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. .

- To complete this part of the exam, download the dataset `annenberg2004.dta` from our class website (from the "Resources" 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.

NOTES:

- In many cases, proper analysis will require recoded versions of variables in the dataset.
- Unless specified otherwise, use an alpha of .05 when performing statistical significance tests.

1. **(2 points)** Run the command that lists the names of variables, their formats and labels.

2. **(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).

   (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? Answer this question under the assumption that

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

   and that the other Gauss-Markov assumptions are met.

      i. Before doing so, 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. **Regardless of your decision, note that some recoding of *income* is necessary here**.

      ii. Interpret your results.

3. **(25 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.*

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

   (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.) 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*.

   (e) Using the proper Stata commands, create a table displaying the estimates from Models I, II and III. The table should display regression coefficients, their standard errors, and markers of statistical significance in three columns, along with four statistics from each of these regressions (N, root MSE, R^2 and adjusted R^2). Variable names should be displayed on the rows .

4. **(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.

5. Close your `.log` file email it to Andrew by 3 p.m.. Then print it out a hard copy and put it in Andrew's mailbox.