking to come up with additional variables that quantify the plays in interesting ways. If relevant, do some plotting that illustrates how the speeches have changed over time in terms of these new variables. Alternatively, particularly clever or thorough treatments of the problem may earn extra credit, such as avoiding

excluding any plays or sophisticated treatment of songs and other items that are hard to handle.

HINTS: In my processing, I did a bit of preprocessing in UNIX and also did some preprocessing on a single character string containing the whole file in R. This allowed me to clean up some of the inconsistent formatting (e.g., "ACT 1" vs. "Act I."). It also allowed me to insert some special symbols of my own that later allowed me to split the chunks of spoken text more easily.