Final Exam — Political Science 599 — Fall 2005

Due 5pm, Monday, December 19 to Bryce's office, 7th floor Haven.

Instructions: Please type your answers to this exam. When presenting tables, do not present raw, "cut-and-pasted" statistical-program output directly but use your word processor or other software to present professional-looking tables. Likewise, when presenting graphs, make sure that graph titles, axes-labels, lines, and important other features of your graph look professional. (Define "professional-looking" as do the top journals in your field.) Embed both graphs and tables directly within your document, in appropriate sizes at appropriate places in the text. If you cannot You may hand-write (neatly) the solutions to the math problems, showing appropriate steps in your work. In responding to the questions, use your own words. If you do quote from books, explain those quotes in your own words. This exam is open-book, open-note, open-web, etc., but do not work together and do not seek or accept interactive human assistance (e.g., in person, phone, e-chat, or e-mail).

Dataset: The dataset for this exam, which is posted on my webpage along with codebooks, includes data from a team of researchers at UC-Berkeley on 2004 Ohio and Florida presidential-election-results, plus data from the 2000 Census on Income and Race for all of the counties in Florida and Ohio, plus 400 further variables, capturing religion and religious participation for these counties as of 1990.

1. **The '04 Bush vote in OH & FL:** The dependent variable in this question is the percent of country votes counted for Bush as of mid-November 2004. The goal is to build and evaluate a small empirical model that explains the percentage Bush vote by county. [total 72 points]

   (a) Describe the dependent variable using appropriate descriptive statistics and graphics. As you do this, relate what you learn about this variable in these two states with what you know about the election in general. [4 points]

   (b) Is the percentage Bush vote in a county independent of the state of that county? Use at least two of the several strategies we discussed to answer such a question. [4 points]

   (c) Pick an explanatory variable that you think might help explain why Bush enjoyed more support in some counties than in others.[1] The variable doesn't have to be the one that you think will prove most powerful, just one of interest to you that you think will prove noticeably relevant empirically. Note that your explanatory *variable* could sum or average several columns in the dataset — e.g., you may want to know about evangelical Christians as a whole rather than particular sects within that group. (You can recode & combine for the other variables you choose in Question 1e, too.) [4 points]

      i. Explain your choice: how & why should this variable matter, and how & why is it interesting (to you)? [2 of 4]

      ii. Describe this variable as you did the dependent variable in Question 1a. [2 of 4]

   (d) Before specifying multivariate OLS or ANOVA models, simple bivariate explorations are often useful [5 points]:

      i. Report appropriate summary statistics for the bivariate relationship of your variable with the Bush vote. [1 of 5]

      ii. Present some useful tabular &/or graphical information about this relationship. [2 of 5]

      iii. Describe what this preliminary investigation suggests for the form of the structural relationship between your explanatory variable and county-level Bush vote. [2 of 5]

[1] Do not choose touch-screen voting as we considered a model with that independent variable already in class.

(e) Choose 1 or 2 *other* variables that seem to you, or might seem to a critic, must also be included in the model, and... [6 points]

i. Describe these "control" variables as well (needn't be as thorough as above; just most notable features of those data) [1 of 6]

ii. Explain the statistical & substantive reasons to include these variables. [2 of 6]

iii. Write your multivariate model in scalar & in matrix notation, & justify, on substantive, statistical, &/or empirical (e.g., maybe you discovered something in Question 1d) grounds, the functional form of the structural relationship your model assumes. [3 of 6]

(f) List formally the assumptions of the estimator you intend to use to estimate your multivariate model [1 of 6], explain those assumptions generically in your own words [2 of 6], and give a substantive interpretation or example or illustration in the context of your model of each assumption [3 of 6]. [6 points]

(g) Derive hypotheses from your theoretical/substantive expectations about the coefficients that you will estimate in this model. Translate these hypotheses back into substantive statements about why Bush would have won more votes in some counties than in others. [4 points]

(h) Estimate, in your favorite statistics package, the model you suggested above. [8 points]

i. Paste the raw estimation output from your software into your exam answers at this point. [1 of 8]

ii. Produce a professional-looking table of coefficients, plus standard errors, t-stats, & p-values, plus a "standard set" (see a top journal in your field) of summary statistics for the estimation (e.g., R^2, regression F-test statistic, etc.). [2 of 8]

iii. For each of five items from your table just mentioned—coefficients, standard errors, (t-stats or p-values), plus two summary statistics (e.g., R^2 or other measure of fit and F-stat)—give the mathematical formula for its calculation from the data (scalar or matrix notation, as you prefer) and state formally and in your own words what it means/what information it conveys. [5 of 8]

(i) Interpret your estimation results statistically and substantively. [15 points]

i. State what each coefficient estimate means in terms of the effect of X's on Y, substantively, and give some clear and accurate statement as to the certainty, precision, statistical-significance of those estimates. [5 of 15]

ii. Present some predicted values or estimated effects for this regression graphically, with revealing substantive labels, to help make your interpretation more vivid. [5 of 15]

iii. Explain for your reader which assumptions you feel more confident about in this context, which less confident, and why. Show diagnostic evidence (appropriate graphics and/or statistics) to support your claims. [5 of 15]

(j) Given these empirical results (Questions 1h and 1i), what do you think about the theoretical/substantive propositions you offered in Question 1g? What have you learned from your empirical analysis? How well does the data do with your empirical model what your theory says it ought? Do you want to revise any of your propositions now? If so, how? [4 points]

(k) Might the effect of some or all of your explanatory variables would vary across states in this country? If so, how & why? If not, why not—on what basis do you contend the effect of those X's should be the same in all states? [3 of 12] [12 points]

i. Pretend a critic emerges who believes the opposite of what you just argued in the last question. Specify a (one) model that will adequately reflect your argument but also allow you to test the critic's hypothesis. [4 of 12]

ii. Estimate this model using the data from Florida and Ohio. Interpret the results statistically and substantive in terms of your debate with your critic. [3 of 12]

iii. For this exam, we only have Florida & Ohio data; if you had data from all 50 states, would you want to estimate a model in the spirit of the one you just estimated here in 2 states or would you suggest some other strategy? Either way, write with formally correct notation the model that you would want to estimate if you had all 50 states. Be sure to define all notation (any new subscripts & variable names, e.g.). [2 of 12]

2. Now, consider electronic touch-screen voting as the dependent variable. [total 28 points]

(a) Offer a theoretically/substantively motivated multivariate model (with, let's say, 3 independent variables, chosen from among the variables in our dataset) to explain which counties (across both Ohio and Florida) went touch-screen in '04 and which remained with some other technology. That is, argue theoretically and substantively what (from the options within our dataset) should determine these choices by Ohio & Florida counties. [6 points]

(b) First, consider estimating this model of which counties choose touch-screen voting by Ordinary Linear-Regression Least Squares (OLS). [7 points]

i. What are the assumptions of C(N)LRM that render OLS estimates BLUE? What do you think of each of those assumptions as applied to this substantive context? [3 of 7]

ii. Estimate the OLS model, report the estimation results in a professional-looking table, & interpret those estimates statistically (discuss certainties & significance-levels of estimates) & substantively (discuss the effects of X's on county's voting-technology choices). [4 of 7]

(c) Now, estimate a model with the same set of independent variables by logit or probit Maximum-Likelihood estimation. [7 points]

i. What are the assumptions that render ML estimates BANC? What do you think of each of those assumptions as applied to this substantive context? [3 of 7]

ii. Estimate the MLE model, report the estimation results in a professional-looking table, & interpret those estimates statistically (discuss certainties & significance-levels of estimates) & substantively (discuss the effects of X's on county's voting-technology choices). [4 of 7]

(d) Comparing MLE & OLS: [8 points]

i. Discuss the *substantive* difference(s)—i.e., in terms of estimated *effects*—in your OLS and logit/probit ML estimates of the voting-technology choice. [4 of 8]

ii. More generally, compare the assumptions of MLE for the linear model with those of OLS. Explain how do they differ in your own words. [2 of 8]

iii. Compare the set of properties BANC with those of BLUE. What are the relative advantages and disadvantages of MLE compared with OLS. [2 of 8]