Introduction to Econometrics: Recitation 7

Seung-hun Lee

November 22nd, 2021

1 Binary Dependent Variables

1.1 Linear Probability Models

We now turn to the case where the dependent variable Y_i takes either 0 or 1. This type of regression can be used to study how independent variable(s) X_i is(are) correlated to yes/no questions in the survey. As always, assume a following regression equation

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

where Y_i is a binary variable. Unlike previous regressions, the complication arises when we attempt to interpret the equation. Especially, what does β_1 now mean? To study this question, we first look into the expected value of Y_i conditional on X_i, $E[Y_i|X_i]$. The conditional mean of Y_i is, by definition

$$E[Y_i|X_i] = 0 \times \Pr(Y_i = 0|X_i) + 1 \times \Pr(Y_i = 1|X_i)$$

In the context of this regression equation, we can obtain the conditional mean of Y_i as follows

$$\begin{aligned} E[Y_i|X_i] &= E[\beta_0 + \beta_1 X_i + u_i|X_i] \\ &= \beta_0 + \beta_1 X_i + E[u_i|X_i] \\ (\because E[u_i|X_i] = 0) &= \beta_0 + \beta_1 X_i \end{aligned}$$

Therefore, we established that $E[Y_i|X_i]$ in this context is the probability of $Y_i = 1$ given X_i.

Now we move back to β_1, notice that $\beta_1 = \frac{\Delta Y_i}{\Delta X_i}$ and $\Delta Y_i = \text{Change in } \Pr(Y_i = 1|X_i)$ with respect to change in X_i , or

$$\Delta Y_i = \Pr(Y_i = 1|X_i = x + \Delta X_i) - \Pr(Y_i = 1|X_i = x)$$

and when we calculate $\Pr(Y_i = 1|X_i = x + \Delta X_i) - \Pr(Y_i = 1|X_i = x) = E[Y_i|X_i = x + \Delta X_i] - E[Y_i|X_i = x]$, we get

$$\beta_0 + \beta_1(x + \Delta X_i) - \beta_0 + \beta_1(x) = \beta_1 \Delta X_i$$

So we get $\Delta Y_i = \beta_1 \Delta X_i \iff \beta_1 = \frac{\Delta Y_i}{\Delta X_i}$. Therefore, β_1 now measures how much the predicted probability of $Y_i = 1$ changes with respect to X_i

The **linear probability model** is the estimation in which you run an OLS on the type of regression equation where Y_i is a binary dependent variable. The advantage is that it is simple - there is no difference in terms of methods between this and the OLS methods we have learned so far. However, there are some critical disadvantages to this model. One is that by setting the regression model as above, we are assuming that the change of predicted probability of $Y_i = 1$ is constant for all values of X_i. But more critically, it is possible that because of the way functional form is specified, the predicted probability $\hat{y}$ may be greater than 1 or strictly less than 0. Given that probability is defined to be in between 0 and 1, this could be a preposterous result. In addition, the distribution of the error term is no longer normal distribution. This could affect the asymptotic (large sample) properties of the OLS estimators.

1.2 Logit and Probit Regressions

Since linear probability models can exceed 1 or fall below 0, we now use a class of **sigmoid estimators**. These estimators are bounded between 0 and 1. Therefore, using these estimators will prevent the predicted probability of $Y_i = 1$ falling out of $[0, 1]$ range.

One of such estimator is **logit regression**. For notational convenience, I write $Z_i = \beta_0 + \beta_1 X_i$. Logit regression assumes that the cumulative probability of Z_i, which is $\Pr(Y_i = 1|X_i)$ is distributed as

$$\Pr(Y_i = 1|X_i) = F(Z_i) = \frac{1}{1 + e^{-Z_i}}$$

In this setup, when $Z_i \to \infty, e^{-Z_i} \to 0$ Therefore, $F(Z_i) \to 1$. Likewise, taking $Z_i \to -\infty$ results in $F(Z_i) \to 0$.

To see the role of the independent variable X_i, we should note that changes in X_i leads to changes in Z_i, since $Z_i = \beta_0 + \beta_1 X_i$. The change in Z_i should also impact $F(Z_i)$. Borrowing the logic from the chain rule, we get

$$\frac{\partial F}{\partial X_i} = \frac{\partial F}{\partial Z_i} \frac{\partial Z_i}{\partial X_i}$$

where $\frac{\partial Z_i}{\partial X_i} = \beta_1$. This says that β_1 alone does not explain how changes in X_i alters probability of $Y_i = 1$ given X_i. Therefore, the coefficient value of β_1 does not mean that much in logit regression (similar for probit regressions). However, Note that $\frac{\partial F}{\partial Z_i}$ is calculated as

$$\frac{\partial F}{\partial Z_i} = \frac{e^{-\beta_0-\beta_1 X_i}}{(1+e^{-\beta_0-\beta_1 X_i})^2} > 0$$

This implies that the sign of $\frac{\partial F}{\partial X_i}$ is determined by the sign of $\frac{\partial Z_i}{\partial X_i}$, which is β_1. Therefore the sign of β_1 still matters. In fact, the interpretation of logit coefficients (and probit) regression matters up to the sign of the coefficients in general.

Probit regression is largely similar with the logit regression, except now that the cumulative probabiltiy of Z_i is assumed to be a standard normal function. Specifically,

$$F(Z_i) = \Phi(Z_i) = \Phi(\beta_0 + \beta_1 X_i)$$

where $\Phi(v)$ means the cumulative normal function $\Pr(Z \leq v)$. Again, the value of β_1 coefficient does not mean as much as the sign. By taking the similar approach with the logit regression, we get

$$\frac{\partial F}{\partial X_i} = \frac{\partial F}{\partial Z_i}\frac{\partial Z_i}{\partial X_i}$$

where $\frac{\partial F}{\partial Z_i}$ is the pdf of a standard normal distribution.

In practice, the two regressions are run together to check whether the signs of the coefficients are the same. There are not many differences between the two.

1.3 Maximum Likelihood Estimation Method

Notice that both logit and probit regressions are nonlinear in the sense that the β_0, β_1 parameters are no longer in linear relationship with the $X_i's$ and subsequently Y_i's. One of the assumptions used in using OLS is that the linear regression assumption. This is no longer a valid option anymore, which requires a different approach. This is where **maximum likelihood estimation** comes in. To understand the maximum likelihood estimators, you must understand what the likelihood function is. A **likelihood function** is the conditional density of $Y_1, ..., Y_n$ given $X_1, ..., X_n$ that is treated as the function of the unknown parameters (β_0, β_1 in our case). In other words, since we have the observations for Y_i's and X_i's, but do not know the values of the parameter β_i's, what we are trying to do here is to find the values of β_i's that best matches the values of X_i's and Y_i's. As a result, the maximum likelihood estimators is the value of β_i's that best describes the data and maximizes the value of the likelihood function

To nail this home, let's not worry about regression equation for the moment and consider a single variable - Y_i. For example, Let's assume that Y_i's are IID normal with mean μ and standard error σ, both of which are unknown. The joint probability of Y_i's are (our likelihood function)

$$\begin{aligned}
\Pr(Y_1 = y_1, ..., Y_n = y_n|\mu, \sigma) &= \Pr(Y_1 = y_1|\mu, \sigma) \times .. \times \Pr(Y_n = y_n|\mu, \sigma) \\
&= \prod_{i=1}^{n} f(y_i|\mu, \sigma) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(Y_i-\mu)^2}{2\sigma^2}} \\
&= (2\pi)^{-\frac{n}{2}} (\sigma^2)^{-\frac{n}{2}} e^{-\sum_{i=1}^{n} \frac{(Y_i-\mu)^2}{2\sigma^2}}
\end{aligned}$$

A convenient way to solve this class of problem is to use the *log-likelihod function*. Taking logs to above equation gets us

$$-\frac{n}{2} \ln (2\pi) - \frac{n}{2} \ln \sigma^2 - \sum_{i=1}^{n} \frac{(Y_i - \mu)^2}{2\sigma^2} \tag{*}$$

To find the maximum likelihood estimator of μ, we differentiate the above with respect to μ. This gets us

$$\begin{aligned}
2\sum_{i=1}^{n} \frac{(Y_i - \mu)}{2\sigma^2} = 0 &\implies \sum_{i=1}^{n} \frac{(Y_i - \mu)}{\sigma^2} = 0 \\
(\because \sigma, \text{ though unknown, will be a constant }) &\implies \sum_{i=1}^{n} (Y_i - \mu) = 0 \\
&\implies n\mu = \sum_{i=1}^{n} Y_i \implies \mu_{MLE} = \frac{1}{n}\sum_{i=1}^{n} Y_i
\end{aligned}$$

We can do the similar for σ^2 by differentiating $(*)$ with respect to σ^2. This leads to

$$\begin{aligned}
-\frac{n}{2\sigma^2} + \sum_{i=1}^{n} \frac{2(Y_i - \mu)^2}{(2\sigma^2)^2} = 0 &\implies \frac{n}{2\sigma^2} = \sum_{i=1}^{n} \frac{(Y_i - \mu)^2}{2\sigma^4} \\
&\implies n = \sum_{i=1}^{n} \frac{(Y_i - \mu)^2}{\sigma^2} \\
&\implies n = \frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - \mu)^2
\end{aligned}$$

By imputing μ_{MLE} in place of μ, we get

$$\sigma^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu_{MLE})^2$$