Assignment 2, SDGB 7840: Multiple Regression

Instructor: Prof. Nagaraja

Due: 3/1 in class

Submit three files: (a) .Rmd R Markdown file with explanations and code, (b) Word document of knitted R Markdown file, and (c) your data file. Email both files to fordhamR-computing@gmail.com by the start of class with the subject line "HW2-[Full Name]-[Class Time]" and include HW 2 and your name in the file names (time of class is either 3:30 PM or 5:45 PM). Please email your solutions only once! Comment your code for full credit and complete the assignments individually. Use the materials in the "Assignments" folder for more on how to use RMarkdown and how to set up your assignment.

The Gini index (also called the Gini coefficient) is a measure of inequality, most often used to describe income inequality in populations. In this assignment you will use multiple regression to model the Gini index across countries. The goal is to understand which factors might be related to income inequality and how. Your grade will depend on both your analysis and the quality of your report.

Use the data provided by the World Bank to determine which explanatory variables to include (link to all World Bank data: `http://data.worldbank.org`; and also a link to Gini index data: `http://data.worldbank.org/indicator/SI.POV.GINI`). (This data requires a lot of cleaning.)

Write your paper as a research report. It should be no longer than 8 pages (this includes graphs and tables but not references) and should include the following six sections. (You can use the posted paper, "The regional dimension of MNEs' foreign subsidiary localization" by Arregle, Beamish, and Hebert (2009) as an example of how to determine what information is important to include in a report.)

1. Executive Summary: short paragraph summarizing your paper (this is like the business version of the abstract in the Arregle, et al paper).
2. Introduction: define the Gini index and the purpose of the study

3. Data: source of your data; discuss which 10 explanatory variables you considered and why (not just the ones you ended up including in your final model); relevant summary information about the explanatory and response variables; which countries are included in your data set; which year(s) are included in your data set; how you cleaned your data

4. Methods: relevant plots; model building summary (transformations, variable selection, collinearity, etc.); check regression assumptions (for now, leave out multicollinearity and autocorrelation since we haven't discussed those topics yet); model evaluation; relevant hypothesis tests (include hypotheses, test statistic, degrees of freedom, p-value, α value and conclusion)

5. Discussion: final regression model; interpretation of model; discussion of usefulness of model; any ideas you have for improvement

6. References: cite the World Bank as your data source and any other publications you may have used to learn about the response variable or which explanatory variables may be helpful, etc. (Note: you do not have to use other sources, but if you do, cite them.) Your reference list is not included in the 8 pages. Finally, DO NOT COPY TEXT from sources; write your report in your own words and add citations.