

## LECTURE 15 BINARY RESPONSE MODELS

This lecture discusses models in which the dependent variable takes only two values, usually zero or one. For example,  $Y_i$  can measure whether or not individual  $i$  participates in the labor force, with  $Y_i = 1$  (yes) and  $Y_i = 0$  (no). The linear regression model may not be appropriate in such cases.

### Bernoulli trials

The econometrician observes  $\{Y_i : i = 1, \dots, n\}$ , where the  $Y_i$  are iid random variables. The distribution of  $Y_i$  is given by

$$Y_i = \begin{cases} 1 & \text{with probability } \theta, \\ 0 & \text{with probability } 1 - \theta. \end{cases}$$

This is the Bernoulli( $\theta$ ) distribution, and

$$\begin{aligned} E[Y_i] &= \theta \\ &= \Pr(Y_i = 1). \end{aligned} \tag{1}$$

The PMF of  $Y_i$  is

$$p(y_i, \theta) = \theta^{y_i} (1 - \theta)^{1 - y_i} \text{ for } y_i = 0, 1.$$

Thus, the log-likelihood function is given by

$$\begin{aligned} \log L_n(\theta) &= \left( n^{-1} \sum_{i=1}^n Y_i \right) \log \theta + \left( 1 - n^{-1} \sum_{i=1}^n Y_i \right) \log(1 - \theta) \\ &= \bar{Y}_n \log \theta + (1 - \bar{Y}_n) \log(1 - \theta), \end{aligned}$$

where  $\bar{Y}_n = n^{-1} \sum_{i=1}^n Y_i$ . The ML estimator of  $\theta$  is the sample mean:

$$\hat{\theta}_n^{ML} = \bar{Y}_n.$$

Computing the second derivative,

$$\frac{d^2 \log p(Y_i, \theta)}{d\theta^2} = -\frac{Y_i}{\theta^2} - \frac{1 - Y_i}{(1 - \theta)^2}.$$

Equation (1) implies that

$$\begin{aligned} E \left[ \frac{d^2 \log p(Y_i, \theta)}{d\theta^2} \right] &= -\frac{1}{\theta} - \frac{1}{1 - \theta} \\ &= -\frac{1}{\theta(1 - \theta)}. \end{aligned}$$

Thus, in this model, the information (scalar) is given by

$$I(\theta) = \frac{1}{\theta(1 - \theta)},$$

and the results in Lecture 14 imply that

$$\begin{aligned} \bar{Y}_n &\rightarrow_p \theta, \\ n^{1/2} (\bar{Y}_n - \theta) &\rightarrow_d N(0, \theta(1 - \theta)). \end{aligned}$$

We can estimate the asymptotic variance consistently by  $\bar{Y}_n (1 - \bar{Y}_n)$ . A  $1 - \alpha$  asymptotic confidence interval for  $\theta$  can be constructed as follows:

$$\left[ \bar{Y}_n \pm z_{1-\alpha/2} \sqrt{\frac{\bar{Y}_n (1 - \bar{Y}_n)}{n}} \right].$$

The Bernoulli trials model is univariate. We now extend it to the case where the probability of  $Y_i$  taking the value 1 is a function of some exogenous explanatory variables.

## Linear probability model

Suppose that the econometrician observes  $\{(Y_i, X_i) : i = 1, \dots, n\}$ , where the  $Y_i$  are binary  $\{0, 1\}$  variables, and the  $X_i$  are  $k$ -vectors of exogenous variables that explain  $\Pr(Y_i = 1 | X_i)$ . Let us assume that

$$\Pr(Y_i = 1 | X_i) = F(X_i^\top \beta),$$

for some function  $F$ . Unlike the Bernoulli trials model, the probability of  $Y_i$  taking the value one or zero is not constant and depends on some observable characteristics  $X_i$ .

Under the linear regression model,

$$E[Y_i | X_i] = X_i^\top \beta,$$

then

$$F(X_i^\top \beta) = X_i^\top \beta.$$

That is, the probability of  $Y_i = 1$  is a linear function of  $X_i$ . Such an assumption is a good starting point for the analysis, but it suffers from a serious flaw. If some of the components of  $X_i$  are continuous,  $X_i^\top \beta$  cannot be restricted to the zero-one interval, and the model may yield probabilities outside  $[0, 1]$ . To avoid such predictions, we require  $F$  to map into  $[0, 1]$ , which forces  $F$  to be nonlinear. Any CDF is a natural choice for  $F$ .

Despite this limitation, the LPM remains useful in practice for estimating average partial effects, particularly when the main interest is in a treatment coefficient. Under correct specification of the conditional expectation, OLS provides a consistent estimator of the average marginal effect without requiring a distributional assumption on the errors.

## Probit and logit models

It is convenient to introduce a latent (unobservable) variable  $Y_i^*$ , which can be interpreted as the net benefit or unobserved utility that determines the agent's binary decision. The usual linear regression model holds for  $Y_i^*$ :

$$Y_i^* = X_i^\top \beta + U_i.$$

We assume further that  $\{(Y_i^*, X_i) : i = 1, \dots, n\}$  are iid, and that

$$Y_i = \begin{cases} 1 & \text{if } Y_i^* > 0, \\ 0 & \text{if } Y_i^* \leq 0. \end{cases}$$

Let  $F$  be the conditional CDF of  $U_i$  given  $X_i$ :

$$F(u | X_i) = \Pr(U_i \leq u | X_i).$$

Then,

$$\begin{aligned} \Pr(Y_i = 1 | X_i) &= \Pr(Y_i^* > 0 | X_i) \\ &= \Pr(X_i^\top \beta + U_i > 0 | X_i) \\ &= \Pr(U_i > -X_i^\top \beta | X_i) \\ &= 1 - F(-X_i^\top \beta | X_i). \end{aligned}$$

Assume further that the distribution of  $U_i$  does not depend on  $X_i$ , so  $F(u)$  replaces  $F(u | X_i)$ . The usual assumption in this framework is that  $F$  is symmetric around zero, that is,

$$F(u) = 1 - F(-u).$$

Under this assumption,

$$\Pr(Y_i = 1 | X_i) = F(X_i^\top \beta).$$

The common choices for  $F$  are

- Probit model:  $F$  is the standard normal CDF, denoted by  $\Phi$ .
- Logit model:  $F$  is the logistic CDF, denoted by  $\Lambda$ :

$$\begin{aligned} \Lambda(X_i^\top \beta) &= \frac{\exp(X_i^\top \beta)}{\exp(X_i^\top \beta) + 1} \\ &= \frac{1}{1 + \exp(-X_i^\top \beta)}. \end{aligned}$$

For either choice of  $F$ , the model is estimated by maximum likelihood. The conditional PMF of  $Y_i$  given  $X_i = x_i$  takes the same form as in the Bernoulli model, with  $\theta$  replaced by  $F(x_i^\top \beta)$ :

$$p(y_i, \beta | x_i) = F(x_i^\top \beta)^{y_i} (1 - F(x_i^\top \beta))^{1-y_i} \text{ for } y_i = 0, 1.$$

Thus, the log-likelihood is given by

$$\log L_n(\beta) = n^{-1} \sum_{i=1}^n Y_i \log F(X_i^\top \beta) + n^{-1} \sum_{i=1}^n (1 - Y_i) \log (1 - F(X_i^\top \beta)).$$

The first-order condition for  $\hat{\beta}_n^{ML}$  is

$$\begin{aligned} 0 &= \frac{\partial \log L_n(\hat{\beta}_n^{ML})}{\partial \beta} \\ &= n^{-1} \sum_{i=1}^n Y_i \frac{\partial F(X_i^\top \hat{\beta}_n^{ML}) / \partial \beta}{F(X_i^\top \hat{\beta}_n^{ML})} - n^{-1} \sum_{i=1}^n (1 - Y_i) \frac{\partial F(X_i^\top \hat{\beta}_n^{ML}) / \partial \beta}{1 - F(X_i^\top \hat{\beta}_n^{ML})} \\ &= n^{-1} \sum_{i=1}^n Y_i \frac{f(X_i^\top \hat{\beta}_n^{ML})}{F(X_i^\top \hat{\beta}_n^{ML})} X_i - n^{-1} \sum_{i=1}^n (1 - Y_i) \frac{f(X_i^\top \hat{\beta}_n^{ML})}{1 - F(X_i^\top \hat{\beta}_n^{ML})} X_i, \end{aligned}$$

where  $f$  is the density corresponding to  $F$ :

$$f(u) = \frac{dF(u)}{du}.$$

No closed-form expression exists for  $\hat{\beta}_n^{ML}$ ; it must be computed numerically. Standard statistical software produces the ML estimates and asymptotic standard errors for both models.

For the logit model, the derivative of  $\Lambda$  satisfies

$$\frac{d\Lambda(u)}{du} = \Lambda(u)(1 - \Lambda(u)). \quad (2)$$

Therefore, for the logit model,

$$\begin{aligned}\frac{f(X_i^\top \beta)}{F(X_i^\top \beta)} &= 1 - \Lambda(X_i^\top \beta), \\ \frac{f(X_i^\top \beta)}{1 - F(X_i^\top \beta)} &= \Lambda(X_i^\top \beta).\end{aligned}$$

The first-order condition simplifies to

$$\begin{aligned}\frac{\partial \log L_n(\hat{\beta}_n^{ML})}{\partial \beta} &= n^{-1} \sum_{i=1}^n \left( Y_i - \Lambda(X_i^\top \hat{\beta}_n^{ML}) \right) X_i \\ &= 0.\end{aligned}$$

Similarly,

$$\frac{\partial \log p(y_i, \beta | x_i)}{\partial \beta} = (y_i - \Lambda(x_i^\top \beta)) x_i,$$

and

$$\frac{\partial^2 \log p(y_i, \beta | x_i)}{\partial \beta \partial \beta^\top} = -\Lambda(x_i^\top \beta) (1 - \Lambda(x_i^\top \beta)) x_i x_i^\top.$$

Thus, the information matrix is given by

$$\begin{aligned}I(\beta) &= -\mathbb{E} \left[ \frac{\partial^2 \log p(Y_i, \beta | X_i)}{\partial \beta \partial \beta^\top} \right] \\ &= \mathbb{E} [\Lambda(X_i^\top \beta) (1 - \Lambda(X_i^\top \beta)) X_i X_i^\top].\end{aligned}$$

Hence, for the logit model,  $\hat{\beta}_n^{ML} \rightarrow_p \beta$ , and

$$n^{1/2} (\hat{\beta}_n^{ML} - \beta) \rightarrow_d N \left( 0, (\mathbb{E} [\Lambda(X_i^\top \beta) (1 - \Lambda(X_i^\top \beta)) X_i X_i^\top])^{-1} \right).$$

The asymptotic variance-covariance matrix of  $\hat{\beta}_n^{ML}$  can be estimated consistently by

$$\left( n^{-1} \sum_{i=1}^n \Lambda(X_i^\top \hat{\beta}_n^{ML}) (1 - \Lambda(X_i^\top \hat{\beta}_n^{ML})) X_i X_i^\top \right)^{-1}.$$

## Marginal effects

In the linear regression model, the marginal effect of  $X_i$  on  $\mathbb{E}[Y_i | X_i]$  is  $\beta$ . In nonlinear binary choice models such as probit or logit,

$$\begin{aligned}\frac{\partial \mathbb{E}[Y_i | X_i]}{\partial X_i} &= \frac{\partial \Pr(Y_i = 1 | X_i)}{\partial X_i} \\ &= \frac{\partial F(X_i^\top \beta)}{\partial X_i} \\ &= f(X_i^\top \beta) \beta,\end{aligned}$$

and

$$\begin{aligned}\frac{\partial \Pr(Y_i = 0 | X_i)}{\partial X_i} &= \frac{\partial (1 - \Pr(Y_i = 1 | X_i))}{\partial X_i} \\ &= -f(X_i^\top \beta) \beta.\end{aligned}$$

Thus, in the case of the probit model, the marginal effect of  $X_i$  on  $\Pr(Y_i = 1 | X_i)$  is given by

$$\phi(X_i^\top \beta) \beta,$$

where  $\phi(u) = d\Phi(u)/du$  is the standard normal density. In the case of logit, equation (2) implies that  $\partial \Pr(Y_i = 1 | X_i) / \partial X_i$  is given by

$$\Lambda(X_i^\top \beta) (1 - \Lambda(X_i^\top \beta)) \beta.$$

The marginal effects can be estimated by replacing the unknown coefficients  $\beta$  with their ML estimators. For a fixed  $X_i = x$ ,

$$\frac{\partial \Pr(\widehat{Y_i = 1} | X_i = x)}{\partial X_i} = f(x^\top \widehat{\beta}_n^{ML}) \widehat{\beta}_n^{ML}.$$

By Slutsky's theorem and the consistency of  $\widehat{\beta}_n^{ML}$ ,

$$\begin{aligned} \frac{\partial \Pr(\widehat{Y_i = 1} | X_i)}{\partial X_i} &= f(x^\top \widehat{\beta}_n^{ML}) \widehat{\beta}_n^{ML} \\ &\rightarrow_p f(x^\top \beta) \beta \\ &= \frac{\partial \Pr(Y_i = 1 | x)}{\partial X_i}, \end{aligned}$$

provided that  $f$  is continuous.

The asymptotic distribution of  $f(x^\top \widehat{\beta}_n^{ML}) \widehat{\beta}_n^{ML}$  can be obtained using the delta method:

$$\begin{aligned} n^{1/2} \left( f(x^\top \widehat{\beta}_n^{ML}) \widehat{\beta}_n^{ML} - f(x^\top \beta) \beta \right) &\rightarrow_d N \left( 0, \frac{\partial (f(x^\top \beta) \beta)}{\partial \beta^\top} I^{-1}(\beta) \frac{\partial (f(x^\top \beta) \beta^\top)}{\partial \beta} \right) \\ &= N \left( 0, (f'(x^\top \beta) \beta x^\top + f(x^\top \beta) I_k) I^{-1}(\beta) (f'(x^\top \beta) \beta x^\top + f(x^\top \beta) I_k)^\top \right). \end{aligned}$$