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:

$$Y_i = x_i^T\beta + u_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + u_i \qquad (i = 1, \ldots, N) \tag{1}$$

where

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

❑ Interpretation of the Regression Coefficients in the LPM

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

Answer:

Under the zero conditional mean error assumption – Assumption A2:

$$E\left(u_i \middle| x_i^T\right) = 0 \qquad \text{(A2)}$$

Implication of A2:

$$E\left(u_i \middle| x_i^T\right) = 0 \quad \Rightarrow \quad E\left(Y_i \middle| x_i^T\right) = x_i^T\beta = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$$

Key Point: When Y_i is a binary dependent variable,

$$\Pr\left(Y_i = 1 \mid x_i^T\right) = E\left(Y_i \mid x_i^T\right) = x_i^T\beta \tag{2}$$

1. $\Pr\left(Y_i = 1 \mid x_i^T\right) = x_i^T\beta$ is called generically the ***response probability*** or probability of "success".

2. $\Pr\left(Y_i = 1 \mid x_i^T\right) + \Pr\left(Y_i = 0 \mid x_i^T\right) = 1$ for all x_i^T.

3. $\Pr\left(Y_i = 1 \mid x_i^T\right) + \Pr\left(Y_i = 0 \mid x_i^T\right) = 1$ implies that

$$\begin{aligned} \Pr\left(Y_i = 0 \mid x_i^T\right) &= 1 - \Pr\left(Y_i = 1 \mid x_i^T\right) \\ &= 1 - E\left(Y_i \mid x_i^T\right) \\ &= 1 - x_i^T\beta \end{aligned} \tag{3}$$

Interpretation of Slope Coefficients β_j in the Linear Probability Model

Let X_j change by ΔX_j ; hold values of all other regressors constant.

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

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

Therefore,

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

= 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)**:

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

where

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

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

$\hat{\beta} = (X^TX)^{-1}X^Ty$ = the OLS estimator of regression coefficient vector β

- **Properties of OLS estimator $\hat{\beta}$**

$\hat{\beta}$ is ***unbiased***: $E(\hat{\beta}) = \beta$

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

$\hat{\beta}$ is ***inefficient***: $\hat{\beta}$ is ***not*** the **minimum variance estimator of β**

❑ Two Major Defects of OLS Estimation of BDV Models

- ***<u>Defect 1</u>:* Predictions outside the unit interval [0, 1]**

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

  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 $\hat{Y}_i > 1$.

- ***<u>Defect 2</u>:* The error terms u_i are heteroskedastic – i.e., have nonconstant variances.**

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

$$\sigma_i^2 \equiv \operatorname{Var}\left(u_i \mid x_i^T\right) = \Pr(Y_i = 1 \mid x_i^T)\left[1 - \Pr(Y_i = 1 \mid x_i^T)\right]$$

$$= x_i^T\beta\,(1 - x_i^T\beta) \qquad (5)$$

$$\neq \text{ a positive constant for all } i = 1, \ldots, N$$

- ***Implications:***

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

- ***One Remedy for Defect 2:*** Use **heteroskedasticity-consistent estimators** of $V_{\hat{\beta}} = V(\hat{\beta}_{OLS})$.

Use either

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

or

$$\hat{V}_{HC1} = \frac{N}{N-K}\hat{V}_{HC} = \frac{N}{N-K}\left(X^T X\right)^{-1} X^T \hat{V} X \left(X^T X\right)^{-1} \tag{7.1}$$

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

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

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

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

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

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

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

$$F_{HC1} = \frac{\left(R\hat{\beta} - r\right)^T \left(R\,\hat{V}_{HC1}\,R^T\right)^{-1} \left(R\hat{\beta} - r\right)}{q} \overset{a}{\sim} F[q, N-K] \text{ under } H_0 \tag{7.2}$$

How to do this in *Stata*: How to compute ***heteroskedasticity-consistent*** **Wald F-statistics** when estimating a linear probability model by OLS.

- Consider the following linear regression equation (which could have a binary regressand Y_i):

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + u_i$$

- We want to perform the following test of two coefficient exclusion restrictions on this model:

H_0: $\beta_3 = 0$ *and* $\beta_4 = 0$
H_1: $\beta_3 \neq 0$ *and/or* $\beta_4 \neq 0$

- The following two *Stata* commands will estimate the above model by OLS and perform a ***heteroskedasticity-consistent*** **Wald F-test** of the two coefficient restrictions specified by the null hypothesis H_0:

```
regress y x1 x2 x3 x4, robust
test x3 x4
```

  - The ***regress*** command with the ***robust*** option computes all coefficient standard errors, t-ratios and confidence intervals using the ***adjusted*** **HC covariance estimator** $\hat{V}_{HC1}$:

$$\hat{V}_{HC1} = \frac{N}{N-K}\hat{V}_{HC} = \frac{N}{N-K}\left(X^T X\right)^{-1} X^T \hat{V} X \left(X^T X\right)^{-1} \qquad (7.1)$$

- The ***test*** command computes the ***heteroskedasticity-consistent*** **(or *heteroskedasticity-robust*) Wald F-statistic** F_{HC1} for the two linear coefficient restrictions specified by the null hypothesis H_0:

$$F_{HC1} = \frac{\left(R\hat{\beta} - r\right)^T \left(R\,\hat{V}_{HC1}\,R^T\right)^{-1}\left(R\hat{\beta} - r\right)}{q} \overset{a}{\sim} F[q, N-K] \text{ under } H_0 \tag{7.2}$$

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 θ 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 $\boldsymbol{\theta} = \boldsymbol{\beta}$, the K×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 θ 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 θ.

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 $\{Y_1, Y_2, \ldots, Y_N\}$** can be written as

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

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

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

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

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

- **<u>The ML Estimator of θ</u>** is that value of the parameter vector θ which maximizes the sample likelihood function (9):

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

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):

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

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

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

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* θ** for given sample values $y = \left(Y_1 \;\; Y_2 \;\; Y_3 \;\; \cdots \;\; Y_N\right)$ of the observable random variable Y.

STEP 3: Maximization of the Sample Log-Likelihood Function

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

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

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

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

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

The reason is that, for any individual parameter θ_j,

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

Thus,

$$\frac{\partial L(\theta; y)}{\partial \theta_j} > 0 \qquad \Rightarrow \qquad \frac{\partial \ln L(\theta; y)}{\partial \theta_j} > 0;$$

$$\frac{\partial L(\theta; y)}{\partial \theta_j} = 0 \qquad \Rightarrow \qquad \frac{\partial \ln L(\theta; y)}{\partial \theta_j} = 0;$$

$$\frac{\partial L(\theta; y)}{\partial \theta_j} < 0 \qquad \Rightarrow \qquad \frac{\partial \ln L(\theta; y)}{\partial \theta_j} < 0.$$

❑ 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***: Asy Var$(\hat{\theta}_{j,ML}) \le$ Asy Var$(\tilde{\theta}_j)$, the asymptotic variance of any other consistent estimator $\tilde{\theta}_j$ of θ_j.

3. ***Asymptotic normality***: $\hat{\theta}_{ML} \overset{a}{\sim} N\left[\theta, \text{AsyV}(\hat{\theta})\right]$.