

## **Lecture 6:**

### **Estimating Probit and Logit Models by Maximum Likelihood**

**(Hayashi, pp. 507-511; Greene 19.1-19.4):**

- 1. Binary outcomes and the linear probability model (LPM).**
- 2. Index Function Models in General**
- 3. Probit**
- 4. Logit**

## Notation note:

So far in the course, we have followed the text pretty consistently in defining  $\mathbf{x}_i$  as the  $K \times 1$  column vector of data for observation  $i$  in the data. Likewise,  $\boldsymbol{\beta}$  has been  $K \times 1$ , so that the outcome for observation is written  $y_i = \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i$ .

From here on, we'll do the more typical thing, where  $\boldsymbol{\beta}$  is  $1 \times K$ , so  $y_i = \boldsymbol{\beta} \mathbf{x}_i + \varepsilon_i$ . In general, somewhat sloppily, both  $\boldsymbol{\beta} \mathbf{x}_i$  and  $\mathbf{x}_i \boldsymbol{\beta}$  will denote the cross-product of  $\mathbf{x}_i$  and  $\boldsymbol{\beta}$ .

## 1. Binary outcomes and the linear probability model (LPM).

Suppose the outcome we are interested in understanding is not a continuous variable (like weekly earnings), but a binary outcome (work/not work; survive for one year/not; ever marry/not; etc.)

One way to study such choices is to define a 0-1 indicator for the outcome, then regress that outcome on a set of  $X$ s by OLS. e.g.:

$$y_i = Work_i = \beta \mathbf{x}_i + \varepsilon_i, \text{ where } Work_i \in \{0,1\} \quad (1)$$

Such a regression is called a *linear probability model (LPM)*.

### *Useful features of LPMs:*

1. as long as assumptions (1)-(3) of the classical model (linearity, exogeneity, no multicollinearity) continue to hold, OLS estimates of (1) still yield unbiased estimates of  $\beta$ , and of the conditional expectation of  $y$ , given  $\mathbf{X}$ .
2. the conditional expectation of  $y$  and the predicted level of  $y$  have an intuitive interpretation as the *probability* that the outcome variable equals 1, given  $\mathbf{X}$ . The estimated coefficients give the derivative of this probability with respect to the  $X$ s, i.e. the “marginal effects”.

### *Drawbacks of LPMs:*

1. *by assumption*, the error term,  $\varepsilon_i$ , must be heteroskedastic. Why?  $y_i$  can take on only two values, 0 and 1. Therefore, given  $\mathbf{x}_i$ ,  $\varepsilon_i$  can only take on two values also:  
also:  $\varepsilon_i = y_i - \beta \mathbf{x}_i$   
=  $1 - \beta \mathbf{x}_i$  when  $y_i = 1$ , which (given  $E(\varepsilon_i) = 0$ ) occurs with prob.  $\beta \mathbf{x}$   
=  $-\beta \mathbf{x}_i$  when  $y_i = 0$ , which (given...) occurs with prob.  $1 - \beta \mathbf{x}$

$$\text{Therefore, } \text{Var}(\varepsilon_i) = E(\varepsilon_i^2) = (1 - \beta \mathbf{x}_i)^2 \beta \mathbf{x}_i + (-\beta \mathbf{x}_i)^2 (1 - \beta \mathbf{x}_i)$$

$$= (1 - \beta x_i) \cdot [(1 - \beta x_i) \beta x_i + (\beta x_i)^2]$$

$$= (1 - \beta x_i) \cdot [\beta x_i - \beta^2 x_i^2 + \beta^2 x_i^2]$$

$$= (1 - \beta x_i) \cdot \beta x_i$$

$$= E(y|\mathbf{x}) \cdot E((1-y)|\mathbf{x})$$

So, the variance of the error term will be higher when the values of  $\mathbf{x}$  are such that the predicted probability of the outcome is close to .5, and lowest when the predicted probability of the outcome is close to 0 or 1.

(On its own, this drawback isn't a game-changer, because (a) we now have estimates of  $\text{Var}(b)$  that are robust to the homoskedasticity assumption, and (b) the form of the heteroskedasticity is known, so we could also correct for it explicitly in the estimation, via GLS).

2. It also follows from (1) above that the residuals cannot possibly be normally distributed. So it's not at all obvious that we can test hypotheses using the 'standard' OLS approach. [turns out also to be less of a problem than one might think—multivariate normality of the  $\mathbf{X}$ s fixes this].

3. If any of the  $\mathbf{x}_i$  are continuous on the real line, then assumption (1) of the classical model (linearity) can't possibly hold. Why? for large enough  $\mathbf{x}_i$ , the LPM asserts that the true probability that  $y_i = 1$  is negative, or exceeds 1. So, even though the LPM has a number of useful and simple features, it can only be seen as a linear *approximation* to a model where the dependent variable is the probability that a binary outcome equals one.

It is this limitation

(plus an advantage of the alternative—the fact that index function models map in a nice and direct way into simple models of utility maximization)

that has led econometricians to develop index function models of discrete choice.

## 2. Index Function Models

Rather than writing the (binary) outcome variable  $y_i$ , as a linear function of  $\mathbf{x}_i$ , index function models all posit the existence of an underlying, *continuous* (“latent”), index  $y_i^*$ , related to  $\mathbf{x}_i$  via:

$$y_i^* = \boldsymbol{\beta}\mathbf{x}_i + \varepsilon_i, \text{ where } \varepsilon_i \text{ is drawn (usually iid) from the pdf } f(\varepsilon_i) \quad (2)$$

The underlying index,  $y_i^*$ , is not observed by the econometrician. The data that we see (a bunch of binary outcomes) is instead assumed to be generated by:

$$y_i = 1 \text{ iff } y_i^* > 0, \text{ (i.e. if } \varepsilon_i > -\boldsymbol{\beta}\mathbf{x}_i)$$

$$y_i = 0 \text{ iff } y_i^* < 0, \text{ (i.e. if } \varepsilon_i < -\boldsymbol{\beta}\mathbf{x}_i)$$

So, if  $y_i^*$  exceeds a threshold (zero), the binary outcome is a one, otherwise it is a zero. Note: as long as  $\mathbf{x}$  includes a constant, the assumption that the threshold is zero is without loss of generality).

One advantage of this formulation is that  $\mathbf{x}$  can have a linear effect on  $y^*$ , which can range from minus to plus infinity, while the outcome variable takes on only two values, 0 and 1.

Another advantage is that it maps in a very intuitive and direct way into a simple model of utility-maximization by the entity (firm, consumer, worker, etc) making the binary choice under study.

For example, imagine

Utility of working  $\equiv U_i^W = \boldsymbol{\beta}^W \mathbf{x}_i + \varepsilon_i^W$ , where  $\mathbf{x}_i$  are an individual's characteristics, e.g. education, available wage rate, number of young children, and  $\varepsilon_i^W$  is the individual's *unobserved tastes for work*.

Utility of not working  $\equiv U_i^N = \boldsymbol{\beta}^N \mathbf{x}_i + \varepsilon_i^N$ , where  $\mathbf{x}_i$  are an individual's characteristics, e.g. education, available wage rate, number of young children, and  $\varepsilon_i^N$  is the individual's *unobserved tastes for "leisure"*.

If the individual chooses the option yielding the highest utility, he/she works iff:

$$U_i^W > U_i^N, \text{ or: } \boldsymbol{\beta}^W \mathbf{x}_i + \varepsilon_i^W > \boldsymbol{\beta}^N \mathbf{x}_i + \varepsilon_i^N, \text{ or:}$$

$$\varepsilon_i^W - \varepsilon_i^N > \boldsymbol{\beta}^N \mathbf{x}_i - \boldsymbol{\beta}^W \mathbf{x}_i, \text{ or: } \varepsilon_i > (\boldsymbol{\beta}^N - \boldsymbol{\beta}^W) \mathbf{x}_i, \text{ or } \varepsilon_i > -(\boldsymbol{\beta}^W - \boldsymbol{\beta}^N) \mathbf{x}_i, \text{ or } \varepsilon_i > -\boldsymbol{\beta} \mathbf{x}_i$$

This is equivalent to the general index function model (2) above, where  $\varepsilon_i = \varepsilon_i^W - \varepsilon_i^N$  and  $\boldsymbol{\beta} = \boldsymbol{\beta}^W - \boldsymbol{\beta}^N$ .

### **So, how do we estimate Index Function Models?**

Their structure lends itself naturally to ML methods. In pretty much all cases, we simply:

- assume a distribution of unobservables,  $f(\varepsilon_i)$
- write out the likelihood function as a function of  $\boldsymbol{\beta}$  and any unknown parameters of  $f(\varepsilon_i)$ .
- maximize the likelihood; then we can derive asymptotic standard errors and perform hypothesis tests as for *any* ML estimation.

### 3. Probit regression.

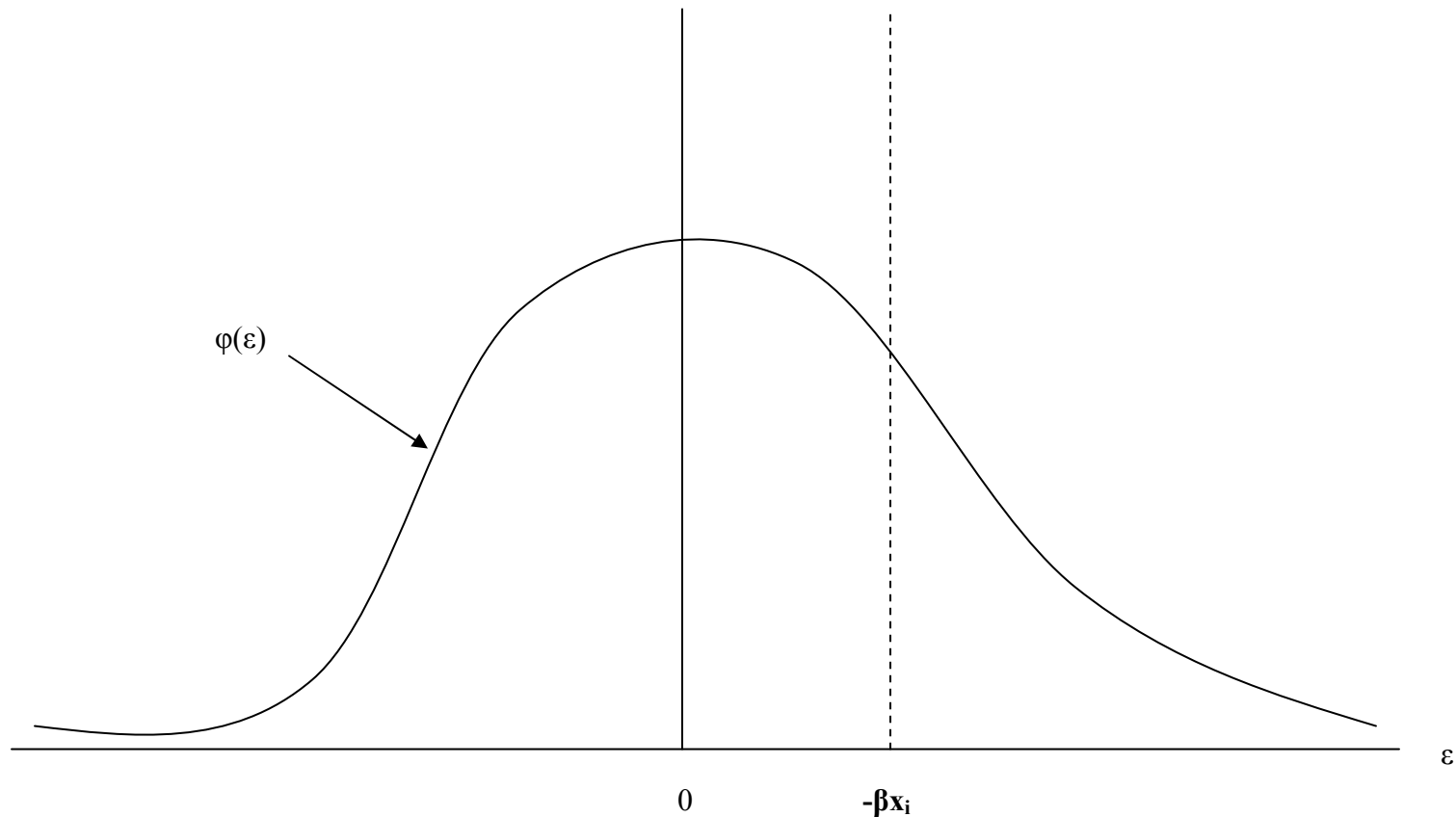
In the probit model,  $f(\varepsilon_i)$  is assumed to be normal, with mean zero and variance  $\sigma^2$ .

Notes:

1. As in the classical linear model, the assumption of  $E(\varepsilon_i)=0$  is without loss of generality, as long as there is a constant term in  $\mathbf{x}_i$ .
2. In the classical linear model, unbiased estimates of  $\sigma^2$  (for example,  $s^2$ ) exist. In probit,  $\sigma^2$  is not identified. (To see this, write the likelihood as a function of  $\boldsymbol{\beta}$  and  $\sigma$ . Then, note that the likelihood depends only on  $\boldsymbol{\beta}/\sigma$ , never on  $\sigma$  or  $\boldsymbol{\beta}$  alone. Thus, *all we can ever know* is  $\boldsymbol{\beta}/\sigma$  --for example, the effect of young children on the relative taste for work *relative to* the dispersion of relative tastes for work in the population). For this reason, we generally just set  $\sigma=1$ , and interpret our estimated coefficients as estimates of  $\boldsymbol{\beta}/\sigma$ .

Thus, in a probit model,

If an observation has  $y_i = 1$ , it must be true that  $\varepsilon_i > -\beta \mathbf{x}_i$  for that observation. Thus we know that this observation's  $\varepsilon_i$  falls to the right of the dashed vertical line in the picture below:



The likelihood of this occurring is just  $1 - \Phi(-\beta \mathbf{x}_i)$ , where  $\Phi$  is the standard cumulative normal distribution.

By the same reasoning, if an observation has  $y_i = 0$ , its likelihood must be just  $\Phi(-\boldsymbol{\beta}\mathbf{x}_i)$ .

Putting these two things together, the likelihood of any individual observation is given by:

$$L_i = [1 - \Phi(-\boldsymbol{\beta}\mathbf{x}_i)]^{y_i} \cdot [\Phi(-\boldsymbol{\beta}\mathbf{x}_i)]^{(1-y_i)}$$

The log likelihood of an individual observation is therefore:

$$\log(L_i) = y_i \log[1 - \Phi(-\boldsymbol{\beta}\mathbf{x}_i)] + (1 - y_i) \log [\Phi(-\boldsymbol{\beta}\mathbf{x}_i)]$$

The log likelihood of the the entire sample is just the sum, over  $i$ , of the individual contributions above.

Maximizing this numerically, using the same methods as when we estimated the classical linear model by ML in the assignment, yields the ML estimates of  $\boldsymbol{\beta}$ .

## Interpreting ML estimates of $\beta$ .

*But what do these estimates tell us?*

Recall that in the classical linear model,  $E(y_i) = \sum_{k=1}^K \beta_k x_{ik}$ ,

so that  $\frac{\partial E(y_i)}{\partial x_k} = \beta_k$ . Thus,  $\beta_k$  gives us the effect on  $y$  of a one-unit change in  $x_k$ ,

holding all the other  $x$ 's constant.

$\beta_k$  has the same interpretation in the linear probability model, with  $E(y_i)$  interpreted as the *probability* that  $y_i = 1$ .

But that is *not true* of the  $\beta_k$  that is estimated in a probit model. Instead, we now

have  $y_i^* = \beta \mathbf{x}_i + \varepsilon_i$ , so that  $\frac{\partial E(y_i^*)}{\partial x_k} = \beta_k$ . Thus,  $\beta_k$  gives us the effect of a one-

unit change in  $x_k$ , on  $y^*$ , where  $y^*$  is not a probability but a standard normal variate ranging in value from minus to plus infinity.

To get the effect of a unit increase in  $x_k$  on the probability that  $y = 1$ , go back to our construction of the likelihood function and recall that the probability (likelihood) that  $y = 1$  is:

$$\Pr(y_i = 1 | \mathbf{x}_i) = [1 - \Phi(-\boldsymbol{\beta}\mathbf{x}_i)]$$

Thus, 
$$\frac{\partial \Pr(y_i = 1 | \mathbf{x}_i)}{\partial x_k} = \beta_k \phi(-\boldsymbol{\beta}\mathbf{x}_i)$$

The desired quantity,  $\frac{\partial \Pr(y_i = 1 | \mathbf{x}_i)}{\partial x_k}$ , is known as the “marginal effect” of  $x$  on the binary outcome  $y$ . To calculate it, we have to multiply our estimate of  $\beta$  by the density of  $\varepsilon$ , which will vary with the level of  $\mathbf{x}$ .

Does this make sense? Sure, in fact it illustrates an important property of the probit model, namely that the effect of any variable,  $x$  on the *probability* that  $y=1$ , *must* vary with the level of  $\mathbf{x}$  if  $\text{prob}(y=1)$  is to remain between 0 and 1.

For example, suppose  $\mathbf{x}$  is such that, at the baseline level of  $x_k$ ,  $\text{prob } y=1$  is very high, e.g. .999999. It follows we are in the right tail of the standard normal distribution, so  $\phi(-\beta\mathbf{x}_i)$  is basically zero. Even if  $\beta_k$  is very large, it is impossible for  $x_k$  to raise  $\text{prob}(y=1)$  by very much.

Now example, suppose  $\mathbf{x}_i$  is such that, at the baseline level of  $x_k$ ,  $\text{prob } y=1$  is very low, e.g. .0000001. It follows we are in the left tail of the standard normal distribution, so  $\phi(-\beta\mathbf{x}_i)$  is once again basically zero. Even if  $\beta_k$  is very large, it is impossible for  $x_k$  to reduce  $\text{prob}(y=1)$  by very much.

It is in the middle of the distribution, when  $\mathbf{x}_i$  is such that the baseline predicted probability is around .5, that changes in  $\mathbf{x}$  have the greatest latitude to affect the probability that  $y=1$ . Viewed in the utility-maximizing interpretation of probits, it is when lots of persons in your sample are close to the margin of indifference between the two choice options that changes in  $\mathbf{x}$  will have their largest impact on the *share* of your sample choosing one choice over the other.

.

## 4. Logit

Recall that when we introduced index function models, we defined them for a general pdf of the error term,  $f(\varepsilon_i)$ .

Probit (short for “probability unit”) regression uses the unit normal distribution for  $f(\varepsilon_i)$ .

Logit simply assumes a logistic distribution for  $f(\varepsilon_i)$ . (This looks a lot like the normal—symmetric, unimodal, with mean and mode 0—, but with fatter tails).

Following the same logic as before, the likelihood of an individual observation is now:

$L_i = [1 - F(-\beta \mathbf{x}_i)]^{y_i} \cdot [F(-\beta \mathbf{x}_i)]^{(1-y_i)}$ , where  $F$  is the cdf of the logistic distribution.

One convenience (minor these days with fast computers) of the logit relative to the probit is that there are closed-form solutions for the logistic cdf.

Using these yields:

$$F(-\beta\mathbf{x}_i) = \frac{1}{1 + \exp(\beta\mathbf{x}_i)}, \quad \text{and:} \quad 1 - F(-\beta\mathbf{x}_i) = \frac{\exp(\beta\mathbf{x}_i)}{1 + \exp(\beta\mathbf{x}_i)}$$

Now that we have the likelihoods, we can estimate the model via ML just as before.

Of course, one can construct and estimate binary choice models using any assumed distribution. Things can, however, get tricky if there are nuisance parameters (like  $\sigma$  in the probit case) in the distribution to be estimated (are they identified?), and when the distribution is not continuously differentiable on the entire real line (e.g. the uniform distribution). Overall, there is not much to be gained by playing with other distributions, and mostly the choice is one of convenience.

As we will see when we look at multinomial choices, the logit has some really convenient features.