Public Policy 779
Applied Econometrics
Problem Set #1

Due on Monday, September 16

Please use the following procedures when turning in problem sets for this class:

- First, please turn in problem sets via file upload to Canvas. A link will be provided in the Assignments section.
- Second, please make a document that contains your written answers to each question and includes the output from Stata or R that is of direct relevance to the question. For example, if you will be interpreting coefficients from a probit, include the probit output in your written answers. If you are asked to make a figure, then include the figure in your answer document. Please note that output from these programs looks better if you use a monospaced font like Courier. If you are familiar with RMarkdown, by all means use that. I will provide some information and a tutorial for using Markdown in both Stata and R, but do not feel the need to learn it for Problem Set 1.
- Third, produce a well-labeled Stata do file or R script file as an appendix to your written answers (or upload it as a separate file). By labeling, I mean that you can use comments and/or visual separators in your code file to clearly demarcate different questions from each other. If you are using Markdown, there is no need for a separate do file/script. Just include all your commands in the output (you can suppress the output from commands if desired by setting `echo = FALSE` in that code block).

1. The purpose of this question is to help you build familiarity with likelihood functions. You may want to use a spreadsheet to help you perform some of the calculations, as it allows you to write an equation in which the unknown is a reference to another cell. Then you can easily change the value of that cell to get different solutions to the equation.

   Let's suppose that we are using a logit to model the probability that a randomly selected voter supports Kamala Harris.

   $$Pr(H) = \frac{\exp(\theta)}{1 + \exp(\theta)}$$

   (a) This function produces a value that falls between 0 and 1. Now let the parameter θ take on the following values: -3, -1, 0, 1, 3. Calculate $Pr(H)$ for each value of θ. What do you notice?

Note: on a calculator, finding exp(θ) involves typing a number for θ and then hitting the e^x button. On a spreadsheet, we could type `= exp(B1)/(1 + exp(B1))` into a cell, where B1 is a cell reference. Then, typing different numbers into cell B1 gives us different solutions for the equation.

(b) Now suppose that you randomly select three voters and observe: H, H, and T, where T means voting for someone else. The joint probability of this set outcomes is:

$$Pr(H,H,T) = \left(\frac{\exp(\theta)}{1+\exp(\theta)}\right)\left(\frac{\exp(\theta)}{1+\exp(\theta)}\right)\left(1-\frac{\exp(\theta)}{1+\exp(\theta)}\right)$$

Note that the third term on the right hand side can be simplified if we rewrite the 1 as $\frac{1+\exp(\theta)}{1+\exp(\theta)}$ and perform the subtraction.

$$Pr(H,H,T) = \left(\frac{\exp(\theta)}{1+\exp(\theta)}\right)\left(\frac{\exp(\theta)}{1+\exp(\theta)}\right)\left(\frac{1}{1+\exp(\theta)}\right)$$

Again, let the parameter θ take on the values of -3, -1, 0, 1, 3. This is where using a spreadsheet will really save some time. What are the associated joint probabilities?

(c) Since a likelihood function is proportional to the probability of observing y given θ, we can write the above as a likelihood function.

$$\mathcal{L}(\theta|y=H,H,T) = \left(\frac{\exp(\theta)}{1+\exp(\theta)}\right)\left(\frac{\exp(\theta)}{1+\exp(\theta)}\right)\left(\frac{1}{1+\exp(\theta)}\right)$$

Try to find the value of θ that maximizes the expression. What is the value of $Pr(H)$ associated with this value of θ? Does this make sense?

(d) If we take the log of both sides of the above, we get the following:

$$\ln\mathcal{L}(\theta|y=H,H,T) = \ln\left(\frac{\exp(\theta)}{1+\exp(\theta)}\right)+\ln\left(\frac{\exp(\theta)}{1+\exp(\theta)}\right)+\ln\left(\frac{1}{1+\exp(\theta)}\right)$$

Find the value of θ that maximizes this new equation. Hint: do the calculations inside the parentheses and then take the natural log. What insights do you draw from your answer?

(e) All of this assumes that $Pr(H)$ is the same for all individuals. It's more likely that different people with different characteristics will have different probabilities of voting H.

Let's assume we measure some characteristic x_i for each person. We now can model the parameter θ_i as βx_i. Re-write the log likelihood function with this change. Note: normally, we would include an intercept as well, so that $\theta_i = \beta_0 + \beta_1 x_i$, but let's keep it simpler for now and ignore the intercept.

(f) Suppose that the value of x_i in each case respectively (in order) is 6, 3, 4. Find the value of β that produces the highest value of the log likelihood function.

(g) Given this value of β, what is the probability that the first voter, with $x_i = 6$, will vote H? The second voter?

2. Use the GSS2014subset dataset for this question. The dependent variable is abany, where 1 indicates the respondent believes a pregnant woman should be able to obtain a legal abortion if she wants, for any reason, and 0 indicates otherwise.

   The independent variables are: age, the respondent's age in years; childs, the number of children a respondent has; educ, the respondent's education in years; polviews, the respondent's score on the 7-point ideology scale in which higher values mean a person is more conservative; and relpersn, a four-point scale in which higher values mean the person is less religious.

   Note: R users likely will want to convert the variables that are in the factor class to numeric class using as_numeric from the sjlabelled library so that they start at 0 post-conversion (see help document). The regular as.numeric function will create variables that start at 1.

   (a) Estimate a probit model using the above-mentioned variables. Report the results and describe the basic substantive findings – the effect of each independent variable and its statistical significance – as best you can without calculating any predicted values or marginal effects.

   (b) By hand, calculate the probability that a person supports an unrestricted right to legal abortion if that person is 35 years old, has two children, has 16 years of education, is a 3 on the ideology scale (starting at 0), and is a 2 on relpersn.

   (c) Now suppose instead the person is 65 years old and is a 5 on the ideology scale. What is the change in the predicted probability of a support for unrestricted abortion rights?

   (d) Using your software, find the predicted probability that a person supports unrestricted abortion rights when all independent variables are set to their medians in the estimation sample. In Stata, this means using the predict command with the (medians) _all option. In R, use the predictions function from the marginaleffects package, which can also be set to use medians from the estimation sample.

   (e) Using your software, set age to 40, childs to 4, educ to 12, relpersn to 3, and let polviews vary from 0 to 6 in increments of 1. Find the predicted probability the person supports unrestricted abortion rights as polviews changes. Then make the marginsplot.

   (f) What is the average marginal effect of the variable relpersn?

   (g) Re-estimate the model as a logit model. Using your software, find the predicted probability that a person supports an unrestricted right to legal abortion if that person is 35 years old, has two children, has 16 years of education, is a 3 on the ideology scale, and is a 2 on relpersn. Compare this to your answer in part (b).

   (h) Now run the model as a logistic regression that produces odds ratios for coefficients. Compare the coefficients to the logit model that you just ran, explaining how they have the same substantive meaning.

3. For this question, use the dataset `anes2020subset`. The dependent variable is `DiversityGood`, which provides an ordinal set of responses to the following question: "Does the increasing number of people of many different races and ethnic groups in the United States make this country a better place to live, a worse place to live, or does it make no difference?" The responses are recorded as "worse," "makes no difference," or "better."

   The independent variables of interest are: `KnowsImmigrant`, a dichotomous variable that indicates whether the respondent knows someone who is an immigrant to the US (1=yes); `LGBTfriends`, a dichotomous variable that indicates whether the respondent has LGBT friends (1=yes); `ICEtherm`, a feeling thermometer score (0-100) for the U.S. Immigration and Customs Enforcement (ICE) agency; `Ideology`, the respondent's placement on a seven-point (0-6) scale in which higher values mean more conservative political beliefs; `WhitesTherm`, the respondent's feeling thermometer score (0-100) for white people.

   (a) Estimate an ordered probit model. Report the results and describe the basic substantive findings, including statistical significance, without calculating predicted or marginal effects. Note: R users should first convert `Ideology` to a numeric variable (see help document). Also, it may also be easier for later – in parts (e) and (f) – to deal with the dichotomous variables if they are converted to numeric.

   (b) By hand, calculate the value of $x_i\beta$ for a person who does not know an immigrant, who does not have LGBT friends, whose feeling thermometer score for ICE is 55, who is a 4 ("slightly conservative") on the ideology scale, and whose feeling thermometer score for white people is 75.

   (c) Put the value you calculated for $x_i\beta$ at the center of a standard normal (z) distribution (i.e. it has a z-score of 0; the standard deviation is already 1). Then, what would be the corresponding z scores for the 2 cut points? What proportion of the area is below cut1? What proportion of area is above cut2? What proportion is between cut1 and cut2?

   (d) What if, instead, the person does know an immigrant? How do the predictions change? Note: keep all other variables at the same values.

   (e) Now use Stata or R. If a person has the median value (in the estimation sample) for all independent variables, with what predicted probabilities does that person give each respective response? For R users: if the dichotomous variables are left as factors, you will need to follow code examples from the help document for specifying factor variable values.

   (f) Again, use your software. For a person with the median values for all independent variables in the estimation sample, what is the predicted effect of a 2-point increase in `Ideology` on the probability of each answer?