Pol Sci 630: Problem Set 11 - Imputation of Missing Data, Regression Diagnostics, and Simulations

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

Due Date: Tuesday, November 10th, 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:

Load the *LDC_IO_replication* dataset. Create a subset that contains the following countries only: Brazil, Jamaica, Kuwait, Hong Kong, and the Kyrgyz Republic. Then, keep only the variables in your dataset that you need for the following regression in this problem (problem 1).

Before you impute any missing data, estimate a linear regression with the following characteristics and show the results:

1. Dependent variable: net changes in FDI as percentage of GNP
2. Independent variables: tariff Level, GDP per capita, Polity score
3. There is no need to use lagged versions of the independent variables or country dummies in this regression.

Then, load the *Amelia* package and the *Zelig* package. In order to limit the length of the PDF document that we produce as homework PDF file, please hide this part of the code entirely by using the following command:

```
<<include=FALSE>>=
```

Use *Amelia* to impute the missing values of all variables in the dataset. In order to limit the length of the PDF document that we produce as homework PDF file, please hide the output of the code by using the following command:

```
<<results=’hide’>>
```

Once you have imputed the missing values, use *Zelig* to run the a new regression with the same characteristics as above and show the regression summary.

Then, answer the following questions: Has the imputation of data changed the results? Are the results that we obtain from the imputed data more or less accurate (i.e. closer to the truth) than the results that we got in the first regression?

Problem 2 (4 points)

Do the following in R:

Load the *LDC_IO_replication* dataset again to make sure that you are not limited to a subset like in the previous task.

Estimate a linear regression with the following characteristics and show the results:

1. Dependent variable: net changes in FDI as percentage of GNP
2. Independent variables: Polity score, GDP per capita, natural log of population
3. Please use lagged versions of the independent variables (lagged by 1 year).

4. There is no need to use country dummies in this regression.

For the regression above, create a component plus residual plot and interpret it.

Then, for the same regression, conduct a *Ramsey RESET Test for Functional Form Misspecification* and interpret the results.

Problem 3 (4 points)

Do the following in R:

Create a subset of the LDC dataset that has complete cases for the variables of the following regression. Then, estimate a linear regression with the following characteristics and show the results:

1. Dependent variable: tariff level
2. Independent variables: Polity score, signing of IMF agreements, the number of years a government has been in office, GDP per capita, natural log of population, economic crisis, balance-of-payment crisis
3. Please use lagged versions of the independent variables (lagged by 1 year).

Create 1000 simulations of this model. Then, use the model simulations to graphically show the linear effect of the Polity score (lagged by 1 year) on tariff levels under all estimated coefficient values. For all of these graphical representations, hold the other variables of the regression at their mean value.

Note: In the tutorial we held all other variables at the value 0. Here you are explicitly asked to hold them at their mean value.

Problem 4 (4 points)

Do the following in R:

Continue to use the full LDC dataset and the regression estimated in problem 3 (with lagged independent variables).

The goal of this problem is to compare the substantive effects of two different independent variables using model simulations and averaging over our entire data. Please create a new plot for the average predictive comparison of the effects of the following variables:

1. The Polity score
2. GDP per capita

Please compare the effect that each of the above variables has when it increases from its 25th percentile value to its 75th percentile value.

Please interpret the findings and make a statement about which variable appears to have the larger substantive effect on tariff level.