Public Policy 571
Econometrics
Problem Set #2

Due on Wednesday, September 25

1. For this question, use the dataset `anes2016subet`. The dependent variable, `HomeOwnership`, consists of four categories: pay rent, pay mortgage, own home with no payments, and some other arrangement. The independent variables are: `Age` in years; `BAplus`, which is a dummy variable in which 1 indicates the person has a degree from a four-year college; and `Ideology`, which is the person's self-reported ideology on a seven-point scale in which higher values indicate more conservative beliefs.

   (a) Estimate a multinomial logit model with "pay mortgage" as the base category. Report the results and describe the basic substantive findings without calculating predicted or marginal effects. R users should convert the `Ideology` variable to a numeric variable starting at 0 and may want to do the same with `BAplus`.

   (b) Show how the χ^2 statistic from the Likelihood Ratio test is calculated. What is the substantive interpretation of this test?

   (c) Find the following by hand using the formulas. Suppose a person is 55, has a college degree, and is a moderate (3) on the ideology scale. With what predicted probabilities is the person in each home ownership category?

   (d) Using your software, find the predicted probabilities for each category for a person with median values of all the independent variables (medians from the estimation sample, which is what the built-in functions use, not the full dataset medians).

   (e) Now suppose that `Age` is one standard deviation higher, while all other variables remain at their medians. What is the *predicted change* in probabilities for each category?

   (f) Using your software, what is the average marginal effect of the variable `BAplus`?

   (g) Change the base category to "own home with no payments" and re-run the model. Examine the coefficients and compare them to the earlier estimation. How do you interpret the differences?

2. This question will use the dataset `anes2016subset`. The dependent variable is `Memberships`.

   (a) Produce a histogram of the dependent variable. Explain why OLS might be not be the best estimation method for these data.

(b) Estimate a Poisson model using the following independent variables: `Age`, `BAplus`, `Ideology`, and `NewsDays`. Report the results and describe the basic substantive findings without calculating predicted or marginal effects. Note: R users should make `Ideology` a numeric variable starting at 0.

(c) Using Stata or R, calculate the predicted number of memberships for a person who is 35 years-old, has a college degree, is liberal (1) on the ideology scale, and who watches/reads the news 5 days a week.

(d) By hand, calculate the probability that this person belongs to 2 groups.

(e) Run the same model using a negative binomial regression. Using Stata or R, calculate the predicted number of memberships for a person with the same characteristics as those described in part (c). Compare this result to the previous prediction.

(f) Now estimate the same model using a zero-inflated Poisson regression. In this model, use the variables `Voted2016` and `Health` as variables that predict whether a person is in the "always zero" category. Interpret the substantive meaning and statistical significance of the coefficients on these two variables. Note: R users should treat `Health` as a numeric variable starting at 0.

(g) Using Stata or R, calculate the predicted number of memberships for a person with the same characteristics as those described in part (c). Assume also that the person voted in 2016 and is in very good health (3).

(h) Now estimate a Hurdle model in which the first stage is a logit and the second stage is a truncated Poisson. In Stata, this will require `suest`.

(i) Calculate the predicted number of memberships for a person with the same characteristics as described in part (g).

(j) When trying to decide which of these models is most appropriate/best for this scenario, what factors do you consider?