New York University
Wilf Family Department of Politics
Fall 2013

Quantitative Research in Political Science I

Professor Patrick Egan

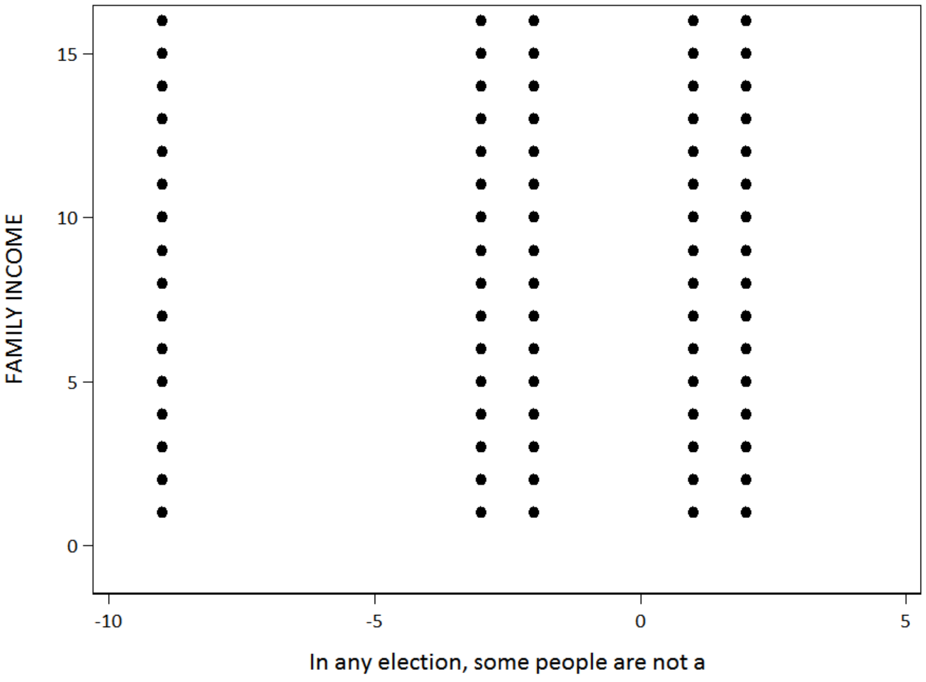
PROBLEM SET 7: Due Monday, November 18 at beginning of class.

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.

For this assigment, you are to use the `cpsnov2012.dta` dataset found on our class website. This is the November 2012 Current Population Survey, conducted by the U.S. Census Bureau with a nationally representative sample of U.S. households. The codebook for the dataset (`cpsnov2012.pdf`) may also be found on our class website.

Note: Most of these questions require that you analyze the CPS's household income variable (`hefaminc`). It is coded at the ordinal level, but the analyses require that it be at an interval level. To do this, create a recoded version of `hefaminc` in which each case is assigned a household income equal to the midpoint of its interval of `hefaminc`. For example, a household with income in the range of $5,000-$7,500 should be assigned the value $6,250, and so on.

1. The following scatterplot was produced by an analyst interested in exploring the relationship between turnout (measured with variable `pes1`) and household income (`hefaminc`). The Stata command generating the scatterplot was `scatter hefaminc pes1` .



(a) There are many, many things wrong with this figure with regard to both accuracy and style. Name as many as you can.

(b) Construct a well-designed figure that best displays the relationship between household income and turnout. This will require recoding variables and thinking carefully about the levels at which both variables are measured. Provide the Stata (or, if you prefer to use it, *R*) commands you used to recode variables and construct the figure. In a brief paragraph, explain why you made the choices that you did.

(c) In a few sentences, describe the relationship you see between income and turnout.

2. Which relationship–that between income and turnout, or education and turnout–best approximates a linear relationship? The variable to use for educational attainment is `peeduca`. Note that it, like `hefaminc`, is coded at the ordinal level but you want to analyze it as an interval-level variable.

3. You want to investigate the relationship between country of birth and current income.

   (a) You wish to divide the sample into three groups: (1) those not born in the U.S.; (2) those born in the U.S. but who have at least one parent not born in the U.S.; and (3) those born in the U.S. with both parents born in the U.S. Using the variables `hefaminc`, `penatvty`, `pemntvty`, and `pefntvty`, construct a boxplot which displays the distribution of household income for each of these three groups. *Hint*: doing this will require creating new variables from `penatvty`, `pemntvty`, and `pefntvty`. You will also need to make choices about how to deal with missing values. Justify any choices you make in a note accompanying the figure.

   (b) Using the proper statistical tests with an $\alpha = .05$, answer the following questions. Note that additional recoding may be necessary.

      i. Do native-born Americans have higher incomes than non-native born Americans?
      ii. Do native-born Americans whose parents were born in the U.S. have higher incomes than native-born Americans with at least one parent *not* born in the U.S.?
      iii. Do Americans with a foreign-born father and a native-born mother have lower incomes than Americans with a foreign-born mother and a native-born father?

   (c) In a few sentences, describe your results.

4. Take some time to acquaint yourself with the dataset. It will be helpful to refer to the codebook as you do. Now pick three variables from the dataset to illustrate omitted variable bias as we discussed in class. That is, illustrate how if the true model is

$$y = \beta_0 + \beta_1 x + \beta_2 z + u$$

but you estimate

$$y = \beta_0 + \beta_1 x + v,$$

you obtain a biased estimate of β_1.[BEFORE PROCEEDING, SEE NEXT PAGE.]

(a) First, show how omitting a variable z can bias estimates of β_1 in a *positive* direction:

   i. Pick three variables from the dataset: a dependent variable (y), an independent variable (x) and a potential confound (z).

   ii. Theoretically justify your designations of x and y as independent and dependent variables, and z as a potential confounder.

   iii. Recode the variables if necessary.

   iv. Generate a correlation matrix with your three variables. Discuss how the matrix indicates that $\widehat{\beta}_1$will be biased upward if you estimate the model $y = \beta_0 + \beta_1 x + v$ when the true model is $y = \beta_0 + \beta_1 x + \beta_2 z + u$.

   v. Estimate the models $y = \beta_0 + \beta_1 x + v$ and $y = \beta_0 + \beta_1 x + \beta_2 z + u$. Discuss how the results confirm your expectations.

(b) Now repeat this process. But this time choose x, y and z that illustrate how omitting a variable z can bias estimates of β_1 in a *negative* direction.