Pol Sci 630: Problem Set 9 - Data Management, Omitted Variable Bias, and Endogeneity

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 variable that differentiates between democracies, anocracies, and autocracies. According to the Polity IV classification, a country is a democracy with an overall 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 run a regression that estimates a curvilinear relationship between net changes in FDI as percentage of GDP and the

Polity IV Score. Interpret the results. Then, show this relationship 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 by the Penn World Tables* and its PDF documentation. Copy it into the same folder in which you work on the problem set.

Note: If you have the dataset in the same folder as your .Rnw file, you do not need to set a working directory to read the file. 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*. The 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 (minimum value: 0, maximum value: 100).

Note: 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.

c) Merge the *National Accounts Data by the Penn World Tables* and the *LDC_IO_replication* datasets. Then estimate the *main* linear regression model we estimated in tutorial 5, section 3 again, this time including both country fixed effects and your newly created measurement of GDP growth as control variable. Interpret the results with respect to GDP growth. What impact, if any, does the inclusion of GDP growth have on the significance of the Polity IV coefficient?

Final Research Paper Data Questions

Problem 3 (4 points)

Please answer the following questions. Note: Please don't include R code in this answer as the file would not compile for other people.

a) Please state which variables you intend to use as variables in your final paper. Explain briefly (2-4 sentences) how these two variables are linked to each other theoretically from your perspective. (1.) What is your dependent variable? (2.) What is your independent variable? (3.) Which dataset(s) contain these variables and how are they coded there? (4.) How many units and which time period do they cover?

Identify at least one (and up to three) important control variables that is not included in the same dataset as your dependent and/or independent variable. Explain briefly (1-2 sentences) why each of these control variables is important for your analysis.

b) For each of the control variables that you have identified in part A of this question, please answer the following questions: (1.) Which dataset contains the variable? (2.) Are the data available for the same units and the same time period as your dependent and independent variables? (3.) Are units and time coded in the same way? If not, how are they coded differently?

Statistical Theory: Omitted Variable Bias and Endogeneity

Problem 4 (4 points)

Do the following problems. Show every step.

a)

b)

c)