Problem Set 1

Francisco Brady

9 Sep 2024

#### 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.

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

##### 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:

Note that the third term on the right hand side can be simplified if we rewrite the 1 as and perform the subtraction.

Again, let the parameter 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 given , we can write the above as a likelihood function.

Try to find the value of that maximizes the expression. What is the value of associated with this value of ? Does this make sense?

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

Find the value of 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 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 for each person. We now can model the parameter as . Re-write the log likelihood function with this change. Note: normally, we would include an intercept as well, so that , but let’s keep it simpler for now and ignore the intercept.

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

##### g) Given this value of , what is the probability that the first voter, with = 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.

. use GSS2014subset, clear  
  
. probit abany age childs educ polviews relpersn  
  
Iteration 0: Log likelihood = -1084.8368   
Iteration 1: Log likelihood = -909.52299   
Iteration 2: Log likelihood = -908.95059   
Iteration 3: Log likelihood = -908.95042   
Iteration 4: Log likelihood = -908.95042   
  
Probit regression Number of obs = 1,575  
 LR chi2(5) = 351.77  
 Prob > chi2 = 0.0000  
Log likelihood = -908.95042 Pseudo R2 = 0.1621  
  
─────────────┬────────────────────────────────────────────────────────────────  
 abany │ Coefficient Std. err. z P>|z| [95% conf. interval]  
─────────────┼────────────────────────────────────────────────────────────────  
 age │ .0019953 .0021531 0.93 0.354 -.0022246 .0062152  
 childs │ -.0465874 .0232186 -2.01 0.045 -.092095 -.0010798  
 educ │ .0945726 .0120598 7.84 0.000 .0709358 .1182094  
 polviews │ -.2348614 .0251512 -9.34 0.000 -.2841568 -.185566  
 relpersn │ -.3304787 .0368112 -8.98 0.000 -.4026273 -.2583302  
 \_cons │ -.2199313 .2089871 -1.05 0.293 -.6295384 .1896758  
─────────────┴────────────────────────────────────────────────────────────────

* Education has a positive and significant association with support for legal abortion
* A higher score on the political ideology index (polviews) is associated with lower support for abortion.
* A higher value on the relpersn scale, which indicates a less religious person, is associated with lower support for abortion
* Age and number of children were not significantly associated with support for abortion

##### 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.

Using the coefficients from the probit regression output and the following values:  
- age = 35 - childs = 2 - educ = 16 - polviews = 3 - relpersn = 2

$$

$$

##### 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?

* age = 65
* childs = 2
* educ = 16
* polviews = 5
* relpersn = 2

$$

$$

Taking the difference of the two values, the difference is: $$

$$

##### 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.

. margins  
  
Predictive margins Number of obs = 1,575  
Model VCE: OIM  
  
Expression: Pr(abany), predict()  
  
─────────────┬────────────────────────────────────────────────────────────────  
 │ Delta-method  
 │ Margin std. err. z P>|z| [95% conf. interval]  
─────────────┼────────────────────────────────────────────────────────────────  
 \_cons │ .4557695 .0111934 40.72 0.000 .4338309 .4777081  
─────────────┴────────────────────────────────────────────────────────────────

##### 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.