Pol Sci 630: Problem Set 9 - Data Management

Prepared by: Jan Vogler (jan.vogler@duke.edu)

Due Date: Tuesday, October 27th, 2015, 10 AM (Beginning of Class)

It is absolutely essential that you show all your work, including intermediary steps, and comment on your R code to earn full credit. Showing all steps and commenting on code them will also be required in future problem sets.

Please use a *single* PDF file created through knitr to submit your answers. knitr allows you to combine R code and LaTeX code in one document, meaning that you can include both the answers to R programming and math problems. Also submit the source code that generates the PDF file (i.e. either .Rnw or .Rmd files)

Make sure that the PDF files you submit do not include any references to your identity. The grading will happen anonymously. You can submit your answer at the following website: `http://ps630-f15.herokuapp.com/`

R Programming

Problem 1 (4 points)

Do the following in R:

a) Load the *LDC_IO_replication* dataset that was introduced in the tutorial. Create a new character variable that differentiates between democracies, anocracies, and autocracies.

Note: According to the Polity IV classification, a country is a democracy when it has an overall Polity IV score of 6 or higher, an autocracy with an overall score of -6 or lower, and anocracy for all remaining scores.

b) Building upon problem set 5, problem number 2, please estimate a regression that estimates a curvilinear relationship between net changes in FDI as percentage of GDP as

dependent variable and the Polity IV Score (lagged by 1 year) as independent variable. Include country fixed effects and show the summary of your regression.

Then, show the relationship between the Polity IV Score (lagged by 1 year) and FDI as percentage of GDP graphically for Algeria.

Problem 2 (6 points)

Do the following in R:

a) Find and download the most recent version of the *National Accounts Data (na_data) by the Penn World Tables* and its PDF documentation. Copy it into the same folder in which you work on the problem set. After you have loaded the dataset into R, display the summary statistics.

Note: If you have the dataset in the same folder as your .Rnw file, you do not need to set a working directory when you compile your PDF. Note that you are not supposed to set a working directory as this might reveal your identity to the grader.

b) Create a new variable named *gdpgrowth*. This new variable is meant to capture the real increase in economic output (gross domestic product, GDP) in a country compared to the last year in **percentage points (not in proportions)**.

Note 1: The real increase in economic output is based on constant prices not current prices. If current prices are used, then inflation could obscure the true increase in economic productivity. Make sure that you read the codebook carefully and use the correct variable.

Note 2: A proportion of 0.5 is 50 percent. The variable you create is meant to be coded in percentage points, not proportions.

c) We are interested in three countries from the *LDC_IO_replication* data set, namely Turkey, South Africa, and Mexico. **For these three countries only**, merge the *LDC_IO_replication* dataset with the *National Accounts* dataset. Then estimate a linear regression model with tariff level as dependent variable and the Polity IV Score (lagged by 1 year) and your new GDP growth variable as independent variables. Also use country-fixed effects. Show the summary of the linear regression.

Final Research Paper: Data Management and Omitted Variable Bias Questions

Problem 3 (6 points)

Please answer the following questions.

The goal of this problem is that you think carefully about the data that is available for your research project, especially how you can merge separate data sets. If data for different variables in your dataset is available for very different actors and time periods, this will create obstacles for your analysis. Additionally, part of this problem is a careful description of potential biases in your model that result from the omission of specific control variables (i.e. omitted variable bias).

Please state which variables you intend to use as variables in your final paper. Explain briefly (2-3 sentences) how these two variables are linked theoretically from your perspective. Then answer the following questions:

1. Which dataset(s) contain these variables and how are they coded there? Is there any need to transform or recode the variables for your analysis?
2. How many units and which time period do the data available cover?

Then, identify at least two (and up to four) important control variables that are not included in the same dataset as your dependent and/or independent variable. Explain briefly (2-4 sentences), with references to the literature that your paper is built upon and the statistical concept of omitted variable bias, why each of these control variables is important for your analysis. Additionally, for each of the control variables, please answer the following questions:

1. Which dataset contains the variable?
2. How are the variables coded in the dataset? Can they easily be used for linear regression or do they have to be transformed?
3. Are the data available for the same units and the same time period as your dependent and independent variables?

4. Are units and time coded in the same way in your control variable datasets? If not, how are they coded differently? How difficult will it be to merge the data sets?

Note: It is perfectly fine if you build upon and extend the answer that you gave in *Problem Set 7: Short Research Outline*.