BIẾN PHỤ THUỘC ĐỊNH TÍNH

GV : Đinh Công Khải – Chương trình Fulbright
Môn: Các Phương Pháp Định Lượng – MPP3

Các tình huống ứng dụng

- Quyết định tham gia vào lực lượng lao động.
- Cả 2 vợ chồng đều tham gia vào lực lượng lao động hay chỉ có một người tham gia.
- Quyết định bầu cho đảng nào.
- Gia đình có sở hữu nhà hay không.
- Công ty có công bố quyết định phân chia cổ tức hay không.

Sự khác biệt giữa mô hình hồi qui với Y là biến định lượng và Y là biến định tính

- Nếu Y là biến định lượng mục tiêu của chúng ta là ước lượng

  $E(Y_i|X_{1i}, X_{2i}, X_{3i},\ldots, X_{Ki})$

- Nếu Y là biến định tính mục tiêu của chúng ta là ước lượng xác xuất một điều gì đó sẽ xảy ra ➔ Mô hình xác suất (probability models).
- *Các vấn đề kinh tế lượng liên quan đến mô hình hồi qui với biến Y định tính?*
  - Có thể sử dụng phương pháp OLS thông thường để ước lượng không?
  - Có thể sử dụng phương thức kiểm định truyền thống không?
  - R^2 có phải là tiêu chí tốt để đánh giá độ thích hợp của mô hình không?

Mô hình xác suất tuyến tính (Linear Probability Models – LPM)

- $\mathbf{Y_i = \beta_1 + \beta_2 X_i + u_i}$ (1)

  X = thu nhập của hộ gia đình;

  Y = 1 nếu hộ gia đình sở hữu nhà, và 0 nếu không sở hữu nhà
- $\mathbf{E(Y_i |X_i) = Pr(Y_i =1|X_i)}$

➔ Xác xuất có điều kiện rằng sự kiện Y sẽ xảy ra với X_i cho trước

➔ Xác xuất để một hộ gia đình sở hữu một căn nhà với thu nhập là X_i.

- $\mathbf{E(Y_i |X_i) = Pr(Y_i =1|X_i) = \beta_1 + \beta_2 X_i}$ (với giả thiết $E(u_i) = 0$)

Mô hình xác suất tuyến tính

- Gọi P_i là xác xuất để $Y_i = 1$ và $(1-P_i)$ là xác xuất để $Y_i = 0$
- → Y_i có phân phối xác xuất Bernoulli
- → $E(Y_i) = 0*(1 - P_i) + 1*P_i = P_i$.
- **$E(Y_i |X_i) = Pr(Y_i =1|X_i) = \beta_1 + \beta_2 X_i = P_i$**
- → $0 \leq E(Y_i |X_i) \leq 1$

Các vấn đề kinh tế lượng của mô hình LPM

1) Sai số ngẫu nhiên u_i không có phân phối chuẩn mà có phân phối Bernoulli

$u_i = Y_i - \beta_1 - \beta_2 X_i$

Y_i	u_i	Xác xuất
$Y_i =1$	$1- \beta_1 - \beta_2 X_i$	P_i
$Y_i =0$	$- \beta_1 - \beta_2 X_i$	$1- P_i$

- u_i không có phân phối chuẩn không phải là quá nghiêm trọng đối với ước lượng OLS vì *ước lượng OLS không bị thiên lệch*
- Với mẫu lớn ước lượng OLS sẽ có phân phối chuẩn.

Các vấn đề kinh tế lượng của mô hình LPM

2) Phương sai thay đổi

$var(u_i) = P_i (1 - P_i) \neq const$ $[P_i = \beta_1 + \beta_2 X_i]$

- Phương pháp khắc phục

$$\frac{Y_i}{\sqrt{w_i}} = \frac{\beta_1}{\sqrt{w_i}} + \frac{\beta_2 X_i}{\sqrt{w_i}} + \frac{u_i}{\sqrt{w_i}} \quad (2)$$

trong đó

$$\sqrt{w_i} = \sqrt{E(Y_i \mid X_i)*[1 - E(Y_i \mid X_i)]} = \sqrt{P_i(1 - P_i)}$$

Các vấn đề kinh tế lượng của mô hình LPM

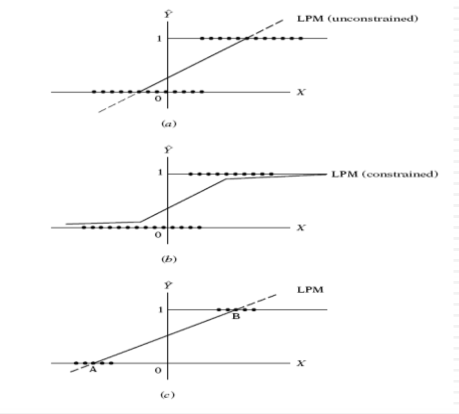
- **Quy trình ước lượng**
- **Bước 1**: Hồi qui (1) bằng OLS, tính $\hat{Y}_i$ [ước lượng của $E(Y_i|X_i)$] và $\hat{Y}_i(1-\hat{Y}_i)$ [ước lượng của w_i].
- **Bước 2**: Dùng w_i để chuyển (1) thành (2), sau đó ước lượng (2) theo OLS.

3) $0 \leq E(Y_i |X_i) \leq 1$ có thể không thỏa

- $E(Y_i |X_i) < 0 \rightarrow E(Y_i |X_i) = 0$;
- $E(Y_i |X_i) > 1 \rightarrow E(Y_i |X_i) = 1$;

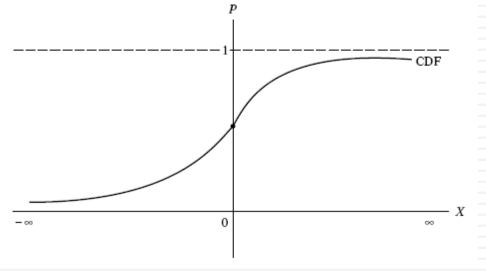
Các vấn đề kinh tế lượng của mô hình LPM

4) R^2 là phải là thước đo độ thích hợp của mô hình?



Hàm phân phối tích lũy (cumulative distribution function - CDF)

- Cần một mô hình thích hợp hơn LPM với các đặc tính sau đây
- P_i và X_i quan hệ phi tuyến tính;
- Khi X_i tăng $E(Y_i| X_i)$ cũng tăng nhưng nằm trong dãy [0;1]



Hàm Logit (Logistic)

- Xây dựng mô hình

$$P_i = E(Y=1|X_i) = \frac{1}{1+e^{-(\beta_1+\beta_2 X_i)}}$$

$$P_i = \frac{1}{1+e^{-Z}} = \frac{e^Z}{1+e^Z}$$

$$Z = \beta_1 + \beta_2 X_i$$

- P_i nằm trong [0;1]; và P_i quan hệ phi tuyến tính với X_i

Hàm Logit (Logistic)

- Tuyến tính hóa mô hình

$$\frac{P_i}{1-P_i} = \frac{1+e^{Z_i}}{1+e^{-Z_i}} = e^{Z_i}$$

$$L_i = \ln(\frac{P_i}{1-P_i}) = Z_i = \beta_1 + \beta_2 X_i$$ (mô hình logit)

Hàm Logit (Logistic)

- Mô hình hồi qui logit

$$L_i = \ln(\frac{P_i}{1-P_i}) = Z_i = \beta_1 + \beta_2 X_i + u_i$$

- ***Ước lượng với thông tin cá nhân***: *không thể dùng OLS*; sử dụng phương pháp maximum-likelihood
- ***Ước lượng với thông tin nhóm (thông tin lặp lại)***

Hàm Logit (Logistic)

HYPOTHETICAL DATA ON X_i (INCOME), N_i (NUMBER OF FAMILIES AT INCOME X_i), AND n_i (NUMBER OF FAMILIES OWNING A HOUSE)

X (thousands of dollars)	N_i	n_i
6	40	8
8	50	12
10	60	18
13	80	28
15	100	45
20	70	36
25	65	39
30	50	33
35	40	30
40	25	20

$$\hat{P}_i = \frac{n_i}{N_i}$$

$$\hat{L}_i = \ln(\frac{\hat{P}_i}{1-\hat{P}_i}) = \hat{\beta}_1 + \hat{\beta}_2 X_i$$

Hàm Logit (Logistic)

- Chúng ta xem xét sai số ngẫu nhiên trước khi hồi qui theo OLS

$$u_i \sim N(0, \frac{1}{N_i P_i (1 - P_i)})$$

- Khắc phục bằng cách hồi qui theo

$$\sqrt{w_i} L_i = \beta_1 \sqrt{w_i} + \beta_2 \sqrt{w_i} X_i + \sqrt{w_i} u_i$$
$$w_i = N_i \hat{P}_i (1 - \hat{P}_i)$$
$$L^*_i = \beta_1 \sqrt{w_i} + \beta_2 X^*_i + v_i$$

Hàm Logit (Logistic)

- **Đánh giá và kiểm định ý nghĩa thống kê mô hình Logit (Probit) khi ước lượng với những thông tin cá nhân**
- Đánh giá độ thích hợp của mô hình

  Psedo R^2 = Mc Fadden R^2= 1 - (LLF_{UR} - LLF_R)
- Kiểm tra ý nghĩa thống kê các hệ số: sử dụng thống kê z thay vì t-student
- Kiểm định ý nghĩa chung của toàn bộ mô hình: sử dụng thống kê chi-square

  LR (Likelihood ratio) = 2(LLF_{UR} - LLF_R)

Hàm Probit

- Mô hình probit sử dụng hàm CDF chuẩn hóa
- Ví dụ về thu nhập và sở hữu nhà, hộ gia đình sẽ sở hữu nhà hay không tùy thuộc vào chỉ số (năng lực) thỏa dụng I_i (utility index).

  $I_i = \beta_1 + \beta_2 X_i$
- Nếu $I_i < I^*$ thì xác xuất mua nhà bằng 0 và nếu $I_i > I^*$ thì xác xuất mua nhà bằng 1.
- I_i và I* không quan sát được, nhưng chúng có phân phối chuẩn

Hàm Probit

- Dựa vào giả thiết phân phối chuẩn

$$P_i = P(Y=1 | X) = P(I^* \le I_i) = P(Z_i \le \beta_1 + \beta_2 X_i) = F(\beta_1 + \beta_2 X_i)$$

F là hàm mật độ tích lũy thường được chuẩn hóa (standardized normal CDF)

$$F(I_i) = \frac{1}{2\pi} \int_{-\infty}^{I_i} e^{-z^2/2} dz$$

$$I_i = F^{-1}(I_i) = F^{-1}(P_i) = \beta_1 + \beta_2 X_i$$

- Tác động biên

$$\frac{dP_i}{dX_i} = \frac{\partial F(\beta_1 + \beta_2 X_i)}{\partial X} = f(\beta_1 + \beta_2 X_i) * \beta_2$$