R Homework #3

Correlations & Bivariate Regression Analysis

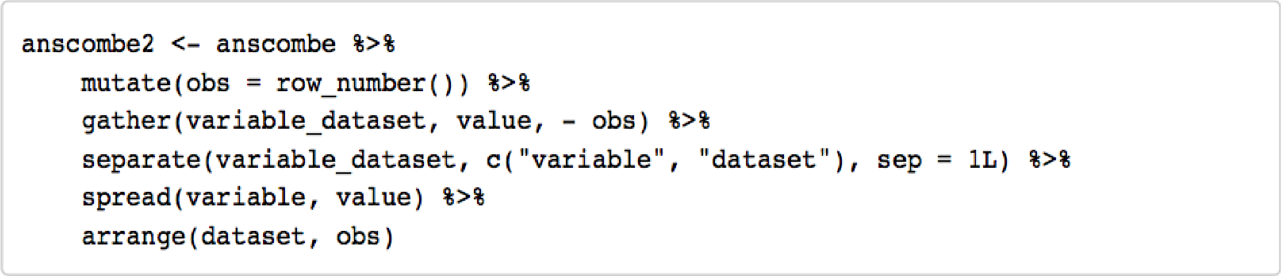
Due: November 27th, 2018 via email before lab; you are also expected to hand in a printed version of script and plots in lab. No late work will be accepted.

Important Details:

- Once again: three files. A script file, .png files of your plots, and a text/word document with responses to questions #5, #6, and #7.
- You may (indeed, I would encourage you to) work with other fellows to solve these problems, but you should list everyone you worked with as an annotation at the top of your R script file (e.g. #Collaborated with Dan and Kim).
- You may also visit the consultants at the Center for Social Science Computation and Research during their copious and friendly drop in hours (http://csscr.washington.edu/)
- You may use google to discover answers to challenges or questions as you are coding.
- A reminder on replication: I should be able to run your script from beginning to end on my own machine and get all the same results that you did. This is what it means for code to be "replicable" – the script file needs to contain everything for me to replicate your results and plots, without me doing any typing.
- You will be graded on – from most to least important – the replicability of your code, the clarity of your code and annotations, the elegance of your plots, and the thoughtfulness and thoroughness of your written responses.

Start a new script file and name it "lastname_HW3.R".

1. Load the dataset "anscombe" from base R using the data( ) function. Load dplyr, tidyr, and ggplot2.
2. Clean this data and make it tidy. Try to do this on your own, and then work from the code below as needed. In annotation, explain clearly what each line is doing (you may need to consult help files and google):



```
anscombe2 <- anscombe %>%
    mutate(obs = row_number()) %>%
    gather(variable_dataset, value, - obs) %>%
    separate(variable_dataset, c("variable", "dataset"), sep = 1L) %>%
    spread(variable, value) %>%
    arrange(dataset, obs)
```

3. There are four datasets within "anscombe2"; these are identified in the "dataset" variable.
    1. Calculate the mean and the standard deviations of X and Y for each dataset using the "group_by" and "summarise" functions; print a 4x5 table of these values in the console.

   2. Calculate the correlation between X and Y using “group_by” and “summarise”; print a 4x2 table of these values in the console.
   3. Run a bivariate linear regression between X and Y for each dataset; report the coefficient estimate for X and the p-value for that coefficient, for each of the four models.
   4. How do these datasets compare in terms of their descriptive statistics, correlation, and regression results? In layman's terms, do they appear roughly similar?

4. Create four plots – one for each dataset – where you plot both a scatter plot of the values of X and Y for that dataset and also overlay a regression line. Your plots should be clean – i.e. you are expected to change labels for axes, add titles, adjust axis tick marks, etc. to make elegant plots that you would feel comfortable putting in a final paper. (For extra credit: produce these four plots in one block of code, as one outputted .png file, using the facet argument.)
5. Based on your plots, what kinds of conclusions about the similarity between these data sets can we draw now? What kinds of concerns does this raise as you think about running regression analyses to answer your real-life research questions?
6. Explain, in 5-7 sentences, what a bivariate regression is and what it is doing. You should describe it clearly and fully as if speaking to someone without any statistical background.
7. Building on the last homework's discussion of control variables, list out all of the control variables you expect to include in your analysis. For each variable, list the data source you expect to use. (Some control variables may be in the data sets where you found your dependent or independent variables; other control variables may require that you find/generate new datasets.)