

## ECON 452\* -- Introduction to NOTES 13 and 14

Introduction to Binary Dependent Variables Models**1. The Linear Probability Model**

The *linear probability model* – or *LPM* – looks exactly like a standard linear regression model, except that **the regressand  $Y_i$  is a binary variable** that takes only two discrete values, 0 and 1.

□ The **population regression equation (PRE)** of the LPM is:

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + u_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + u_i \quad (i = 1, \dots, N) \quad (1)$$

where

$$Y_i = \{0, 1\}.$$

□ **Interpretation of the Regression Coefficients in the LPM**

**Question:** How should the slope coefficients  $\beta_j$  ( $j = 1, \dots, k$ ) be interpreted when  $Y_i$  is a binary dependent variable?

**Answer:**

Under the zero conditional mean error assumption – Assumption A2:

$$E(u_i | \mathbf{x}_i^T) = 0 \quad (A2)$$

Implication of A2:

$$E(u_i | \mathbf{x}_i^T) = 0 \Rightarrow E(Y_i | \mathbf{x}_i^T) = \mathbf{x}_i^T \boldsymbol{\beta} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$$

**Key Point:** When  $Y_i$  is a binary dependent variable,

$$\Pr(Y_i = 1 | x_i^T) = E(Y_i | x_i^T) = x_i^T \beta \quad (2)$$

1.  $\Pr(Y_i = 1 | x_i^T) = x_i^T \beta$  is called generically the **response probability** or probability of “success”.
2.  $\Pr(Y_i = 1 | x_i^T) + \Pr(Y_i = 0 | x_i^T) = 1$  for all  $x_i^T$ .
3.  $\Pr(Y_i = 1 | x_i^T) + \Pr(Y_i = 0 | x_i^T) = 1$  implies that

$$\begin{aligned} \Pr(Y_i = 0 | x_i^T) &= 1 - \Pr(Y_i = 1 | x_i^T) \\ &= 1 - E(Y_i | x_i^T) \\ &= 1 - x_i^T \beta \end{aligned} \quad (3)$$

### Interpretation of Slope Coefficients $\beta_j$

Let  $X_j$  change by  $\Delta X_j$ ; hold values of all other regressors constant.

The resulting change in  $\Pr(Y_i = 1 | x_i^T) = x_i^T \beta$  is:

$$\Delta \Pr(Y_i = 1 | x_i^T) = \Delta E(Y_i | x_i^T) = \beta_j \Delta X_j.$$

Therefore,

$$\beta_j = \left( \frac{\Delta \Pr(Y_i = 1 | x_i^T)}{\Delta X_j} \right)_{\Delta X_g = 0, \forall g \neq j}$$

= the change in the probability that  $Y_i = 1$  associated with a one-unit increase in  $X_j$ , holding constant the values of all other explanatory variables

### □ OLS Estimation of the LPM

- OLS estimation of the PRE  $Y_i = x_i^T \beta + u_i$  yields the **OLS sample regression equation (OLS SRE)**:

$$Y_i = x_i^T \hat{\beta} + \hat{u}_i = \hat{Y}_i + \hat{u}_i \quad (4)$$

where

$$\hat{Y}_i = x_i^T \hat{\beta} = \text{the estimated or predicted value of } Y_i$$

$$\hat{u}_i = Y_i - \hat{Y}_i = Y_i - x_i^T \hat{\beta} = \text{the OLS residual for the } i\text{-th observation}$$

$$\hat{\beta} = (X^T X)^{-1} X^T y = \text{the OLS estimator of regression coefficient vector } \beta$$

### • Properties of OLS estimator $\hat{\beta}$

$$\hat{\beta} \text{ is } \textit{unbiased}: \quad E(\hat{\beta}) = \beta$$

$$\hat{\beta} \text{ is } \textit{consistent}: \quad \text{plim}(\hat{\beta}) = \beta$$

$$\hat{\beta} \text{ is } \textit{inefficient}: \quad \hat{\beta} \text{ is not the minimum variance estimator of } \beta$$

### □ Two Major Defects of OLS Estimation of BDV Models

- **Defect 1: Predictions outside the unit interval [0, 1]**

$$\hat{Y}_i = x_i^T \hat{\beta} \text{ is an estimator of } \Pr(Y_i = 1 | x_i^T) = x_i^T \beta$$

But it is nonetheless possible for the values of  $\hat{Y}_i = x_i^T \hat{\beta}$  to lie outside the unit interval [0, 1]: i.e., for  $\hat{Y}_i < 0$  and  $\hat{Y}_i > 1$ .

- ***Defect 2:*** The error terms  $u_i$  are heteroskedastic – i.e., have nonconstant variances.

- ***Result:*** It can be shown that

$$\begin{aligned}\sigma_i^2 &\equiv \text{Var}(u_i | x_i^T) = \text{Pr}(Y_i = 1 | x_i^T) [1 - \text{Pr}(Y_i = 1 | x_i^T)] \\ &= x_i^T \beta (1 - x_i^T \beta)\end{aligned}\tag{5}$$

$\neq$  a positive constant for all  $i = 1, \dots, N$

- ***Implications:***

1. OLS estimators of  $\text{Var}(\hat{\beta}_j)$  are ***biased*** and ***inconsistent***.
2. **t-tests** and **F-tests** based on  $\hat{V}_{\text{OLS}} = \hat{\sigma}^2 (\mathbf{X}^T \mathbf{X})^{-1}$  are ***invalid***.

- ***One Remedy:*** Use **heteroskedasticity-consistent estimators** of  $V_{\hat{\beta}} = V(\hat{\beta}_{\text{OLS}})$ .

Use either

$$\hat{V}_{\text{HC}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \hat{V} \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1}\tag{6.1}$$

or

$$\hat{V}_{\text{HCl}} = \frac{N}{N-K} \hat{V}_{\text{HC}} = \frac{N}{N-K} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \hat{V} \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1}\tag{7.1}$$

where  $\hat{V}$  is the  $N \times N$  diagonal matrix

$$\hat{V} = \text{diag}(\hat{u}_1^2 \quad \hat{u}_2^2 \quad \hat{u}_3^2 \quad \dots \quad \hat{u}_N^2) = \begin{bmatrix} \hat{u}_1^2 & 0 & 0 & \dots & 0 \\ 0 & \hat{u}_2^2 & 0 & \dots & 0 \\ 0 & 0 & \hat{u}_3^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \hat{u}_N^2 \end{bmatrix}$$

$\hat{u}_i^2 = (Y_i - x_i^T \hat{\beta})^2 =$  the squared unrestricted OLS residuals for  $i = 1, \dots, N$

---

Then **perform hypothesis tests on the coefficient vector  $\beta$**  using either of the following *heteroskedasticity-consistent* Wald F-statistics:

$$H_0: R\beta = r \Leftrightarrow R\beta - r = \underline{0}$$

$$H_1: R\beta \neq r \Leftrightarrow R\beta - r \neq \underline{0}$$

$$F_{HC} = \frac{(R\hat{\beta} - r)^T (R\hat{V}_{HC} R^T)^{-1} (R\hat{\beta} - r)}{q} \sim F[q, N - K] \text{ under } H_0 \quad (6.2)$$

$$F_{HCI} = \frac{(R\hat{\beta} - r)^T (R\hat{V}_{HCI} R^T)^{-1} (R\hat{\beta} - r)}{q} \sim F[q, N - K] \text{ under } H_0 \quad (7.2)$$

---

## 2. The Basics of Maximum Likelihood Estimation

This section introduces the basic principles of maximum likelihood estimation.

- **ML estimation** involves **joint estimation of all the unknown parameters** of a statistical model.

Let  $\theta$  denote **the vector of all unknown parameters** of the statistical model in question.

For example, for the linear probability model  $Y_i = x_i^T \beta + u_i$  where  $Y_i = \{0, 1\}$ , the parameter vector  $\theta = \beta$ , the  $K \times 1$  vector of regression coefficients.

ML estimation therefore requires that the model in question be completely specified. Complete specification of the model includes specifying the specific form of the probability distribution of the model's random variables.

- Derivation and computation of the ML estimator the parameter vector  $\theta$  consists of three main steps:

**Step 1: Formulation of the joint probability density function (pdf) and sample likelihood function** of the statistical model.

**Step 2: Formulation of the sample log-likelihood function** of the statistical model.

**Step 3: Maximization of the sample log-likelihood function** with respect to the unknown parameters in the vector  $\theta$ .

## STEP 1: Formulation of the Sample Likelihood Function

- **Assumption A4 of independent random sampling** implies that the *joint pdf* of all N sample values of  $Y_i$  is simply the **product** of the **pdf's of the individual  $Y_i$  values**.
- Under the assumption of *independent random sampling*, the *joint pdf of the N sample values*  $\{Y_1, Y_2, \dots, Y_N\}$  can be written as

$$f(y; \theta) = f(Y_1; \theta) \cdot f(Y_2; \theta) \cdots f(Y_N; \theta) = \prod_{i=1}^N f(Y_i; \theta). \quad (8)$$

*Note:* The joint pdf  $f(y; \theta)$  is a function of the N sample values of Y,  $\{Y_i : i = 1, \dots, N\}$ ; the parameter vector  $\theta$  is assumed to be known.

- Define the *likelihood function of the sample data*  $\{Y_i : i = 1, \dots, N\}$  as

$$L(\theta; y) = L(\theta; Y_1, Y_2, \dots, Y_N) = \prod_{i=1}^N f(Y_i; \theta) \quad (9)$$

where the sample likelihood function  $L(\theta; y)$  treats the parameters in the vector  $\theta$  as the unknowns and the sample values  $(Y_1, Y_2, \dots, Y_N)$  of the random variable Y as the knowns.

- **The ML Estimator of  $\theta$**  is that value of the parameter vector  $\theta$  which maximizes the sample likelihood function (9):

$$\hat{\theta}_{ML} = \max_{\theta} L(\theta; y) = \max_{\theta} \prod_{i=1}^N f(Y_i; \theta) \quad (10)$$

---

## STEP 2: Formulation of the Sample Log-Likelihood Function

- **Computation of  $\hat{\theta}_{ML}$** : It is often easier to maximize the natural logarithm of the sample likelihood function  $L(\theta; y)$  than it is to maximize  $L(\theta; y)$  directly.
- The *sample log-likelihood function* is obtained by simply taking the natural logarithm of the sample likelihood function  $L(\theta; y)$  in (9):

$$L(\theta; y) = L(\theta; Y_1, Y_1, \dots, Y_N) = \prod_{i=1}^N f(Y_i; \theta) \quad (9)$$

The **sample log-likelihood function** is therefore:

$$\ln L(\theta; y) = \ln L(\theta; Y_1, Y_1, \dots, Y_N) = \sum_{i=1}^N \ln f(Y_i; \theta) \quad (11)$$

*Note:*

1. Because  $0 < L(\theta; y) < 1$ ,  $\ln L(\theta; y) < 0$ .
2. The **sample log-likelihood function** is interpreted as **a function of the parameters  $\theta$**  for given sample values  $y = (Y_1 \ Y_2 \ Y_3 \ \dots \ Y_N)$  of the observable random variable  $Y$ .

### STEP 3: Maximization of the Sample Log-Likelihood Function

- The ML estimator of  $\theta$  is that value of the parameter vector  $\theta$  which *maximizes the sample log-likelihood function (9)*:

$$\hat{\theta}_{ML} = \max_{\theta} \ln L(\theta; y) = \max_{\theta} \sum_{i=1}^N \ln f(Y_i; \theta) \quad (12)$$

- **Equivalence of maximizing the likelihood and log-likelihood functions.**

Since the sample *log-likelihood function*  $\ln L(\theta; y)$  is a *positive monotonic transformation* of the sample *likelihood function*  $L(\theta; y)$ , that value of the parameter vector  $\theta$  which maximizes  $L(\theta; y)$  also maximizes  $\ln L(\theta; y)$ :

$$\hat{\theta}_{ML} = \max_{\theta} L(\theta; y) = \max_{\theta} \ln L(\theta; y). \quad (13)$$

The reason is that, for any individual parameter  $\theta_j$ ,

$$\frac{\partial \ln L(\theta; y)}{\partial \theta_j} = \frac{1}{L} \frac{\partial L(\theta; y)}{\partial \theta_j} = \frac{\partial L(\theta; y) / \partial \theta_j}{L} \quad \text{where } L > 0. \quad (14)$$

Thus,

$$\begin{aligned} \frac{\partial L(\theta; y)}{\partial \theta_j} > 0 & \Rightarrow \frac{\partial \ln L(\theta; y)}{\partial \theta_j} > 0; \\ \frac{\partial L(\theta; y)}{\partial \theta_j} = 0 & \Rightarrow \frac{\partial \ln L(\theta; y)}{\partial \theta_j} = 0; \\ \frac{\partial L(\theta; y)}{\partial \theta_j} < 0 & \Rightarrow \frac{\partial \ln L(\theta; y)}{\partial \theta_j} < 0. \end{aligned}$$

**□ Statistical Properties of the ML Parameter Estimators**

All **ML estimators** exhibit **three large sample properties**: consistency, asymptotic efficiency, and asymptotic normality.

1. **Consistency**: the probability limit of  $\hat{\theta}_{ML} = \theta$ ;  $\text{plim}(\hat{\theta}_{ML}) = \theta$ .
2. **Asymptotic efficiency**:  $\text{Asy Var}(\hat{\theta}_{j,ML}) \leq \text{Asy Var}(\tilde{\theta}_j)$ , the asymptotic variance of any other consistent estimator  $\tilde{\theta}_j$  of  $\theta_j$ .
3. **Asymptotic normality**:  $\hat{\theta}_{ML} \stackrel{a}{\sim} N[\theta, \text{AsyV}(\hat{\theta})]$ .