Lecture 17

Discrete choice models (or qualitative response models)

Suppose y_t is the probability that a household will purchase a car in a given year and x_t is the household's income.

Model:

$$y_t = \alpha + \beta x_t + u_t$$

For this regression, all we observe is whether the household purchased the car or not, that is, $y_t \in \{0, 1\}$.

Consequences:

1. The discreteness makes the errors non-normal;
2. This introduces heteroskedasticity.

Demonstration:

Let $P_t = \Pr(y_t = 1)$. Then:

$$P_t = \Pr(u_t = 1 - \alpha - \beta x_t)$$
$$\therefore 1 - P_t = \Pr(y_t = 0) = \Pr(u_t = -\alpha - \beta x_t)$$

- For a given x_t, u_t can only take one of two values. For that reason, u_t cannot be normal; it rather follows a binomial distribution.
- We know that $\mathrm{E}(u_t) = 0$.
  - $E(u_t) = P_t(1 - \alpha - \beta x_t) + (1 - P_t)(-\alpha - \beta x_t) := 0$.
  - From that, $P_t = \alpha + \beta x_t$.
- We know that $\sigma_u^2 = \mathrm{E}(u_t^2 - \bar{u}) = \mathrm{E}(u_t^2)$.
  - $\mathrm{E}(u_t^2) = P_t(1 - \alpha - \beta x_t)^2 + (1 - P_t)(\alpha - \beta x_t)^2$.
  - From that, $\sigma_u^2 = P_t(1 - P_t)^2 + (1 - P_t)(-P_t)^2$.
  - Then, $\sigma_u^2 = P_t(1 - P_t) = (\alpha + \beta x_t)(1 - \alpha - \beta x_t)$, and it means that the variance of the errors varies with x_t (heteroskedasticity).

For those reason, OLS of α and β would be *unbiased* and *consistent*, but *inefficient*.

- Due to the non-normality, the test statistics would be invalid.

Solution:

Step 1: Estimate using OLS. Save $\hat{\boldsymbol{y}}_{\boldsymbol{t}}$.

Step 2: Estimate the variance by $\hat{\sigma}_t^2 = \hat{y}_t(1 - \hat{y}_t)$.

Step 3: Divide dependent and independent variables by $\hat{\sigma}_t$.

- That will produce y_t^* and x_t^*.

Step 4: Regress y^* on $\dfrac{1}{\sigma_t}, x^*$. The estimate will be BLUE.

Possible problems:

- In step 2, there is no guarantee that $\hat{\sigma}_t^2 > 0$.

- In other terms, there is no guarantee $0 \leq \hat{y} \leq 1$.

The Probit model

Consider the model:

$$y_t^* = \alpha + \beta x_t + u_t$$

where x_t is observable, but y_t^* is not.

- An example: y_t^* is the difference between the wage and the reservation wage.

What one would actually observe would be:

$$y_t = \begin{cases} 1, & \text{if } \alpha + \beta x_t + u_t > 0 \quad (y_t^* > 0) \\ 0, & \text{if } \alpha + \beta x_t + u_t < 0 \quad (y_t^* < 0) \end{cases}$$

Let $F(z)$ be the cumulative normal distribution.

- Then, $F(z) = \Pr(Z \leq z)$.

Thus:

$$\begin{aligned} \Pr(y_t = 1) &= \Pr(u_t > -\alpha - \beta x_t) \\ &= 1 - F\left(\frac{\alpha - \beta x_t}{\sigma_t}\right) \\ \Pr(y_t = 0) &= \Pr(u_t \leq -\alpha - \beta x_t) \\ &= F\left(\frac{\alpha - \beta x_t}{\sigma_t}\right) \end{aligned}$$

Likelihood function:

$$L = \prod_{y_t=0} F\left(\frac{-\alpha - \beta x_t}{\sigma_t}\right) \prod_{y_t=1} \left[1 - F\left(\frac{-\alpha - \beta x_t}{\sigma_t}\right)\right]$$

Maximum Likelihood estimation: pick α and β that maximize L.

- It is always consistent and always efficient.
- It reaches the Cramér-Rao bound.

The Logit model

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta x + u$$

It works best when $0 < y < 1$.

For this model, $p_t = \dfrac{1}{1 + e^{-(\alpha + \beta x_t)}}$.

Thus:

$$\begin{aligned} \beta x &\to \infty, \qquad & p \to 1 \\ \beta x &\to -\infty, \qquad & p \to 0 \end{aligned}$$

Probit model: errors are assumed normal ($F(z)$).

Logit model: errors are assumed to follow a logistic distribution $f(x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}$.

Estimation

- If $0 < y < 1$, use OLS on $y^* = \ln \frac{y}{1-y}$.
- If $y \in \{0, 1\}$, use MLE.