New York University
Wilf Family Department of Politics
Fall 2013

Quantitative Research in Political Science I

Professor Patrick Egan

FINAL EXAMINATION: WRITTEN PART

(70 POINTS TOTAL)

This exam is open-book, open-note.

1. **(15 points)** Consider the following (admittedly simple) example. A DGP defined by the population model $y = \beta_0 + \beta_1 x + u$ gives rise to the following dataset of 4 observations:

$$\mathbf{y} = \begin{bmatrix} 4 \\ 2 \\ -6 \\ 1 \end{bmatrix} \qquad \text{and} \qquad \mathbf{X} = \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & -1 \\ 1 & 4 \end{bmatrix},$$

where the second column of $\mathbf{X}$ is composed of observations of the variable x. In answering the following questions, be sure to show all your work.

(a) Show that the OLS estimates $\widehat{\beta}_0 \approx -2.89$ and $\widehat{\beta}_1 \approx 1.57$. You will be glad to know that

$$(\mathbf{X}'\mathbf{X})^{-1} \approx \begin{bmatrix} .536 & -.143 \\ -.143 & .071 \end{bmatrix}.$$

(b) Show that $\widehat{\sigma} \equiv SEE \approx 3.33$.

(c) Show that $R^2 \approx .61$.

2. **(15 points)** Consider four random variables W, X, Y and Z, where

$$cov(W, Y) > 0; \quad cov(W, X) = 0;$$
$$cov(Z, Y) = 0; \quad cov(Z, X) < 0,$$
$$\text{and } cov(X, Y) \text{ is unknown.}$$

Say whether the following statements are TRUE or FALSE, and explain why. Assume we have a large number of observations of the joint distribution of all four variables from an i.i.d. random sample.

(a) If we estimate the equation

$$\widehat{y_i} = \widehat{\beta}_0 + \widehat{\beta}_1 x_i,$$

$\widehat{\beta}_1$ is a *biased* estimate of the parameter β_1 due to the omission of w and z.

(b) The estimate of the parameter β_1 we obtain from the estimated equation

$$\widehat{y_i} = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$$

will be *more efficient* than the estimate of the parameter β_1 obtained from the equation

$$\widehat{y_i} = \widehat{\beta}_0 + \widehat{\beta}_1 x_i + \widehat{\beta}_2 z_i.$$

(c) The estimate of the parameter β_1 we obtain from the estimated equation

$$\widehat{y_i} = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$$

will be *more efficient* than the estimate of β_1 obtained from the equation

$$\widehat{y_i} = \widehat{\beta}_0 + \widehat{\beta}_1 x_i + \widehat{\beta}_2 w_i.$$

3. **(15 points)** Consider three variables X, Y and Z, where in the population

- X takes on the value zero 50 percent of the time and the value one 50 percent of the time, while
- Z takes on the value zero 3 percent of the time and the value one 97 percent of the time.

You are interested in estimating the *ceteris paribus* association of X with Y as well as the *ceteris paribus* association of Z with Y. To do so, you use the model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + u_i.$$

Assume that this model is properly specified, and the Gauss-Markov assumptions hold.

(a) One or more of the following four statements is true. In a few sentences, identify the correct statement(s) and explain:

$$var(\beta_1) = \frac{\sigma^2}{n \cdot var(x) \cdot (1 - R_x^2)} \qquad \widehat{var\left(\widehat{\beta}_1\right)} = \frac{\widehat{\sigma}^2}{n \cdot var(x) \cdot (1 - R_x^2)}$$

$$var\left(\widehat{\beta}_1\right) = \frac{\sigma^2}{n \cdot var(x) \cdot (1 - R_x^2)} \qquad \widehat{var\left(\widehat{\beta}_1\right)} = \frac{\sigma^2}{n \cdot var(x) \cdot (1 - R_x^2)}$$

(b) It is the case that $R_x^2 = R_z^2$. Why can we say for sure that

$$var\left(\widehat{\beta}_1\right) < var\left(\widehat{\beta}_2\right) \quad ?$$

(c) All things being equal, with which of the two findings should you be more comfortable? Why?

- A failure to reject the null that the *ceteris paribus* association between Z and Y is zero.
- A failure to reject the null that the *ceteris paribus* association between X and Y is zero.

4. **(25 points)** Consider the Stata output on the following page. It is an OLS analysis of "feeling thermometer" ratings given to Barack Obama (on a zero to 100 scale) in the 2012 American National Election Studies by a nationally representative sample of American adults. Be sure to explain your answers and show your work.

   (a) What proportion of the respondents in the sample own guns?

   (b) What is the rating predicted to be given to Obama by a (non-Hispanic) African American man born in the U.S. whose education and age are equal to the American average, whose household income is $60,000, and who is a military veteran and a union member but who is not a gun owner?

   (c) How many standard deviations away from y is the typical prediction $\widehat{y}$?

   (d) The constant term in the regression ≈ 72. Describe the hypothetical American whose predicted rating of Obama is indicated by this term (however nonsensical the prediction may be).

   (e) What is the approximate predicted difference in ratings given to Obama between someone with a household income of $30,000 and someone with an income of $45,000, holding all other covariates constant?

   (f) What is $\frac{\partial ObamaFT}{\partial AGE}$? Your response should include both a mathematical expression and a few sentences of explanation.

   (g) What is $\frac{\partial ObamaFT}{\partial EDUC}$? Your response should include both a mathematical expression and a few sentences of explanation.