Homework 8-10 – Intro. to Computational Statistics

This assignment covers selected topics from three modules, and counts for double the usual assignment. Be sure to get started early so that any initial data difficulties can be resolved in time to do the analysis. I would suggest trying to get at least half done in the first week just to make sure you have time for the rest in the second week.

For this assignment, I'd like you to find and use some data of your own: select a topic of substantive interest to you and a dependent variable you would like to explain, find some existing data on the internet that pertains to that topic, and test your hypotheses using multiple regression. (Or do it the opposite way: find some convenient data, and formulate some interesting questions to ask of it.) This is not a full-scale paper by any means, though – relatively brief answers to each question are ok.

For example, two potential sources of many datasets from my own field (social science) include ICPSR, at http://www.icpsr.umich.edu/icpsrweb/ICPSR/index.jsp , and Dataverse, at https://dataverse.harvard.edu/ . But there are many other convenient data sets available online, and hopefully you can find something in a domain of interest to you.

One last thing: please don't use temporal data (where the observations are taken over different times), and please don't use a dataset where your dependent variable is categorical (including binary); both require more complex methods than basic linear regression.

1. Describe your substantive interest and the general questions(s) you would like to answer (eg, "Does more education cause people to become more liberal?"). Be sure to frame it in a such a way that you are proposing a hypothesis (or multiple hypotheses) that might be either confirmed or disproven by the results of your analysis.
2. Describe the data set you have found, including its source, its contents, and why it was collected originally.
3. What is your dependent variable? Why are you interested in explaining it? What do you hypothesize are the major factors that influence or cause it?
4. What are your independent variables, and why have you chosen these? Prior to running your regression, what effects do you expect them to have on the dependent variable? Which of these variables do you think affect other of the independent variables, and how might that affect your final results?
5. Explain and show in detail how you rename and recode the variables you are examining, and what units each are measured in.
6. Before running a multiple regression, run a few bivariate regressions of Y on some of your X variables. What do you infer? Which of these do you think might change with the addition of multiple variables?
7. Run your full multiple regression using `lm()` and present your results using the output from the `stargazer` R package. Interpret the coefficients. What do they tell you substantively? Which variables seem to have the biggest substantive impact? Which ones could you actually change with some intervention, and how big a difference do you think that could make?
8. How have any of the coefficients changed from the bivariate regressions? What can you infer from that? How do you think your various independent variables interact and affect each other? Try to find an example where a variable appears signficant in the bivariate regression, but not in the full regression. Is this an example of a spurious or a chained causal pathway?

9. How does what you see match, or not, your hypotheses from (4)? Why did/didn't it match what you expected?
10. What do the R^2 and adjusted R^2 tell you about your model?
11. How would you use one of the variable selection methods to choose a model with fewer variables? Select one of the methods (either one of the stepwise or criterion-based methods) and show which variables it would lead you to keep. Do you agree with its results?
12. What are your overall conclusions? What are the weaknesses of your results, and how could you improve them with better or different data?
13. Calculations (using R):
    a. Derive the coefficients from your regression using the $(X'X)^{-1}X'Y$ formula. (If you run into problems using `solve()`, try using `ginv()` instead, which does the same thing but is a bit more robust.)
    b. For one of the coefficients, confirm its p value as shown in the regression output using the coefficient, its standard error, and `pt()` in R.
    c. Calculate the R^2 and adjusted R^2 using R, and confirm that your results match the regression output.
    d. Calulate the F statistic using R and confirm it against the regression output.
14. Add at least one quadratic term into your model and interpret the results. Is it significant? What is the effect of a 1-unit increase in that variable at its mean value?
15. Add at least one interaction term to you model and interpret the results. Is it significant? What is the effect of a 1-unit increase in one of those interacted variables holding the other at its mean value?
16. Test either the model in 14 or the model in 15 using the F test for nested models. That is, estimate the full model with the variable and quadratic term, or the variable and interaction, and then estimate the reduced model without either, and run the F test to establish whether those variables significantly improve your model.