Pol Sci 630: Problem Set 8 - Data Management and Omitted Variable Bias

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

Due Date: Wednesday, October 26th, 2016, 1.25 PM (Beginning of Class)

Note 1: It is absolutely essential that you show all your work, including intermediary steps, in your (mathematical) calculations and that you comment on your R code to earn full credit (you can comment on your R code both with the use of # in the R code and in the LaTeX code). Showing all steps and commenting on code will also be required in future problem sets.

Note 2: Please submit a PDF file created through knitr containing all your answers to the problem set. 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. the .Rnw file).

Note 3: 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 ordinal-level, 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 with 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.

Note: Make sure that you add +10 to the Polity IV Score to estimate the curvilinear relationship correctly.

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

Problem 2 (4 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* (Stata format) and its PDF documentation. Copy it into the same folder in which you work on the problem set. Load the dataset into R and 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 represent the real increase in economic output (gross domestic product, GDP) of 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* 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 OVB

Problem 3 (4 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 variables you plan to use in your research is available for different units 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 (OVB).

a) Please state which variables you intend to use as dependent variable and key independent variable 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 a statistical analysis (linear regression)?
2. How many/which units and which time period do the data available cover?

Note: It is not necessary to list all the units for which the data is covered. It is sufficient to provide an overview. For example, one could write here: "Data is available for the major developing countries, such as India and China, but not the OECD countries. The data covers the period 1970-1999 for most countries but shorer time periods for others."

b) 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 (2-3 sentences for each), with references to the literature that your paper is built upon and (if applicable) 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(s) contain these variables and how are they coded there? Is there any need to transform or recode the variables for a statistical analysis (linear regression)?
2. Are the data available for the same units and the same time period as your dependent and key independent variables?
3. Are units and time coded in the same way in your control variable datasets? If not, how are they coded differently?

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

Statistical Theory: Omitted Variable Bias

Problem 4 (4 points)

Please answer the following questions.

a) Write down the formula of your empirical model for your final paper, including your key independent variable and up to four control variables that you identified in the previous section.

b) Assuming that there is omitted variable bias for **one** of the control variables, please explain (for this variable only) using mathematical notation: (1) which influence the control variable potentially has on your dependent variable and key independent variable. (2) Which consequence the omission of this variable would therefore have, i.e. upwards or downwards bias on the estimated coefficient of your key independent variable.

Note 1: For this problem, it does not matter whether or not the control variable has this influence in reality. The purpose of this assignment is to illustrate the effect of omitted variable bias using mathematical notation.

Note 2: Do not just state whether there is upward or downward bias but also support your statement with a (brief) mathematical explanation.