R Homework #4

Multivariate Regression Analysis

Due: December 4th, 2018 via email by 7 pm; you are also expected to hand in a printed version of script and plots to me on December 4th – since there is no lab on December 4th, please leave printed copies on my desk in the CAPPP suite. 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 question #2.
- 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_HW4.R".

#0. Load ggplot, lindia, tidyr, and dplyr. Import the democracy data I've sent to you using this code:

```
democracy <- read.csv(file = "hw4_democracy.csv",
                      header = TRUE,
                      stringsAsFactors = FALSE,
                      na.strings = ".")
```

#1. Look at the data. It contains the following variables:

- COUNTRY: numeric code for each country
- CTYNAME: name of each country
- REGION: region of each country
- YEAR: data year for a given country
- BRITCOL: 1 if former British colony; 0 if not
- CATH: percent of population that is Catholic
- CIVLIB: range of civil liberties score from 1 to 7, where 1 is least free and 7 is most free
- EDT: years of education of the average member of the labor force
- ELF60: index of ethnolinguistic fractionalization in country, where 0 is homogenous and values

closer to 1 represent more multi-ethnic or multi-lingual countries
- GDPW: real GDP per worker
- MOSLEM: percent of population that is Muslim
- NEWC: new country (independent after 1945) = 1; otherwise, 0
- OIL: 1 if ratio of fuel exports to total exports exceed 50%; otherwise, 0
- POLLIB: range of political liberties score from 1 to 7, where 1 is least free and 7 is most free
- DICTATOR: if country under authoritarian regime, 1; otherwise, 0

Select one dependent variable from among the following variables: GDPW, EDT, CIVLIB, POLLIB. Select one single year to study. Create a new dataset with only your chosen year. How many rows do you have? Next, use the drop.na( ) function in tidyr to remove all rows in your data that have missing information. How many rows do you have now?

#2. Develop a research question that explains the variation in this dependent variable in your chosen year, and that can be answered using this data set. Articulate the independent (one variable), dependent (one variable from above list), and control variables (at least two), and briefly (2-3 sentences) describe your theory and your hypothesis. Lay out the null hypothesis.

#3. Plot a scatter plot of the dependent and independent variables.

#4. Test your research question using multivariate regression. Report the results of your model, including a discussion of the statistical significance and substantive effects of the independent and control variables on your dependent variable, and your adjusted R squared value. Do you have sufficient evidence to reject your null hypothesis? Why or why not?

#5. Is OLS appropriate for your data? Consider the five assumptions of OLS, and present evidence that all five of those assumptions hold for your data.