Intro to R for Epidemiologists

Lab 2 (1/22/15)

Many of these questions go beyond the information provided in the lecture. Therefore, you may need to use R help files and the internet to search for answers. Feel free to ask questions of the instructor, the TAs, or your classmates, but try to work through as much as you can independently.

For the lab, you are expected to create an R script (.R file in the R editor) with your code corresponding to each question. Begin each question with a commented line of code indicating the question. As an example:

```
# Jenna Krall

# Question 1.
head(iris)
```

Part 1. Google flu trends

This data, introduced in lecture, estimates 2013 US flu activity using google searches. The data includes estimated daily flu activity for the whole United States, Georgia, Atlanta, and Health and Human Services Region 4 (which includes Alabama, Florida, Georgia, Kentucky, Mississippi, North Carolina, South Carolina, and Tennessee). Google published their results in Nature (http://www.nature.com/nature/journal/v457/n7232/full/nature07634.html). An article in Science also discusses problems with using this data to predict flu activity (http://www.sciencemag.org/content/343/6176/1203).

Data Source: Google Flu Trends (http://www.google.org/flutrends)

1. Download the dataset "googleflumissing.txt" from www.jennakrall.com/IntrotoRepi/googleflumissing.txt . Open the file in a text editor (e.g. TextEdit on a Mac, Notepad in Windows). What separates the entries in this dataset? See the help page in R for `read.table` and read this dataset into R.
2. Find the mean Georgia flu activity. What additional argument do you need?
3. Apply the `is.na` function to Georgia flu activity. Save the results of `is.na` to a new vector `na_georgia`. What class is `na_georgia`?
4. Take the mean of your vector `na_georgia`. What value does it give and what does this correspond to?
5. Create a vector `notna_georgia` that indicates the values that are not missing Georgia flu activity. Use your vector to create a vector of the Georgia flu activity for only those days that are not missing.
6. Create a revised Google flu dataset to include only the days where Atlanta had at least 1,000 flu cases.
7. Find the mean flu activity in Georgia when the Atlanta flu activity is above its median.

   · Hint: break this question down into steps,

   a. Find the median of Atlanta flu activity
   b. Find the rows when Atlanta flu activity exceeds its median
   c. Create a vector of Georgia flu activity with the rows in (b.).
   d. Take the mean of the vector in (c.).

8. The flu season is generally from October through April. Create a new variable indicating when it is flu season. Specifically, your variable should be TRUE when it is flu season and FALSE when it is not. Hint: To create a date, use `as.Date`. You may also need to look again at logical operators in R (`?"|"`).

Part 2. FIFA World Cup goals

The FIFA (Fédération Internationale de Football Association) World Cup is an international men's soccer (football) competition held every four years. We are interested in the total number of goals scored in the past four touraments for the USA, Germany, Brazil, and Italy.

1. Create a matrix in R of total goals scored that has the following information,

$$\begin{bmatrix} 7 & 14 & 18 & 5 \\ 2 & 14 & 10 & 12 \\ 5 & 16 & 9 & 4 \\ 5 & 18 & 11 & 2 \end{bmatrix}$$

   Hint: Use the `matrix` function.

2. Name the columns of your matrix "USA", "Germany", "Brazil", "Italy." Name the rows "2014", "2010", "2006", "2002" using the `seq()` command.

3. Remove the column corresponding to Brazil and print the new matrix.

4. Provide the mean and median of total goals scored over the entire time period for Germany.

5. Order the matrix you created so that Italy's goals are organized in decreasing order.