New York University
Wilf Family Department of Politics
Fall 2013

Quantitative Research in Political Science I
Professor Patrick Egan

PROBLEM SET 9: Due Friday, December 13 at 10 a.m.

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. Consider three random variables X, Y and Z, where

$$\begin{aligned} cov(Z,Y) &< 0; \quad cov(Z,X) = 0; \\ \text{and } cov(X,Y) &> 0. \end{aligned}$$

Assume we have a large number of observations of the joint distribution of all variables from an i.i.d. random sample yielded by a data generating process governed by the linear model

$$y = \alpha + \beta x_i + \delta z_i + u_i.$$

If we estimate the equation

$$\widehat{y_i} = \widehat{\alpha} + \widehat{\beta} x_i,$$

(a) What is the formula OLS uses to generate $\widehat{\beta}$?

(b) What is $E\left(\widehat{\beta}\right)$?

(c) Is $\widehat{\beta}$ is a biased estimate of the parameter β due to the omission of z? Explain your answer in both words and mathematics.

Now instead if we estimate the equation

$$\widehat{y_i} = \widehat{\alpha} + \widehat{\delta} z_i,$$

(d) What is the formula OLS uses to generate $\widehat{\delta}$?

(e) What is $E\left(\widehat{\delta}\right)$?

(f) Is $\widehat{\delta}$ is a biased estimate of the parameter δ due to the omission of x? Explain your answer in both words and mathematics.

Now consider a different DGP governed by the model

$$y = \alpha + \beta x_i + \gamma w_i + u_i,$$

where W is a random variable with $cov(W, X) < 0$ and $cov(W, Y) < 0$. If we estimate the equation

$$\widehat{y_i} = \widehat{\alpha} + \widehat{\gamma} w_i,$$

(g) What is the formula OLS uses to generate $\widehat{\gamma}$?

(h) What is $E(\widehat{\gamma})$?

(i) Is $\widehat{\gamma}$ is a biased estimate of the parameter γ due to the omission of x? Explain your answer in both words and mathematics.

Let's say that a measure of X is not available and we thus have no choice but to estimate $\widehat{y_i} = \widehat{\alpha} + \widehat{\gamma} w_i$ even though we know that the proper model is $y = \alpha + \beta x_i + \gamma w_i + u_i$. Say whether the following statements are TRUE or FALSE and explain why.

(j) Our estimate of γ will be biased upward.

(k) An estimate of $\widehat{\gamma} > 0$ leaves us quite confident that $\gamma > 0$.

(l) An estimate of $\widehat{\gamma} < 0$ leaves us quite confident that $\gamma < 0$.

2. For this question, please use the *counties.dta* dataset you used for Problem Set #8. Analyze the following questions using Stata, but answer them with a few sentences in plain English. Provide appropriate Stata output as an attachment to your assignment.

(a) You are interested in the relationship between population density and crime in the state of California. Estimate the linear model $\widehat{crimerate} = \widehat{\beta}_0 + \widehat{\beta}_1 density$ using `regress`, limiting your analysis only to counties in California. Find $\widehat{\beta}_1$, and provide a one-sentence English-language interpretation of this coefficient.

(b) Construct a plot of the estimated linear model $\widehat{crimerate} = \widehat{\beta}_0 + \widehat{\beta}_1 density$ using Stata's `twoway, scatter` and `lfit` commands, again limiting your analysis only to counties in California.. For legibility purposes, you may want to remove the legend using the `legend(off)` option.

(c) A glance at the plot suggests that there is an outlier that may be affecting your results in a substantial fashion. Describe what you see and say how this might influence our estimate of $\widehat{\beta}_1$.

(d) Modify the graph you constructed in (b) in a way that allows you to identify the outlier. (Hint: to label points with a variable name, use the option `mlab(`*`varname`*`)` with the `scatter` command, where *`varname`* is the name of the variable you want to use to create the labels).

(e) Re-run your estimation, eliminating the outlying county. How did your estimates change? Did they do so in the way that you expected?

(f) Counties typically consist of many municipalities. If we were to run this analysis using California *municipalities* instead of California *counties*, $VAR\left(\widehat{\beta}_1\right)$ would definitely change in two very helpful ways for us. What are they and why would they be helpful?