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. You will need a calculator.
Use of a computer on this part of the exam is not needed but is permitted.

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

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

(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:

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

(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 $\approx$ 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.

Variable	Variable label
obamaFT	Obama feeling thermometer (0 = cold, 100 = warm)
ln_inc	Household income (logged $)
educ	Educational attainment (1-5 scale)
black	Black/African American (no = 0, yes = 1)
hispanic	Hispanic/Latino (no = 0, yes = 1)
female	Female (no = 0, yes = 1)
age	Age (in years)
gunowner	Owns gun (no = 0, yes = 1)
veteran	Military veteran (no = 0, yes = 1)
union	Member of labor union (no = 0, yes = 1)
foreign_born	Born outside of U.S. (no = 0, yes = 1)

```
. sum obamaFT ln_inc educ black hispanic female age gunowner veteran
union foreign_born

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
     obamaFT |       5188    57.98207    34.49865          0        100
      ln_inc |       5188    10.49084    1.206994   7.824046   12.61154
        educ |       5188    3.010023    1.156159          1          5
       black |       5188    .1808019     .384891          0          1
    hispanic |       5188    .1661527    .3722535          0          1
-------------+---------------------------------------------------------
      female |       5188    .5144564    .4998391          0          1
         age |       5188     49.8325    16.52954         19         80
    gunowner |       5188    .3257517    .4687002          0          1
     veteran |       5188    .1337702    .3404381          0          1
       union |       5188    .1002313    .3003371          0          1
-------------+---------------------------------------------------------
foreign_born |       5188     .093485    .2911389          0          1
```

```
. reg obamaFT female c.age##i.black foreign_born hispanic veteran gunowner union
ln_inc c.educ##c.educ, robust

Linear regression                                      Number of obs =    5188
                                                       F( 12,  5175) =  252.98
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.2567
                                                       Root MSE      =  29.777

------------------------------------------------------------------------------
             |               Robust
     obamaFT |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   3.023606   .8835436     3.42   0.001     1.291488    4.755725
         age |  -.1058921   .0308458    -3.43   0.001     -.166363   -.0454212
     1.black |   24.15863   2.418359     9.99   0.000     19.41762    28.89963
             |
 black#c.age |
          1  |   .2807063   .0463736     6.05   0.000     .1897944    .3716182
             |
foreign_born |   2.621286   1.478578     1.77   0.076    -.2773525    5.519924
    hispanic |   14.24586   1.283816    11.10   0.000     11.72903    16.76268
     veteran |  -3.144224   1.400031    -2.25   0.025    -5.888876    -.399571
    gunowner |  -11.38375   .9791201   -11.63   0.000    -13.30324    -9.46426
       union |   4.904637   1.435476     3.42   0.001     2.090497    7.718776
      ln_inc |  -.5553546   .3858159    -1.44   0.150    -1.311717    .2010075
        educ |  -8.292665   1.749165    -4.74   0.000    -11.72177   -4.863563
             |
c.educ#c.educ |   1.44211   .2818197     5.12   0.000     .8896241    1.994595
             |
       _cons |   71.74481   4.770451    15.04   0.000     62.39271    81.09691
------------------------------------------------------------------------------
```

New York University
Wilf Family Department of Politics
Fall 2013

Quantitative Research in Political Science I

Professor Patrick Egan

FINAL EXAMINATION: STATISTICAL COMPUTING PART

(70 POINTS TOTAL)

This exam is open-book, open-note.
Use a computer on this part of the exam.

- To complete this part of the exam, download the dataset annenberg2004.dta from our class Blackboard site (from the "Course Documents" section). If you have trouble doing this, see Andrew or call me (646-808-7271).

- Your responses to the questions on this part of the exam should be included in a .log file that you will print out and submit for your exam grade. The file must include the commands you use in the analysis, any output, and answers to questions (entered in the .log file as comments).

- You must answer the substantive questions on this part of the exam in words. **You will not earn full credit for Stata output submitted without written interpretation.**

The dataset `annenberg2004.dta` is a sample of respondents from the 2004 National Annenberg Election Survey (NAES), conducted during the 2004 U.S. Presidential election campaign.

NOTE: Unless specified otherwise, use an alpha of .05 when performing statistical significance tests.

1. **(5 points)** Begin your `.log` file with an identification header, as suggested by Jonathan Nagler's "Coding Style and Good Computing Practices" and used by James in your lab sessions.

2. **(2 points)** Run the command that lists the names of variables, their formats and labels.

3. **(18 points)** *bush_favorability* is a measure of respondents' favorability ratings of George W. Bush on a scale of zero (least favorable) to ten (most favorable). (NOTE: For reasons that should be obvious, **you will need to create a recoded version of this variable**.)

   (a) What is the mode of *bush_favorability*?

   (b) What is its mean?

   (c) What is the 95% confidence interval about this mean?

   (d) What is the 90% confidence interval about this mean?

   (e) Who rates Bush more favorably, men or women?

      i. Answer this question with a *t*-test. Interpret your results.

      ii. Answer this question with a bivariate regression. Interpret your results.

   (f) Are higher income Americans more likely to assess Bush favorably than lower income Americans?

      i. Answer this question under the assumption that

$$bush\_favorability = \beta_0 + \beta_1 income + u,$$

      and that the other Gauss-Markov assumptions are met. (Have a close look at *income*; **some substantive recoding is needed here**.) Interpret your results.

      ii. Explore this question in a way that does not require this assumption.

4. **(20 points)** *kerry_favorability* is a measure of respondents' ratings of John Kerry on the same scale as *bush_favorability*. Create a new variable (as done in class) called *rating_diff* that is equal to *bush_favorability* minus *kerry_favorability.* (**You'll again need to do some recoding.**)

   (a) Run a regression of *rating_diff* on the variables *age*, *female*, and *income*. We'll call this Model I.

      i. Before estimating Model I, use diagnostics (as we did in class) to examine carefully how *income* should enter into the model as a predictor. Justify and document the decision you make regarding this. **Recoding of *income* is necessary here**.

   (b) How well does Model I explain variation in the dependent variable? Cite two estimated statistics in your response.

(c) Now run a regression of *rating_diff* on the same variables as well as *ideology*–a variable in which respondents rate themselves very conservative (1) to very liberal (5) on a five-point scale. (For the moment, treat *ideology* as an interval-level variable. Note that **some recoding may be necessary**.) We'll call this Model II.

   i. In a few sentences, explain what happens to the coefficient on *female* between Model I and Model II and your substantive interpretion of this change.

   ii. How well does Model II explain variation in the dependent variable compared to Model I?

   iii. What does Model II predict is the difference in *rating_diff* between those who are very conservative and those who are very liberal, holding the other factors constant? Approximately what percentage of the range of the dependent variable is this difference equal to?

   iv. What do your answers to (ii) and (iii) suggest about the importance of *ideology* in explaining the dependent variable compared to the other variables in Model II?

(d) Finally, run the estimation in a way that does not require that we assume *ideology* is an interval-level variable. Call this equation Model III.

   i. Now think carefully: what would we need to see in Model III that would give us confidence in treating *ideology* as an interval-level variable in the present context? Do we see this here?

   ii. In a few sentences, say what Model III tells us about the *ceteris paribus* relationship between *ideology* and *rating_diff*.

5. **(25 points)** Explore the following question using OLS.

   - We expect that voters who rated their personal economic situation as poor in 2004 would be likely to blame the incumbent (Bush), and thus have lower values of *rating_diff* than those in better economic circumstances. However, we might theorize that the *ceteris paribus* association between a voter's personal economic situation and *rating_diff* is even stronger for Independent voters (who are not affiliated with either the Democratic or Republican parties), as these Independents do not have the cue of party identification to rely upon when rating the two candidates.
   - To explore this question, use the variables *pid* (a categorical variable that is a measure of voters' party identification) and *economic_situation* (a variable measuring how voters rate their personal economic situation on a scale of 1 (excellent) to 4 (poor)). For purposes of this question, you may treat this variable as interval-level.
   - From here, you're on your own. You will need to create new variables, estimate the appropriate model, present results both in tabular and graphical form, and interpret the results. Do your estimates confirm the theory? Make sure to describe your results in plain English.

6. Close your `.log` file, print it out, and hand it in with the written part of the exam.