Econ 140 - Spring 2016
Section 10

GSI: Fenella Carpena

April 13, 2016

Regressions with Binary Dependent Variables

Let us first review binary variables (also known as dummy variables). It is a variable that takes on only two values: 0 and 1. For example, Y can be defined to indicate whether a student passed a midterm; or Y can indicate whether an individual's loan application was approved. In each of these examples, we can let $Y = 1$ denote one of the outcomes and $Y = 0$ the other outcome.

We have previously seen binary variables as independent variables (i.e. X's) in our regression, but for this part of the course, we consider the case where our dependent variable Y is a dummy variable. How can we estimate a regression model with a binary dependent variable? As you will see in this section, we can use the following 3 models: (1) Linear Probability Model, (2) Probit, and (3) Logit.

Throughout this section, we will use the following example. Suppose we are interested in investigating the determinants of women's labor force participation. Our dependent variable is $inlf$ ("in the labor force") which is a binary variable equal to 1 if the woman reports working for a wage outside the home at some point during the year, 0 otherwise. Our independent variables are *educ* (years of education) and $kidslt6$ (number of children less than 6 years old).

1 Linear Probability Model (LPM)

Specification. The **Linear Probability Model** is given by

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i$$

where Y_i is a binary variable. In other words, the LPM is just the name that we use for a multiple linear regression model with a binary dependent variable. It is called a Linear Probability Model because it gives us the *probability* that Y equals 1, and this probability is *linear* in the parameters β_j.

Why is it the case that the LPM gives us the probability that the dependent variable is equal to 1? Consider a multiple linear regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$ where Y is a binary variable. By the first least squares assumption,

$$E(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

and since Y is a binary variable, we know that

$$E(Y|X) = 1 \cdot P(Y = 1|X) + 0 \cdot P(Y = 0|X) = P(Y = 1|X)$$

so putting the above two equations together, we get the important equation

$$P(Y = 1|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2.$$

In summary, this means that for a binary dependent variable, the expected value from the population regression is the probability that $Y = 1$, given X.

Estimation Method. OLS, which as before minimizes the sum of squared residuals (SSR). For example, in the case of 2 explanatory variables:

$$\min_{\widehat{\beta}_0,\widehat{\beta}_1,\widehat{\beta}_2} \sum_{i=1}^{n} \widehat{u}_i^2 = \min_{\widehat{\beta}_0,\widehat{\beta}_1,\widehat{\beta}_2} \sum_{i=1}^{n} (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_{1i} - \widehat{\beta}_2 X_{2i})^2$$

Example. Consider the results of the following LPM:

$$\widehat{inlf} = 0.052 + 0.046 \cdot educ - 0.224 \cdot kidslt6$$

1. **What is the predicted probability of labor force participation for a woman who has 12 years of education and 2 children under the age of 6 years old? 3 children under the age of 6 years old?** To get the predicted probability when $educ = 12$ and $kidslt6 = 2$, we compute $\widehat{\beta_0} + \widehat{\beta_1} * 12 + \widehat{\beta_2} * 2$. That is, $0.052 + 0.046 * 12 - 0.224 * 2 = 0.156$. Thus, for a woman who has 12 years of education, and 2 children under the age of 6, we would predict that her probability of participating in the labor force is 15.6%. To get the predicted probability when $educ = 12$ and $kidslt6 = 3$, again we compute $0.052 + 0.046 * 12 - 0.224 * 3 = -0.068$. Note that we are getting a negative probability, which is non-sensical.

   This example illustrates one of the disadvantage of using the LPM. We may get predicted probabilities that are less than 0 or above 1, since there are is nothing in the model that constrains the predicted values to be between 0 and 1.

2. **Interpret the coefficient on *educ*.** To interpret the coefficient, we need to remember that because we have an LPM, we are looking at the probability that $Y = 1$. Hence, the coefficient on *educ* means that holding *kidslt6* constant, another year of education is associated with an increase in the *probability* of being in the labor force by 0.034.

3. **For a woman with 16 year of education, what is the predicted change in probability of labor force participation when going from 0 to 1 young child? From 1 young child to 2?** In both cases, the change in the predicted probability is $\widehat{\beta}_2 = -0.224$.

Advantages/Disadvantages. The example regression above illustrates the advantages and disadvantages of using the LPM. Specifically, the advantage is that it is very simple to estimate and use, since it is basically a multiple regression model.

However, the disadvantages of the LPM are that: (1) the fitted probabilities can be greater than 1 or less than 0, as seen in Example # 1 above, and (2) it assumes constant marginal effects; for example, as seen in Example # 3 above, the predicted drop in the probability of working when going from 0 children to 1 young child is exactly the same as wehn going from 1 young child to 2. It seems more realistic that the first small child would reduce the probability by a large amount, but then subsequent children would have a smaller marginal effect.

Statistical Inference. Confidence intervals, t-test, and F-test that we've learned in the past still apply and can be constructed in the same way (assuming large sample size). However, errors in the LPM are always heteroskedastic, so robust standard errors should always be used.

Why are errors in LPM always heteroskedastic? Consider a regression $Y = \beta_0 + \beta_1 X + u$ where Y is a binary variable. Then, $var(u|X) = var(Y|X) = P(Y = 1|X) \cdot [1 - P(Y = 1|X)] = (\beta_0 + \beta_1 X) \cdot (1 - \beta_0 - \beta_1 X)$. This means that the $var(u|X)$ is not constant, as its value depends on X; hence u is heteroskedastic.

Measures of Fit. In the LPM, the R^2 is not a particularly useful statistic. One way to see this is that the $R^2 = ESS/TSS$, where $ESS = \sum(\widehat{Y}_i - \overline{Y})^2$, and with the LPM, we can get non-sensical fitted values $\widehat{Y}$, as shown in Example # 1 above. As a result, the R^2 is of limited interest here.

2 Probit

Specification. The **Probit Regression Model** with k regressors is given by

$$P(Y = 1|X) = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)$$

where Φ is the standard normal CDF. Since a CDF is always between 0 and 1, the probit forces the predicted probabilities to be between 0 and 1 as well.

Estimation Method. We can no longer use OLS since the probit is not linear in the parameters β_j (the β's appear "inside" the function Φ). Instead, we use the **Maximum Likelihood Estimator (MLE).** Specifically, we chooses $\widehat{\beta}_0, \ldots, \widehat{\beta}_k$ that maximizes the log-likelihood function

$$\max_{\widehat{\beta}_0,\ldots,\widehat{\beta}_k} \sum_{i=1}^{n} Y_i \cdot ln[\Phi(\beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki})] + (1 - Y_i) \cdot ln[\Phi(\beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki})]$$

There are no closed-form solutions to the above maximization problem, so the solution must be found through numerical algorithms. Also note that the maximum likelihood estimator is consistent and normally distributed in large samples.

Example. Consider the following probit results:

$$\widehat{P(inlf|educ, kidslt6)} = \Phi(-1.259 + 0.129 \cdot educ - 0.621 \cdot kidslt6)$$

1. **What is the predicted probability of labor force participation for a woman who has 12 years of education and 2 children under the age of 6 years old? 3 children under the age of 6 years old?** To find the predicted probability when $educ = 12$ and $kidslt6 = 2$, we first calculate the z-value, $z = \widehat{\beta}_0 + \widehat{\beta}_2 * 12 + \widehat{\beta}_3 * 2 = -1.259 + 0.129 * 12 - 0.621 * 3 = -0.953$. Then we use the normal table to find $\Phi(-0.953) = 0.17$.

   To find the predicted probability when $educ = 12$ and $kidslt6 = 3$, we find $\Phi(-1.574) = 0.058$. Note that in comparison to the LPM case, we are getting a value that makes sense (it is greater than zero), since we have used the Normal CDF which is always between 0 and 1.

2. **Interpret the coefficient on *educ*.** We need to be careful when interpreting coefficients in a probit model. We can say something about the sign of the coefficient and its relationship with the probability that $Y = 1$. However, the coefficient has no direct interpretation in terms of the probability that $Y = 1$.

   Hence, we can say education is positively related to the probability of being in the labor force. However, all we can say about the size of the coefficient is that a 1 year increase in education is associated with a 0.129 increase in the $\underline{z\text{-value}}$, holding all $kidslt6$ constant.

3. **For a woman with 16 year of education, what is the predicted change in probability of labor force participation when going from 0 to 1 young child? From 2 young children to 3?** We first calculate the probabilities for the following:

   - Predicated probability when $educ = 16$, $kidslt6 = 0$: $\Phi(-1.259 + 0.129 * 16 - 0.621 * 0) = \Phi(0.805) = 0.790$
   - Predicated probability when $educ = 16$, $kidslt6 = 1$: $\Phi(-1.259 + 0.129 * 16 - 0.621 * 1) = \Phi(0.184) = 0.573$

   Hence, taking the difference between the two, the predicted change in probability in labor force participation when going from 0 to 1 young child is a decline of .217 points. Similarly, we can calculate:

   - Predicated probability when $educ = 16$, $kidslt6 = 2$: $\Phi(-1.259 + 0.129 * 16 - 0.621 * 2) = \Phi(-0.437) = 0.331$

- Predicated probability when $educ = 16$, $kidslt6 = 3$: $\Phi(-1.259 + 0.129 * 16 - 0.621 * 3) = \Phi(-1.058) = 0.145$

Taking the difference between the two, the predicted change in probability in labor force participation when going from 0 to 1 young child is a decline of 0.186 points.

Let's compare our results here to that of LPM. As mentioned above, one of the disadvantages of the LPM is that the marginal effect of X is constant. In the LPM example above, whether we went from 0 to 1, or 2 to 3 children, the change in the predicted probability is -0.224. In comparison, the probit is non-linear, so the effect of a change in X depends on the starting value of X. In the probit example here, when going from 0 to 1, the change was 0.217, but when going from 2 to 3, the change was 0.186.

Advantages/Disadvantages. The above example again illustrates the advantages and disadvantages of probit. The advantage is that it overcomes the challenges of LPM: predicted probabilities from probit are always between 0 and 1, and the probate incorporates non-linear effects of X as well. However, a potential disadvantage is that the coefficients are difficult to interpret. We cannot directly interpret the size of the coefficient from a probit model, in the same way that we do in LPM.

Statistical Inference. Confidence intervals, t-test, and F-test that we've learned in the past still apply and can be constructed in the same way (assuming large sample size). Again use heteroskedastic SEs.

Measures of Fit. Two commonly used measures of fit are the following.

- Fraction correctly specified: This approach uses the following rule. If $Y_i = 1$ and the predicted probability exceeds 50%, or if $Y_i = 0$ and the predicted probability is less than 50%, then Y_i is said to be correctly predicted. Otherwise Y_i is said to be incorrectly predicted. Then, the "fraction correctly specified" is the fraction of n observations $Y_1, \ldots, Y_n$ that are correctly specified.
- Pseudo R^2: $1 - \frac{\mathcal{L}_{ur}}{\mathcal{L}_o}$ where $\mathcal{L}_{ur}$ is the maximized value of the log-likelihood function for the estimated model, and $\mathcal{L}_o$ is the maximized value of the log-likelihood function in a model with only an intercept.

3 Logit

Specification. The **Logit Regression Model** with k regressors is given by

$$P(Y = 1|X) = \Lambda(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)$$

where Λ is the CDF of the standard logistic distribution $\Lambda(z) = \frac{1}{1+exp(-z)}$, so that

$$\Lambda(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k) = \frac{1}{1 + exp[-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)]}.$$

As before, since a CDF is always between 0 and 1, the logit forces the predicted probabilities to be between 0 and 1 as well.

Estimation Method. Again, we can no longer use OLS since the logit is not linear in the parameters β_j (the β's appear "inside" the function Λ). As before, we use the maximum likelihood estimator, which is the same as in the probit except that we replace the CDF Φ with Λ.

Example. Consider the following probit results:

$$\widehat{P(inlf|educ, kidslt6)} = \Lambda(-2.053 + 0.210 \cdot educ - 1.010 \cdot kidslt6)$$

1. **What is the predicted probability of labor force participation for a woman who has 12 years of education and 2 children under the age of 6 years old? 3 children under the age of 6 years old?** If $educ = 12$ and $kidslt6 = 2$, $z = \widehat{\beta}_0 + \widehat{\beta}_2 * 12 + \widehat{\beta}_3 * 2 = -2.053 + 0.210 * 12 - 1.010 * 2 = -1.553$. Then $\Lambda(-1.553) = 1/(1 + exp(-(-1.553))) = 0.174$.

   If $educ = 12$ and $kidslt6 = 2$, $z = -2.053 + 0.210 * 12 - 1.010 * 3 = -2.563$. Then, $\Lambda(-2.563) = 1/(1 + exp(-(-2.563))) = 0.072$.

2. **Interpret the coefficient on *educ*.** Same as in the probit case.

3. **For a woman with 16 year of education, what is the predicted change in probability of labor force participation when going from 0 to 1 young child? From 2 young children to 3?** We first calculate the probabilities for the following:

   - Predicated probability when $educ = 16$, $kidslt6 = 0$: $\Lambda(-2.053 + 0.210 * 16 - 1.010 * 0) = \Lambda(1.307) = 0.787$
   - Predicated probability when $educ = 16$, $kidslt6 = 1$: $\Lambda(-2.053 + 0.210 * 16 - 1.010 * 1) = \Lambda(0.297) = 0.573$

   Hence, taking the difference between the two, the predicted change in probability in labor force participation when going from 0 to 1 young child is a decline of 0.214. Similarly, we can calculate:

   - Predicated probability when $educ = 16$, $kidslt6 = 2$: $\Lambda(-2.053 + 0.210 * 16 - 1.010 * 2) = \Lambda(-.713) = 0.329$
   - Predicated probability when $educ = 16$, $kidslt6 = 3$: $\Lambda(-2.053 + 0.210 * 16 - 1.010 * 3) = \Lambda(-1.723) = 0.152$

   Taking the difference between the two, the predicted change in probability in labor force participation when going from 0 to 1 young child is a decline of 0.177 points.

Advantages/Disadvantages. Same as probit.

Statistical Inference. Same as probit.

Measures of Fit. Same as probit.

4 Exercises

Exercise 1. (Review the Concepts, Chapter 11.) Suppose that a linear probability model yields a predicted value of Y that is equal to 1.3. Explain why this is non-sensical.

Since Y is binary, its predicted value is the probability that $Y = 1$. A probability must be between 0 and 1, so the value of 1.3 is nonsensical.

Exercise 2. (Review the Concepts, Chapter 11.) On of your friends is using data on individuals to study the determinants of smoking at your university. She asks you whether she should use a probit, logit or linear probability model, what advice do you give here? Why?

She should use a logit or probit model. These models are preferred to the linear probability model because they constrain the regression's predicted values to be between 0 and 1. Usually, probit and logit regressions give similar results, and she should use the method that is easier to implement with her software (though with current computers and software, both are very easy to implement). Nevertheless, would be good to estimate LPM as well, and check that the predicted probabilities obtained across all three models are similar (and if not, to investigate why).

Exercise 3. (Review the Concepts, Chapter 11.) Why are the coefficients of the probit and logit models estimated by maximum likelihood instead of OLS?

OLS cannot be used because the regression function is not a linear function of the regression coefficients (the coefficients appear inside the nonlinear functions Φ or Λ).

Exercise 4. You want to estimate how income affects the probability of voting for a republican candidate. The result from your logit model is the following:

$$\widehat{P(Republican} = 1|Income) = \Lambda(-1.00 + 0.02Income)$$

where $Income$ is measured in thousands of dollars.

(a) What is the probability of voting republican if the income is 10,000 dollars?

(b) What is the marginal effect of income, evaluated at the sample mean of income of 50,000 dollars?

(a) The probability is given by $\Lambda(-1.00 + 0.02 * 10) = \Lambda(-0.8) = 1/(1 + exp(0.8)) = 0.31$

(b) We can take the derivative

$$\frac{\partial \widehat{P(Republican} = 1|Income)}{\partial Income} = \frac{\partial}{\partial Income}\Lambda(\widehat{\beta}_0 + \widehat{\beta}_1 * Income) = \frac{\widehat{\beta}_1 * exp(\widehat{\beta}_0 - \widehat{\beta}_1 Income)}{(1 + exp(\widehat{\beta}_0 - \widehat{\beta}_1 Income))^2}$$

$$= \frac{-0.02 * exp(1 - 0.02Income)}{(1 + exp(1 - 0.02))^2} = \frac{0.02 * exp(1 - 0.02 * 50)}{(1 + exp(1 - 0.02 * 50))^2}$$

where in the last step we have plugged in $Income = 50$, since the problem asks us to use the sample mean of income. Doing so yields $0.02/4 = 0.005$.

Final Exam 2009, Question 3. In order to better target their marketing, a cell phone company wants to assess the determinants of cell phone subscribership. Their consultant estimates the following Linear Probability Model (LPM) using a sample of individuals older than 18 years of age:

$$CELL_i = \underset{(0.121)}{0.30} - \underset{(0.005}{0.02}\, AGE_i + \underset{(0.020)}{0.05}\, EDUC_i + \underset{(0.006)}{0.10}\, ln(INCOME_i)$$

where $CELL$ is a dummy variable equal to 1 if person i subscribes to a cell service, 0 otherwise.

(a) What is the interpretation of the estimate of the coefficient on ln(INCOME)?

The coefficient gives the change in the probability of owning a cell phone caused by a 1% increase in individual's income holding other variables constant. Since the coefficient is 0.10, a 1% increase in income is associated with an increase in CELL of 0.001, or 0.1 percentage points in probability of cell ownership, holding all other variables constant.

(b) Suppose you want to forecast the probability that a 50-year-old individual with 12 years of education, and income of $40,000 will own a cell phone. What prediction does this model make and does it make economic sense? [HINT: ln(40,000) = 10.6.]

Plugging in these values we predict: 0.30 + (-0.02)(50) + 0.05(12) + (0.1)(10.6) = 0.96. Makes sense since this demographic is likely to have high rates of cell phone subscription, and also since the estimate is between 0.0 and 1.0.

(c) Describe two problems with least squares estimation of the coefficients in the LPM when some variables, such as ln(INCOME), can get very large or even infinite.

The first assumption of least squares estimation, $E[u|X] = 0$, is violated in this situation, creating a correlation between the regressor(s) and the error term, leading to biased and inconsistent coefficient estimates using least squares. Also, the variance of the error terms is going to be correlated with the values of X, causing heteroskedasticity.

(d) Instead of using a LPM, you recommend to the consultant to estimate a probit model using the dummy dependent variable CELL. Give the specification of the model and explain how to compute the estimates of the coefficients in the probit model.

The specification is $CELL_i = \Phi(\beta_0+\beta_1 AGE_i+\beta_2 EDUC_i+\beta_3 ln(INCOME_i))$ where Φ is the standard normal CDF. We estimate the coefficients using the maximum likelihood estimator. Specifically, we choose $\widehat{\beta}_0, \ldots, \widehat{\beta}_3$ that maximizes the log-likelihood function

$$\max_{\widehat{\beta}_0,\ldots,\widehat{\beta}_3} \sum_{i=1}^{n} CELL_i \cdot ln[\Phi(\beta_0 + \beta_1 AGE_i + \beta_2 EDUC_i + \beta_3 ln(INCOME_i)]$$

$$+\sum_{i=1}^{n}(1 - CELL_i) \cdot ln[\Phi(\beta_0 + \beta_1 AGE_i + \beta_2 EDUC_i + \beta_3 ln(INCOME_i)].$$

There are no closed-form solutions to the above maximization problem, so the solution must be found through numerical algorithms.