Lecture 15: Limited Dependent Variables

Jaakko Meriläinen

5304 Econometrics @ Stockholm School of Economics

Limited Dependent Variables

- Many outcomes, as observed in the real world, reflect a binary choice:
  - You either enroll in school/college or you do not
  - You either go to the doctor or not
  - You either accept a job offer or not
- We have analyzed some outcomes like these in some of the examples we have seen
- Today we will review the application of OLS to such problems but also introduce non-linear models of binary choice

Plan for This Lecture

1. Introduction
2. The linear probability model revisited
3. LPM versus probit/logit
4. Conclusions and references

Applying OLS to a Binary Choice

- We have previously looked at work investigating the effect of various regressors on a binary dependent variable
- We modeled this as:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u$$

- With a binary outcome, $E(y|x) = P(y = 1|\mathbf{x}) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$
- The linear regression model with a binary outcome is called the Linear Probability Model (LPM)
- The OLS estimator looks the same as always: $\beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$

Problems with the LPM

- The LPM is widely-used in economics (for some good reasons!)
- But it has important limitations:
  - Conceptually, a probability cannot be related to independent variables linearly for all possible values
  - Predicted values of the outcome ($\widehat{y}$) can be outside the interval between 0 and 1
  - The LPM is naturally heteroskedastic
- We will discuss the implications of these issues in a bit but let us first look at some alternatives

A Latent Variable Motivation: Binary Response

- Let us look at the case of a binary outcome such as whether you apply for a PhD at SSE or not
  - Let us say this depends on your characteristics (grades, motivation, sex, age, place of residence and so on)

- One simple way of formulating this choice model is to think of a latent variable (utility) from the two options: (a) applying to PhD and (b) your reservation option (a high-paying consultancy job), both of which are a function of your grades x_i

- Let the utility from the two options be:
  - PhD: $U_i^S = \alpha_0^S + \alpha_1^S x_i + \mu_i^S$
  - McKinsey: $U_i^H = \alpha_0^H + \alpha_1^H x_i + \mu_i^H$

- This is a very simple example of an additive random utility model

A Latent Variable Motivation: Binary Response

- Define
  - $\beta_0 \equiv \alpha_0^S - \alpha_0^H$
  - $\beta_1 \equiv \alpha_1^S - \alpha_1^H$
  - $\epsilon_i \equiv \mu_i^S - \mu_i^H$

- You choose to apply for a PhD if:

$$y^* \equiv \beta_0 + \beta_1 x_i + \epsilon_i > 0$$

i.e., if the utility from a PhD is greater than the utility from going to work at McKinsey

A Latent Variable Motivation

- Define the "latent variable" y^* as:

$$y^* \equiv \beta_0 + \beta_1 x_i + \epsilon_i$$

- The outcome y_i is observed as follows:

$$y_i = \begin{cases} 0 & \text{if } y^* < 0 \\ 1 & \text{if } y^* \geq 0 \end{cases}$$

- In order to estimate this model, we will need to make some distributional assumptions about the error term

Assume Normally Distributed Errors...

- **Assumption:** ϵ_i is i.i.d with a standard normal distribution, independent of x_i:

$$\epsilon_i | x_i \sim \mathcal{N}(0, 1)$$

- The probability density function of a standard normal distribution is denoted by $\phi(x)$:

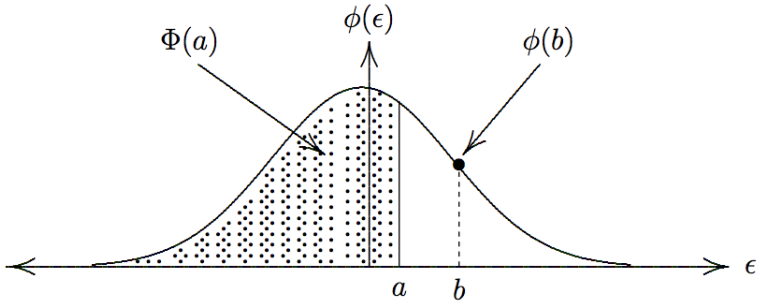
$$\phi(x) = \frac{1}{\sqrt{2\pi}} exp(-\frac{x^2}{2})$$

- Then the cumulative distribution function ($\Phi(x)$) would be:

$$\Phi(x) = \int_{-\infty}^{x} \phi(z) dz$$

The Standard Normal Distribution

Figure 1.2: **The standard normal: probability density ($\phi(\cdot)$) and cumulative density ($\Phi(\cdot)$)**



The Probit Model

- With our distributional assumption, we can write:

$$\begin{aligned} \Pr(y_i = 1|x_i) &= \Pr(\beta_0 + \beta_1 x_i + \epsilon_i > 0 \mid x_i) \\ &= \Pr(-\epsilon_i \leq \beta_0 + \beta_1 x_i \mid x_i) \\ &= \Phi(\beta_0 + \beta_1 x_i) \end{aligned}$$

  where we utilize the symmetry of the normal distribution

- Since probabilities sum to 1, $\Pr(y_i = 0|x_i) = 1 - \Phi(\beta_0 + \beta_1 x_i)$

Estimation of the Probit Model: Maximum Likelihood

- Estimation of non-linear models is typically done by **maximum likelihood**
  - See Appendix C in Wooldridge (2013) for a very basic introduction

- Write down the likelihood for the ith individual as:

$$\begin{aligned}\mathcal{L}_i(\beta_0, \beta_1; y_i|x_i) &= \Pr(y = 1|x_i)^{y_i} \Pr(y_i = 0|x_i)^{1-y_i} \\ &= \Phi(\beta_0 + \beta_1 x_i)^{y_i}[1 - \Phi(\beta_0 + \beta_1 x_i)]^{1-y_i}\end{aligned}$$

- The log likelihood therefore is:

$$\ell_i(\beta_0, \beta_1; y_i|x_i) = y_i ln\Phi(\beta_0 + \beta_1 x_i) + (1 - y_i) ln[1 - \Phi(\beta_0 + \beta_1 x_i)]$$

Estimation of the Probit Model

- Under the assumption of random sampling, we can stack the log-likelihood across the N individuals as follows:

$$\ell(\beta_0, \beta_1; \mathbf{y}|\mathbf{x}) = \sum_{i=1}^{N} \{y_i \ln \Phi(\beta_0 + \beta_1 x_i) + (1 - y_i) \ln[1 - \Phi(\beta_0 + \beta_1 x_i)]\}$$

- The parameters are estimated numerically using various algorithms (such as Newton-Raphson, BHHH etc.)

- Maximum likelihood estimators are consistent, asymptotically normally distributed and efficient

- If you want to read more on maximum likelihood, pitched at a PhD level, see Cameron and Trivedi (2005) or Wooldridge (2002)

A General Formulation

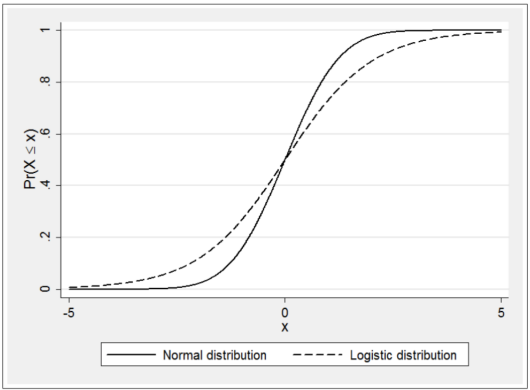
- The probit model is an example of a larger class of models
- The variables combine linearly into a single index, which then passes through a separate function (called a **link function**) which translates this index into predicted probabilities
  - In the probit case, this is the CDF of the standard normal distribution
  - If instead, we had adopted a logistic function, this would give us the logit model
- The logit model relies on the assumption that

$$\Pr(\epsilon \leq Z|x) = \Lambda(Z) = \frac{exp(Z)}{1 + exp(Z)}$$

- The link function comes from our assumption about the distribution

The Normal and Logistic Distributions

Figure 2.1: **Cumulative density functions: normal and logistic distributions**



Interpreting Results from a Probit Model

- Interpreting the coefficients from a probit (logit) model is not straightforward
- The "single index" passes through the link function (G) to generate the effect on predicted probability
  - What this means is that the effect of x on $Pr(y = 1)$ differ across the distribution of x
  - Unlike the LPM, you cannot just read off the coefficient of the probit (logit) model to get the marginal effect
- The marginal effect of a small change in x on $Pr(y = 1)$ is given by

$$\frac{\partial \Pr(y = 1)}{\partial x_j} = g(\beta_0 + \mathbf{x}\beta)$$

where $g(z) \equiv \frac{dG}{dz}(z)$

Interpreting Results from a Probit Model

- The **sign** of the marginal effect is the same as the coefficient

- But the **magnitude** will differ depending on the value of x and the value of the other variables in the model
- Thus, to make any sense of coefficients from probit and logit models, we need to compute and examine marginal effects
- The most common computation is for marginal effects at the mean of all variables

- This is different, e.g., from the average marginal effect!

Should you use an LPM?

- This is a surprisingly divisive subject!
- We have spoken about the cons of the LPM:
  - A linear in parameters model is not attractive for modelling probabilities over an unbounded range!
  - Predictions could be outside the interval between 0 and 1
  - There is heteroskedasticity
- To enough people, these reasons are adequate for ex ante always preferring a probit or a logit for a binary outcome over LPM as a matter of principle

Should You Use a LPM?

- But the LPM also carries some important advantages
- Interpretation is **much** easier:
  - We typically care about marginal effects;
  - In the LPM, I can read these off the coefficient!
- Fixed effects are much easier to include
  - There are non-linear special cases (Poisson models etc.) but in general fixed effects do not work the same way
- Panel data in general is often easier to work with in a linear framework
- As indeed is dealing with endogenous variables
- So the question is: how serious are the issues with the LPM?

Should You Use a LPM?

- If your purpose is to get at the marginal effect of some variable, the advantages of the LPM will frequently outweigh the potential issues:
  - For most of the range, probit and logit distributions are also very flat (except at the tails)
  - So the marginal effects really look like LPM most of the time anyway!
  - Whereas the cost for interpretation is incurred every time!

- Remember: one of the justifications by Angrist and Pischke for the linear regression was as the best linear approximation to a non-linear CEF
  - Turns out that this is often pretty good
  - Even though in theory it is always objectionable!

- But this is not a consensus view by any means!

Should You Use a LPM? Angrist and Pischke (2009) Say...

Why then, should we bother with nonlinear models and marginal effects? One answer is that the marginal effects are easy enough to compute now that they are automated in packages like Stata. But there are a number of decisions to make along the way (e.g., the weighting scheme, derivatives versus finite differences) while OLS is standardized. Nonlinear life also promises to get considerably more complicated when we start to think about IV and panel data. Finally, extra complexity comes into the inference step as well, since we need standard errors for marginal effects.

Should You Use a LPM?

- What if you are looking for predictions of individual $\widehat{y}$?
- The situation is different when you are interested not in average effects or marginal effects but in the predictions for each individual i
- In those situations, we would usually always turn towards some non-linear model (whether logit or probit)
- The reason is that then predicted probabilities outside the unit interval are quite objectionable
  - One example is when we will look at the probability of attrition in a panel in order to "weight" observations
  - We want the predicted value to be in the unit interval for everyone in these cases
  - And so, we would almost always end up using a probit (or a logit)
- These applications are, however, much rarer than the simple case of looking for average marginal effects

A Side Note About IV

- One thing you may be tempted to do is to...
  - use a probit or a logit to estimate the first-stage in an IV specification
  - and then use the fitted values for the second stage

- **DO NOT DO THIS!**

- This is what is called a "forbidden regression"

- The reason is that 2SLS with the linear first stage (and satisfying IV assumptions) is consistent
  - With a non-linear version in general it is not!
  - See the section on 2SLS mistakes in MHE for details

Readings

Limited Dependent Variables

Important references

- Wooldridge (2013) Introductory Econometrics: A Modern Approach, Chapter 17 (Section 17.1)
- Angrist and Pischke (2009): Mostly Harmless Econometrics, Section 4.6.1 (On forbidden regressions)

Other

- Angrist and Pischke (2009): Mostly Harmless Econometrics, Section 3.4.2 (skip part about Tobit)