New York University
Wilf Family Department of Politics
Fall 2013

Quantitative Research in Political Science I

Professor Patrick Egan

PROBLEM SET 8: Due Monday, November 25 at beginning of class.

A reminder: you may work with others in the class on this problem set, and you are in fact encouraged to do so. However, the work you hand in must be your own. Your work must be word-processed in order for you to receive credit for the assignment.

1. Show (using matrix notation) that by construction in OLS
$$\mathbf{i}'\widehat{\mathbf{u}} = 0.$$

2. Show (using matrix notation) that by construction in OLS for any N x 1 vector of observations of the k'th regressor $\mathbf{x}_k$, it is the case that
$$cov\left(\mathbf{x}_k, \widehat{\mathbf{u}}\right) = 0.$$

For the rest of this assignment, use the *counties.dta* dataset found on our class Blackboard site. It comes from the *City and County Data Book* produced by the U.S. Census Bureau. Provide appropriate Stata output as an attachment to your assignment.

3. Analyze the following questions using Stata, but answer them with a few sentences in plain English.
   (a) Controlling for household income, is the percentage of a county's residents who have college degrees associated with crime rates? How so?
   (b) Construct a plot displaying the relationship between $\widehat{crimerate}$ and *college*. To do this, write a loop that generates the value of $\widehat{crimerate}$ when *college* is equal to 5, 7,...25. (In doing so, you may find it helpful to use the coefficients saved by the `regress` command in the vector `e(b)`.) Then plot these connected points.
   (c) You are interested in how well this model applies to counties in New York State. Construct a plot showing at a glance that Tompkins County, has a crime rate much lower than predicted by a regression of *crimerate* on *college* among counties in New York State.
   (d) Look up Tompkins County on the Internet. Why might it be an outlier? Does this suggest an additional control variable?

The following two questions require that you use two techniques not yet covered in class: OLS with indicator variables and with interaction terms. Many of you are already familiar with these techniques, and Andrew will be covering them in lab on this Friday. I encourage you to consult the cited sections of Wooldridge if needed. We'll cover them in detail next week in class.

4. Is unemployment higher in counties located in the South than in those located outside the South?

   (a) Answer this question first with a *t*-test to compare group means.

   (b) Now answer this question by running a bivariate specification regressing *unemprate* on the indicator variable *south.* If you are unsure about how to interpret the coefficient on *south* in this context, consult Wooldridge pp. 225-231.

   (c) Identify two important similarities found between the two analyses.

   (d) We required a lot more assumptions to perform the second analysis than the first. In a few sentences, reflect upon why they are unnecessary to answer the question posed at (4) above.

5. Is the association between poverty and crime by county stronger in the South or outside the South?

   (a) Answer this question first by running two separate regressions.

   (b) Now create a term *povxsouth* that incorporates the interaction between *south* and *povrate*. Run the proper analysis that includes the interaction term. If needed, consult Wooldridge pp. 238-243.

   (c) Identify two important similarities found between the estimates generated by the two analyses.

   (d) Construct a plot with two lines: one showing $\widehat{crimerate}$ by *povrate* among counties in the South, and the other showing $\widehat{crimerate}$ by *povrate* among non-Southern counties.

6. Perform the necessary analyses to assess the accuracy of the following statements. In a few sentences, discuss each of your findings.

   (a) The bivariate relationship between poverty and crime is weaker where local government spending is higher.

   (b) The bivariate relationship between unemployment and crime is stronger in densely populated counties.

   (c) Federal spending per capita plays a greater role in reducing poverty than local spending per capita, controlling for variables that may confound these relationships.

7. Pretend that you are a Marxist-Socialist who wants to show that crime is caused by poverty. To do so, you are hell-bent on finding the combination of regressors (including *povrate*) that results in a very large coefficient on *povrate* when *crimerate* is regressed upon these regressors.

   (a) Using the proper Stata commands, choose 90 percent of your observations at random and set them aside. Do not include these observations in the analysis.

   (b) Find the combination of variables that yields the highest possible coefficient on *povrate* as a predictor of *crimerate* that you can obtain. If you're feeling creative, you should be able to write a Stata loop that accomplishes this, but that is not necessary. Just have fun with it. [ONE MORE QUESTION ON NEXT PAGE]

(c) Once you've arrived at a specification, note the results. Now, take the rest of your dataset out of "cold storage." How well does your model do with these other data? What is the lesson learned?