ØAMET4100 · Spring 2019
Lecture Note 4B

Instructor: Fenella Carpena

January 31, 2019

This lecture note provides a review of regressions with binary dependent variables (Stock & Watson, Chapter 11). This lecture note is not intended to be a comprehensive review of lecture or the textbook, since there is a lot more material than we have time to cover. However, I have tried to focus on the concepts which I believe are necessary to be successful in our class.

1 Regressions with Binary Dependent Variables: Overview

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, we can use the following 3 models: (1) Linear Probability Model, (2) Probit, and (3) Logit.

Throughout the discussion, 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).

2 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. To estimated the LPM, we can use 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. To better understand how LPM works, let's consider Exercise 1.1 from Worksheet 4B. The results of the LPM regression are:

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

(a) 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.

(b) 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 4.6 percentage points.

(c) 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 $\widehat{\beta}_2 = -0.224$ (i.e., a decreased in the predicted probability of being in the labor force by 22.4 percentage points).

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 disadvantage of the LPM is that the fitted probabilities can be greater than 1 or less than 0, as seen above.

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 $\widehat{Y}$. Another way to see this is that R^2 is a measure of how the close the data points are to the line, so that when $R^2 = 1$, the data points are exactly on the line. When the dependent variable is binary, the it is possible to have $R^2 = 1$ (unless the independent variables are also binary). For these reasons, the R^2 is of limited interest here.

3 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.

What is the intuition for MLE? The likelihood function is the joint probability distribution of the data, as a function of unknown coefficients. The maximum likelihood estimator chooses β's to maximize the likelihood function, which in turn is the joint probability distribution. Thus, MLE chooses β's to maximize the probability of drawing the data we actually observe. Another way to say this is that maximum likelihood estimates the parameters that are "most likely" to have produced the data we observe.

Example. To better understand how probit works, let's consider Exercise 1.2 from Worksheet 4B. The results of the probit regression are:

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

(a) 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.

(b) 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 z-value, holding all $kidslt6$ constant. Note that the z-value refers to the value that goes into the Φ function, i.e., $-1.259 + 0.129 \cdot educ - 0.621 \cdot kidslt6$.

(c) For a woman with 16 years 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 need to carry out the following calculations.

- Predicted probability when $educ = 16$, $kidslt6 = 0$: $\Phi(-1.259+0.129*16-0.621*0) = \Phi(0.805) = 0.790$
- Predicted 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 0.217 points. Similarly, we can calculate:

- Predicted probability when $educ = 16$, $kidslt6 = 2$: $\Phi(-1.259+0.129*16-0.621*2) = \Phi(-0.437) = 0.331$
- Predicted 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 2 to 3 young children is a decline of 0.186 points.

Let's compare our results here to that of LPM. In Exercise 1.1, 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 case here, when going from 0 to 1 child, the change was 0.217, but when going from 2 to 3 children, the change was 0.186.

Advantages/Disadvantages. The above exercise 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 probit 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 we have a large sample size). Again, we must 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.

4 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. To better understand how logit works, let's consider Exercise 1.3 from Worksheet 4B. The results of the logit regression are:

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

(a) 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, $\Lambda(-2.563) = 1/(1 + exp(-(-2.563))) = 0.072$.

(b) Interpret the coefficient on *educ*.

Same as in the probit case: a one-year increase in education is associated with a 0.129 increase in the z-value, holding $kidslt6$ constant.

(c) 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:

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

- Predicted probability when $educ = 16$, $kidslt6 = 2$: $\Lambda(-2.053+0.210*16-1.010*2) = \Lambda(-.713) = 0.329$
- Predicted 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 2 to 3 young children is a decline of 0.177 points. Notice that the results we get here is very similar to the probit case.

Advantages/Disadvantages, Statistical Inference, and Measures of Fit. All of these are the same as with probit.