Pol Sci 630: Problem Set 8 - Data Management, Measurement Error, 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 *VOTE1* dataset. Please estimate a regression with the incumbent vote share as dependent variable and the party strength of the incumbent and the expenditures

of both parties as independent variables. Please add a new variable to the model that allows you to estimate a curvilinear influence of incumbent party expenditures. Then, estimate this model, including the new variable, and display the summary of the regression.

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.

b) Show the relationship between incumbent party strength and incumbent vote share graphically, holding all other variables at their mean values.

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

Note 1: 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.

Note 2: 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.

Problem 2 (4 points)

Do the following in R:

a) Download the *National Accounts Data (na_data) by the Penn World Tables* and its PDF documentation which you can find on the course website under "Meetings". 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.

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, Measurement Error, and Omitted Variable Bias

Problem 3 (4 points)

Please answer the following questions.

Please note that we will not be able to completely maintain anonymity here.

The goal of this problem is that you think carefully about the data that is available for your final research project, especially how you can merge separate data sets. In general, you should ask yourself the following questions whenever you begin a new research project.

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 problems with your model and results that are due to measurement error or 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 shorter time periods for others."

b) Then, explain briefly (2-3 sentences for each) whether there could be measurement error in your data and why. If you believe that there could be measurement error in your variables, please make sure that you also answer the following questions:

1. Is the possible measurement error systematic or stochastic?
2. Which consequences can the measurement error have for your results?

If you do not believe that there could be measurement error in the specific variables that you intend to use, please generally describe the consequences of either a type of systematic or a type of stochastic measurement error.

c) Then, identify at least two (and up to four) important control variables (that are potentially 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 1: It is perfectly fine if you build upon and extend the answer that you gave in Problem Set 7.

Note 2: If you have a dataset for which you cannot gather additional data, such as a random sample of individuals that cannot be contacted again, please nevertheless think about potentially omitted variables and how the omission might affect your results.

Statistical Theory: Omitted Variable Bias

Problem 4 (4 points)

Please answer the following questions.

Assume that you want to find out what determines a country's military expenditures. You estimate the following model:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + u$

Where the variables represent the following concepts:

Y	Military Expenditures (Percent of GDP)
X_1	Regime Type (Polity IV)
X_2	External Military Threat
X_3	Militaristic Ideology
X_4	Size of the Arms Industry
X_5	No. of Armed Conflicts in the Last Decade

It is likely that militaristic ideology has an impact on both regime type and military expenditures. We would expect that it has a negative effect on regime type and a positive effect on military expenditures. If this expectation is true, what would happen if we omit X_3 from the regression? How would it affect our estimate of β_1 specifically? Show mathematically and explain carefully.