Final Exam

Due Tuesday, December 17, 2024

Insert Name Here

Instructions

Answer the questions within this document. Keep the questions included below. Make sure to answer the questions in sufficient detail.

- Report all regression results in table form using **texreg**.
- Submit your rendered .pdf in Blackboard.

Number 1: Unusual and Influential Data; Normality of Errors

For this problem, use the "influence" data.

A. Estimate the following linear regression model and report your results:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$$

B. Use studentized residuals to assess which observations are outliers. Produce a graphical plot capable of identifying outliers, and identify outlying observations. Include your graph.

C. Use DFBETAS to assess which observations are highly influential. Produce graphical plots capable of identifying influential observations. Include your graphs. Report which observations are influential. [Note: On your own, feel free to use Cook's D as well.]

D. Though you have no theoretical background on this data, discuss one robustness check that might be appropriate for dealing with influential observations in this data.

E. Use the Jarque-Bera test to evaluate the normality of the residuals. What do the results suggest about additional data transformations or model specifications?

Number 2: Binary Dependent Variables

Let's say you're interested in studying the individual-level predictors of *opposition to torture* for terrorist suspects. You have a random sample of American adults from the NES data. Open the "torture" data. Your dependent variable is:

- **tor2**: opposition to torture; 1=opposed to using torture, 0=not opposed (either in favor or moderately in favor)

You include the following independent variables:

- **pid**: Party identification; 7-point scale ranging from 0=Strong Republican to 6=Strong Democrat
- **Female**: 1=female, 0=male
- **educ4**: 1=HS or less, 2=some college or Assoc. degree, 3=B.A., 4=Grad or more

- **Race3**: 1=white, 2=black, 3=Hispanic; reminder: this is nominal variable, so you need to "dummy out" the categories and include two out of the three.
- **income5**: Income; 1=<=$22.5k, 2=$22.5-$45k, 3=$45-$75k, 4=$75-$125k, 5=>$125k
- **Age**: Age in years, ranging from 18-90

A. Estimate the model using both logit and the linear probability model (LPM, with robust standard errors). Report your results. Report goodness-of-fit statistics for each model. Interpret the results in terms of sign and significance. Also interpret goodness-of-fit in each model.

B. For each model, show the effects of *party id*, *education*, and *race* in terms of predicted probabilities. For logit, use the "observed value" approach. Report your results in graphical form. Interpret those results.

C. In general, how does the LPM compare to the logit model?

D. A critic contends that education surely attenuates partisan divisions in opposition to torture. Estimate a logit model to test whether education attenuates the effect of party identification. Report your model results, and also report a post-estimation analysis in graphical form. Use the observed value approach for predicted probabilities. Is the critic right or wrong?

Number 3: Probability and Distributions Review

A state's public university system admits students only if they score above 400 on a standardized achievement test. Applicants from group A have a mean of 500 and a standard deviation of 100 on this test, and applicants from group B have a mean of 450 and a standard deviation of 100. Both distributions are approximately normal, and both groups have the same size.

A. Find the proportion *not admitted* for each group.

B. Of the students who are not admitted, what proportion are from group B?

C. A state legislator proposes lowering the cutoff to 300, thinking that the proportion of students admitted who are from group B will decrease. What's the answer to part B under this policy? Would the proportion of students admitted from group B decrease?

Number 4: Comp Exam Style Question I

Answer **one** of the following two questions.

A. What is the value of "data visualization" in political science research? Discuss the role and benefits of data visualization in both data exploration ("pre-estimation") and post-estimation interpretation of regression model results.

B. We usually think of "data visualization" in terms of summarizing our data frame with the use of scatterplots, boxplots, and histograms. In other words, we're looking for ways to graphically display both *central tendency* and *dispersion* of our sample observations. Discuss how the use of *geographic maps* in data visualization (particularly the use of choropleth maps) can enhance our communication of central tendency and dispersion and uncover new patterns in our data? How would you plan on using geographic maps in your own research?

Number 5: Comp Exam Style Question II

Answer **one** of the following two questions.

A. Let's say you estimate a model with a nonlinear functional form for the effect of one of your variables (e.g., quadratic, diminishing marginal returns). A critic cries foul, claiming that "this model makes no sense at all since *linear* regression assumes *linearity*." How would you respond to the critic's claim?

B. A critic suggests that models with interaction terms ruin the elegance of the linear regression model. According to the critic, "The whole point of regression is *generalization*, which entails estimates of *average marginal effects*. What you get from interactions is *particularization*, which entails a bunch of conditional marginal effects that don't make any sense." Critically review this critic's arguments about interactions.