

Assignment 1**First draft is due for peer evaluation by October 15**

For this assignment, you will utilize a World Development Indicators (WDI) dataset for 2000, located at <http://www.sarkisian.net/socy7704/datasets.html>. You can find more information about this dataset at <http://www.worldbank.org/data>. Alternatively, you can use a dataset of your choice; make sure that the data are cross-sectional.

- 1) Devise a research question that you can answer using OLS regression on these data, with approximately 4-6 independent variables. Clearly state your research question as well as your hypotheses; include a brief justification for each hypothesis.

Start a running log that will contain the commands and output for all of your models and diagnostics, with brief comments. (In order to be able to do this, make sure you open a log file each time you open Stata to work on this assignment!)

- 2) Investigate the univariate distributions for each of the variables involved; consider normalizing transformations and deal with univariate outliers.
- 3) Investigate the bivariate relationships between the dependent and each of the independent variables; evaluate linearity and consider potential transformations; identify bivariate outliers.

Fit the OLS regression model (using the dependent and independent variables with those transformations and modifications you find appropriate). Then conduct the following diagnostics and attempt to remedy any problems that you find. This process can sometimes be circular as you may need to recheck things after applying new transformations etc.

- 4) Examine the model for potential nonlinearities.
- 5) Consider potential 2-way interactions. If any of these interactions are statistically significant, include them in your model.
- 6) Use hypothesis testing to evaluate whether you can omit non-significant variables, combine some categories of dummy variables (if relevant), and whether groups of dummies are jointly significant (also if relevant); try to make the model more parsimonious.
- 7) Check for possible multicollinearity problems using variance inflation factors.
- 8) Examine the residuals for departures from normality.
- 9) Identify possible outlying and influential data.
- 10) Examine residuals for possible heteroscedasticity. If you detect problems, determine whether any of the independent variables might be responsible.
- 11) Write up your interpretation of the results of your final model (using both unstandardized and standardized coefficients). If necessary, use graphs to assist the interpretation (graphs can be useful if your model includes nonlinear relationships or interactions.)
- 12) When submitting the assignment, make sure to include all the steps specified in items 1-11. There is no page limit for your assignment but please edit it to contain only the relevant commands and output, and include the relevant graphs as well (you can copy and paste them into your word processor).

Once your assignment is completed, please fill out and include the evaluation table that will contain a summary of what you did and will be used to evaluate your work, first by a peer and then by the instructor.