Professor: Pam Paxton
TA: Nicholas Reith

SOC 385L Lab Assignment 7

Limited Data Analysis: Logit and Probit Regression Models for Binary Outcome Variables

The data set (*binlfp2.dta*) contains information about 753 married women's family characteristics, childbearing experiences, and their personal and spousal demographic characteristics (i.e., education and age). In this lab assignment, we will use this data set to understand how these family and personal and spousal characteristics predict married women's labor market participation. Our outcome variable (lfp) is a binary variable with 0 indicating not participating in the paid work and 1 indicating participating in the labor market. This assignment requires you to accomplish the following tasks:

In this assignment, we use 7 variables to predict married women's labor force participation. These 7 predictor variables are **k5, k618, age, wc, hc, lwg, and inc**. The binary outcome variable is **lfp**. As always, use Stata commands such as *desc, summarize, tab (with options like 'm' and 'nolabel'* to explore and understand the data and these before you start regression analysis.

Overall, you need to know how to interpret the coefficients substantively and answer attached questions.

Coding and Questions:

1. Estimate two types of regression models (i.e., logit and probit) for binary outcome variables and compare the results from these two modeling methods.
   - Do some simple math and divide some logit coefficients by their corresponding probit coefficients. Note that the ratio ranges between 1.6 and 1.7.

2. Obtain predicted probabilities of working outside of home for each observation (i.e., each sampled woman) in the data. (use **predict prlogit**)
   - Sum these predicted probabilities
   - Can you tell me what the average probability of working outside of the home is for the married women in this sample?
   - Make a histogram
   - Look at the datatable for individual probabilities
   - Comment on each of the above

Then, you will employ different approaches to discussing and exhibiting your results:

3. Obtain predicted probabilities of labor force participation for certain types of women by specifying values for the independent variables.

3.1. Independent variables take their mean values

- What is the predicted probability of labor force participation for an "average woman"?

3.2. Change values of the independent variables representing different types of women whose labor market participation you are interested in and contrast the probabilities of participating in labor market for one type of women to another.

Specify the characteristics of women for their predicted probabilities. According to the output from using the above command (prvalue), you see that Stata, by default, uses the mean values of independent variables to compute the predicted probabilities. However, variables such as wc and hc are dummy variables (1 or 0) indicating whether or not wives and husbands have a college degree. Categorical variables, such as hc and wc, can only take on two values, 0 and 1. It wouldn't make much sense to compute how probability would change if, say, wc changed from 0 to .6, because that cannot happen. Therefore, we may use the same command "**prvalue**" but using its options to specify values for the independent variables that reflect the "real world" to obtain predicted probabilities.

- What is the predicted probability of participating in the labor market for women who are *young and have low family income, live in a family of low education, and have young children* at home?
- What is the predicted probability of the labor force participation for women who are 35 years of age, have two kids below age 5, have no college education and neither does her husband, and have family income below the median level and set the rest independent variables in the parenthesis after **x** to their mean values**?**
- What is the predicted probability of participating in the labor market for women who are *older, both her and her husband are highly educated, and have no children* at home?
- Try one other combination on your own.
- Comment on each of the above.

3.3. Display changes in the probabilities when the values of certain variables change (else are set equal to the means). For example, marginal changes, also known as *marginal effect, partial effect*, and discrete changes. (Hint: use **prtab** and **prchange** for this section)

- What are the predicted probabilities of participating in the labor market for women who have different numbers of young kids at home? What are the differences in the probabilities (i.e., discrete changes in the probabilities) between women with different numbers of young kids at home?
- Combine 2 or 3 variables to see how probabilities change with changes in multiple variables. For example **prvalue k5 k618 wc**

- Comment on the above.
- Note how **prtab** and **prchange** differ in the results they report.

4. Obtain "odds ratio" (exponential of the coefficient) to help you interpret the effects of independent variables (for logit regression models). Odds ratios are more interpretable than the logit coefficients. You can get these odds ratios even more easily when you are fitting a logit regression model: logit lfp k5 k618 age wc hc lwg inc, or

Instead of reporting coefficients, you will only get odds ratios in the Stata output table.

Using listcoef is beneficial in the sense that you can get both coefficients and odds ratios at the same time in the same output table.

- Please substantively interpret the effects of these predictor variables on labor force participation for married women.

5. Obtain the "odds" ratio as a percentage change instead of a factor change, and interpret for the logit model.

- Comment/interpret these effects and note how they differ from those in #4 above.

6. I won't make you do a probit interpretation, which is more complicated. However, please take a look at this website: http://www.ats.ucla.edu/stat/stata/output/Stata_Probit.htm

The particular section copied below helps with understanding the interpretation of coefficients.

However, interpretation of the coefficients in probit regression is not as straightforward as the interpretations of coefficients in linear regression or logit regression. The increase in probability attributed to a one-unit increase in a given predictor is dependent both on the values of the other predictors and the starting value of the given predictors. For example, if we hold ***gre*** *and* ***topnotch*** *constant at zero, the one unit increase in* ***gpa*** *from 2 to 3 has a different effect than the one unit increase from 3 to 4 (note that the probabilities do not change by a common difference or common factor):*

and the effects of these one unit increases are different if we hold ***gre*** *and* ***topnotch*** *constant at their respective means instead of zero:*

However, there are limited ways in which we can interpret the individual regression coefficients. A positive coefficient means that an increase in the predictor leads to an increase in the predicted probability. A negative coefficient means that an increase in the predictor leads to a decrease in the predicted probability.