

Mô hình với Biến Phụ thuộc bị Giới hạn (Models with Limited Dependent Variables)

Lê Việt Phú
Chương trình Giảng dạy Kinh tế Fulbright

Ngày 21 tháng 11 năm 2015

Table of contents

- Thế nào là biến phụ thuộc không bị giới hạn và bị giới hạn
 - Một số mô hình sử dụng biến phụ thuộc bị giới hạn
 - Sử dụng hồi quy tuyến tính đối với biến phụ thuộc bị giới hạn
 - Phương pháp tối đa hoá xác suất - MLE
- Ước lượng mô hình hồi quy Logit và Probit
- Thực hành trên STATA

Thế nào là biến phụ thuộc không bị giới hạn và bị giới hạn

- ▶ Các loại biến phụ thuộc trong mô hình hồi quy:
 - ▶ Liên tục và rời rạc: tăng trưởng GDP là liên tục, có thể có con số bất kỳ, ví dụ 6.1025%; số lần đi học muộn là rời rạc, ví dụ đi muộn 0, 1, 2 lần.
 - ▶ Không bị giới hạn và bị giới hạn: lợi nhuận của công ty là không giới hạn (lỗ thì nhận giá trị âm, lãi là dương); số nhân viên là bị giới hạn (bị chặn dưới, ít nhất 1 nhân viên trong một công ty).
 - ▶ Biến phụ thuộc định tính và định lượng: có hút thuốc lá hay không là biến định tính; hút bao nhiêu điếu thuốc một ngày là định lượng và bị giới hạn (ít nhất là một điếu).
- ▶ Hầu hết các biến số kinh tế đều bị giới hạn.
- ▶ Sử dụng hồi quy tuyến tính đối với dữ liệu bị giới hạn thì kết quả có thể bị sai lệch, hoặc khó giải thích ý nghĩa về mặt kinh tế.

Một số mô hình sử dụng biến phụ thuộc bị giới hạn (1)

- ▶ Mô hình xác suất xảy ra một sự kiện hay một biến cố nào đó. Ví dụ đối tượng vị thành niên hút thuốc, đi học đại học, phụ nữ dân tộc thiểu số tham gia lao động chính thức. Biến phụ thuộc là có hoặc không (mã hoá 1 cho câu trả lời có, 0 cho câu trả lời không). Đối với biến phụ thuộc định tính thì không có cách xếp hạng câu trả lời (có/không) như biến phụ thuộc định lượng (nhiều/ít).
- ▶ Mô hình xác suất có thể là đa lựa chọn thay vì hai lựa chọn, ví dụ anh/chị đến trường bằng phương tiện gì: ô-tô, xe máy, xe đạp, đi bộ.

Một số mô hình sử dụng biến phụ thuộc bị giới hạn (2)

- ▶ Mô hình số lần xảy ra một sự kiện nào đó. Ví dụ số lần một học viên MPP đi học muộn, số con trong một gia đình, số sản phẩm bị hỏng trong một ngày, số lần đi khám bệnh một năm. Biến phụ thuộc sẽ có giá trị 0 và số nguyên dương (1, 2, 3...).
- ▶ Mô hình mô tả xếp hạng của một sự kiện, ví dụ cảm quan của anh/chị về một môn học có thể là quá khó/khó/trung bình/tương đối dễ/quá dễ.
- ▶ Mô hình với biến phụ thuộc bị chặn trên hoặc dưới. Ví dụ thu nhập chỉ có thể là 0 hoặc dương; số tiền một người đã làm từ thiện trong một năm tối thiểu là 0 hoặc dương; số giờ làm việc trong một tuần không thể quá $24 \times 7 = 168$ giờ.

Tên gọi mô hình sử dụng biến phụ thuộc có giới hạn

- ▶ Mô hình xác suất (Logit, Probit, Multinomial Logit)
- ▶ Mô hình số lần xảy ra sự kiện (Poisson)
- ▶ Mô hình với biến phụ thuộc bị chặn (Tobit, Censored, Truncated Regression)

Điều gì xảy ra nếu sử dụng công cụ OLS cùng các giả định của mô hình CLRM vào dữ liệu có biến phụ thuộc bị giới hạn?

Xem xét mô hình:

$$SMOKING_i = \beta_0 + \beta_1 * PRICE_i + u_i \quad (1)$$

trong đó $SMOKING_i$ là biến định tính cho hành vi hút thuốc lá của trẻ vị thành niên, nhận giá trị 1 nếu có hút thuốc và 0 nếu không. Biến giải thích là giá bán lẻ.

$$SMOKING_i = \begin{cases} 1 & \text{for smoker} \\ 0 & \text{for non-smoker} \end{cases}$$

- ▶ Trong mô hình thông thường, β_1 là thay đổi của biến phụ thuộc $SMOKING$ nếu biến giải thích $PRICE$ tăng một đơn vị.
- ▶ Đối với biến phụ thuộc nhị phân, $SMOKING_i$ chỉ nhận giá trị 0 hoặc 1, ý nghĩa của β_1 là gì?

Mô hình xác suất tuyến tính - Linear Probability Model (LPM)

- ▶ Với giả thiết kỳ vọng của biến dư bằng 0, $E[u|PRICE] = 0$:

$$E[SMOKING|PRICE] = \beta_0 + \beta_1 * PRICE \quad (2)$$

- ▶ Đồng thời:

$$\begin{aligned} E[SMOKING] &= \\ 1 * P(SMOKING = 1) + 0 * P(SMOKING = 0) &= \\ P(SMOKING = 1) \end{aligned}$$

$$\Rightarrow P(SMOKING = 1|PRICE) = \beta_0 + \beta_1 * PRICE$$

- ▶ Điều này có nghĩa là xác suất quan sát được một vị thành niên hút thuốc là mô hình tuyến tính của biến giải thích $PRICE$. Ví dụ $\beta = -0.1$, nếu giá bán tăng 1 đơn vị thì xác suất vị thành niên hút thuốc sẽ giảm 10%.

Mô hình xác suất tuyến tính (2)

Những vấn đề của mô hình xác suất tuyến tính:

- ▶ Nếu $\beta_1 = -0.1$ thì tăng giá bán thêm 20 đơn vị có làm cho xác suất hút thuốc giảm về 0 hay thậm chí âm không?
- ▶ Tác động biên của giá bán là cố định có hợp lý không? Ví dụ nếu giá thuốc lá tăng từ 10.000đ lên 20.000đ/bao có khác so với tăng từ 100.000đ lên 110.000đ/bao không?
- ▶ Giả định về phương sai không đổi trong mô hình CLRM, $Var(u_i) = \sigma^2$, bị vi phạm. Khi này:

$$Var(u_i|X_i) = P_i * (1 - P_i) , \text{ với}$$

$$P_i = \beta_0 + \beta_1 * PRICE_i$$

$\Rightarrow Var(u_i|PRICE_i) \in PRICE_i$, hay nói cách khác, phương sai của sai số thay đổi.¹

¹Biến phụ thuộc Y_i phân phối Bernoulli với xác suất $P_i = \beta_0 + \beta_1 * X_i$ nên u_i cũng phân phối Bernoulli với xác suất $P_{u_i} = 1 - \beta_0 - \beta_1 * X_i$. Phương sai của phân phối Bernoulli là $Var(u_i) = P_{u_i} * (1 - P_{u_i})$.

Phương pháp xác suất tối đa - Maximum Likelihood Estimation (MLE)

- ▶ Khắc phục các nhược điểm đã nêu trên, để (a) ước lượng xác suất luôn nằm trong khoảng $[0,1]$ với mọi giá trị của biến giải thích PRICE, và (b) tác động biên của biến giải thích không cố định, chúng ta cần cách tiếp cận mới không sử dụng phương pháp OLS.
- ▶ Giả định xác suất của việc hút thuốc được xác định bởi hàm phân phối xác suất tích lũy $G(\cdot)$:

$$P(SMOKING_i = 1 | PRICE) = G(\beta_0 + \beta_1 * PRICE_i) \quad (3)$$

Với hàm $G(\beta_0 + \beta_1 * PRICE_i)$ nhận giá trị nằm trong khoảng $[0,1]$ với mọi giá trị của biến giải thích PRICE.

- ▶ Hàm phân phối xác suất $G(\cdot)$ thường không biết trước, và phải dựa vào giả định hoặc các lý thuyết kinh tế.

Các hàm phân phối xác suất thông dụng (1)

- ▶ Nếu $G(\cdot)$ có phân phối tích lũy Logistic, khi đó ta có hồi quy "Logit":

$$G(z) = \frac{e^z}{1 + e^z}$$

với hàm mật độ phân phối Logistic $g(z) = G'(z) = \frac{e^z}{(1+e^z)^2}$

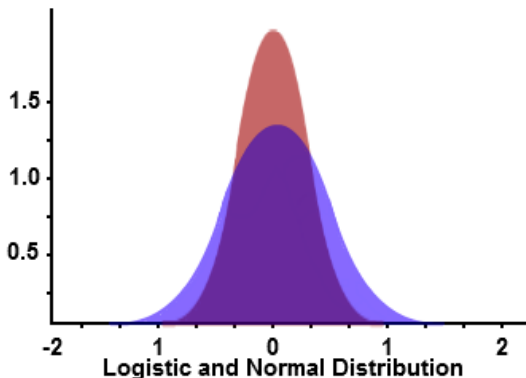
- ▶ Nếu $G(\cdot)$ có phân phối tích lũy chuẩn $\Rightarrow$ hồi quy Probit:

$$G(z) = \Phi(z) = \int_{-\infty}^z \phi(x) dx$$

với hàm mật độ phân phối chuẩn $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

Các hàm phân phối xác suất thông dụng (2)

Đồ thị Hàm Mật độ Phân phối Logit (Tím) và Chuẩn (Cam)



Hàm Logistic có mức độ phân tán cao hơn so với phân phối chuẩn.

Ước lượng mô hình hồi quy Logit và Probit (1)

- ▶ Khác với phương pháp bình phương phần dư tối thiểu OLS, mô hình hồi quy dựa trên hàm phân phối xác suất như Logit hay Probit dùng phương pháp xác suất tối đa (Maximum Likelihood Estimation-MLE).
- ▶ Hàm mục tiêu của phương pháp OLS là tối thiểu tổng bình phương phần dư của biến phụ thuộc, còn hàm mục tiêu của phương pháp MLE là tối đa xác suất quan sát được mẫu với thuộc tính cho trước.

Ước lượng mô hình hồi quy Logit và Probit (2)

- ▶ Xác suất quan sát được vị thành niên i có hút thuốc hay không có thể viết như sau:

$$P(SMOKING_i | PRICE_i) = [G(.)]^{SMOKING_i} \times [1 - G(.)]^{1 - SMOKING_i} \quad (4)$$

- ▶ Nếu $SMOKING_i = 1$ thì $P(SMOKING_i | PRICE_i) = G(.)$
- ▶ Nếu $SMOKING_i = 0$ thì $P(SMOKING_i | PRICE_i) = 1 - G(.)$
- ▶ $G(.)$ là hàm đơn điệu (do $G(.)$ là hàm phân phối xác suất tích lũy, $G(.)$ chỉ tăng hoặc giảm theo biến giải thích), có thể đơn giản hoá bằng cách chuyển đổi từ hàm tích (4) sang hàm logarithm :

$$\ell_i = \ln[P(.)] = SMOKING_i \times \ln[G(.)] + [1 - SMOKING_i] \times \ln[1 - G(.)] \quad (5)$$

Ước lượng mô hình hồi quy Logit và Probit (3)

- ▶ Nếu mẫu dữ liệu có N thành viên thì hàm xác suất tổng thể được tính bằng cách lấy tổng của xác suất của các quan sát:

$$\mathbf{L} = \sum_{i=1}^N \ell_i \quad (6)$$

và việc ước lượng theo phương pháp MLE được thực hiện bằng cách tối đa hoá tổng xác suất $\mathbf{L}$.

$$\text{Max } \mathbf{L} = \sum_i \left\{ S_i * \ln[G(.)] + [1 - S_i] * \ln[1 - G(.)] \right\} \Rightarrow \hat{\beta}_{MLE} \quad (7)$$

với S_i là biến phụ thuộc $SMOKING_i$, và $G(.)$ là hàm phân phối xác suất tích lũy $G(\beta_0 + \beta_1 * PRICE_i)$.

Ước lượng mô hình hồi quy Logit và Probit (4)

- ▶ Để tìm tham số β_0 và β_1 nhằm tối đa giá trị $\mathbf{L}$, sử dụng điều kiện tối ưu bậc nhất (first-order condition). Ví dụ với β_1 , sử dụng quy tắc chuỗi (chain-rule) khi lấy đạo hàm bậc nhất:
 - ▶ $G'(\beta_0 + \beta_1 * X_i) = g(.) * X_i$
 - ▶ $\frac{\partial \ln[G(.)]}{\partial \beta_1} = \frac{1}{G(.)} * g(.) * X_i$
- ▶ Do đó, điều kiện bậc nhất với β_1 :

$$\frac{\partial \mathbf{L}}{\partial \beta_1} = \sum_i \left\{ \frac{S_i}{G(.)} * g(.) * X_i - \frac{1 - S_i}{1 - G(.)} * g(.) * X_i \right\} = 0 \quad (8)$$

Ước lượng mô hình hồi quy Logit và Probit (5)

- ▶ Ví dụ đối với hồi quy Logit, $G(z) = \frac{e^z}{1+e^z}$ và $g(z) = \frac{e^z}{(1+e^z)^2}$.
Sau khi biến đổi, điều kiện bậc nhất đối với β_1 là:

$$\frac{\partial \mathbf{L}}{\partial \beta_1} = \sum_i S_i * X_i - \sum_i \frac{e^{\beta_0 + \beta_1 * X_i}}{1 + e^{\beta_0 + \beta_1 * X_i}} * X_i = 0 \quad (9)$$

và áp dụng đối với β_0 :

$$\frac{\partial \mathbf{L}}{\partial \beta_0} = \sum_i S_i - \sum_i \frac{e^{\beta_0 + \beta_1 * X_i}}{1 + e^{\beta_0 + \beta_1 * X_i}} = 0 \quad (10)$$

- ▶ Trong phương pháp MLE, do tính phi tuyến của điều kiện bậc nhất (9) và (10) nên không có công thức cụ thể để tính $\hat{\beta}_0$ và $\hat{\beta}_1$ như phương pháp OLS.
- ▶ Việc ước lượng β_0 và β_1 phải sử dụng các phần mềm chuyên dụng.
- ▶ Với hàm Probit thì phương pháp ước lượng cũng tương tự.

Giải thích ý nghĩa của mô hình Logit và Probit (1)

- Từ giả định xác suất của hành vi hút thuốc (3):

$$P(SMOKING_i = 1 | PRICE) = G(\beta_0 + \beta_1 * PRICE_i) \quad (11)$$

Với những thay đổi nhỏ của giá bán lẻ $PRICE$ thì tác động biên lên xác suất hút thuốc có thể được tính như sau:

$$\frac{\partial P(SMOKING)}{\partial PRICE} = g(\beta_0 + \beta_1 * PRICE_i) * \beta_1 \quad (12)$$

với $g(\beta_0 + \beta_1 * PRICE_i)$ là hàm mật độ phân phối xác suất.

- Trong phương pháp MLE, tác động biên của giá lên hành vi hút thuốc thay đổi tùy thuộc vào giá trị của hàm mật độ $g(.)$ tại giá bán gốc, khác với tác động biên cố định trong phương pháp hồi quy tuyến tính OLS!

Giải thích ý nghĩa của mô hình Logit và Probit (2)

- ▶ Thông thường chúng ta tính tác động biên tại mức giá trung bình, tại các tứ phân vị, giá trị tối đa/tối thiểu.
- ▶ Nếu biến giải thích là biến rời rạc (ví dụ có thêm biến giới tính trong hồi quy Logit đa biến) thì không áp dụng được công thức (12). Khi đó, tác động của giới tính đến hành vi hút thuốc có thể ước lượng trực tiếp:

$$\Delta P = P(SMOKING|MALE) - P(SMOKING|FEMALE) \quad (13)$$

$$= G(\beta_0 + \beta_1 * PRICE + D) - G(\beta_0 + \beta_1 * PRICE)$$

với D là biến giả đại diện cho giới tính.

Thực hành trên STATA (1)

- Sử dụng bộ dữ liệu MROZ.DTA để ước lượng mô hình giải thích nhân tố ảnh hưởng việc tham gia lao động chính thức của phụ nữ đã có gia đình.

```
sum inlf nwifeinc educ exper age kidslt6 kidsge6
```

Bảng mô tả dữ liệu

Variable	Description	Mean	Std. Dev.	Min	Max
inlf	Tham gia lao động	0.57	0.50	0	1
nwifeinc	Thu nhập ròng hộ gia đình	20.13	11.63	-0.03	96
educ	Số năm đi học	12.29	2.28	5	17
exper	Số năm kinh nghiệm	10.63	8.07	0	45
age	Tuổi	42.54	8.07	30	60
kidslt6	Số con dưới 6 tuổi	0.24	0.52	0	3
kidsge6	Số con từ 6 tuổi trở lên	1.35	1.32	0	8

N = 753

Thực hành trên STATA (2)

- ▶ Xác suất tham gia được giả định bị ảnh hưởng bởi các yếu tố như thu nhập ròng (tổng thu nhập gia đình trừ đi mức lương tối thiểu), trình độ học vấn, số năm kinh nghiệm làm việc (với tác động phi tuyến qua kinh nghiệm bình phương), tuổi, số con nhỏ dưới và trên 6 tuổi:

$$\begin{aligned} \text{inlf}_i = & \beta_0 + \beta_1 * \text{nwifeinc}_i + \beta_2 * \text{educ}_i + \beta_3 * \text{exper}_i + \beta_4 * \text{exper}_i^2 \\ & + \beta_5 * \text{age}_i + \beta_6 * \text{kidslt6}_i + \beta_7 * \text{kidsge6}_i + u_i \end{aligned}$$

- ▶ Mô hình trên có thể được ước lượng bằng phương pháp OLS cho mô hình xác suất tuyến tính (LPM), và phương pháp MLE cho mô hình Logit và Probit.

Thực hành trên STATA (3)

```
gen exper2 = exper*exper  
reg inlf nwifeinc educ exper exper2 age kidslt6 kidsge6, robust  
logit inlf nwifeinc educ exper exper2 age kidslt6 kidsge6  
probit inlf nwifeinc educ exper exper2 age kidslt6 kidsge6
```

	(1) LPM	(2) Logit	(3) Probit
nwifeinc	-0.00341* (-2.23)	-0.0213* (-2.53)	-0.0120* (-2.48)
educ	0.0380*** (5.23)	0.221*** (5.09)	0.131*** (5.18)
exper	0.0395*** (6.80)	0.206*** (6.42)	0.123*** (6.59)
exper2	-0.000596** (-3.14)	-0.00315** (-3.10)	-0.00189** (-3.15)
age	-0.0161*** (-6.71)	-0.0880*** (-6.04)	-0.0529*** (-6.23)
kidslt6	-0.262*** (-8.24)	-1.443*** (-7.09)	-0.868*** (-7.33)
kidsge6	0.0130 (0.96)	0.0601 (0.80)	0.0360 (0.83)
_cons	0.586*** (3.85)	0.425 (0.49)	0.270 (0.53)
N	753	753	753

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Diễn giải và so sánh giữa LPM với MLE như thế nào?

- ▶ Các ước lượng của mô hình LPM có thể diễn giải trực tiếp là tác động biên của các biến giải thích, ví dụ có thêm một con dưới 6 tuổi khiến xác suất tham gia lao động của phụ nữ giảm 26%.
- ▶ Tác động biên của ước lượng MLE phải tính bằng công thức (12) và (13). Do tác động biên thay đổi tùy thuộc vào giá trị gốc nên trong ví dụ này chúng ta tính tác động biên tại giá trị trung bình của các biến giải thích. Sử dụng lệnh:

```
logit inlf nwifeinc educ exper exper2 age kidslt6 kidsge6  
mfx, at(mean)
```

- ▶ Lưu ý: trị kiểm định của mô hình LPM là t-test, của mô hình Logit hoặc Probit là z-test.

Giải thích mô hình Logit (1)

```
Marginal effects after logit  
y = Pr(inlf) (predict)  
= .58277201
```

variable	dy/dx	Std. Err.	z	P> z	[	95% C.I.	]	X
nwifeinc	-.0051901	.00205	-2.53	0.011	-.009204	-.001176		20.129
educ	.0537773	.01056	5.09	0.000	.033078	.074476		12.2869
exper	.0500569	.00782	6.40	0.000	.034721	.065393		10.6308
exper2	-.0007669	.00025	-3.10	0.002	-.001252	-.000281		178.039
age	-.021403	.00354	-6.05	0.000	-.028341	-.014465		42.5378
kidslt6	-.3509498	.04964	-7.07	0.000	-.448241	-.253658		.237716
kidsge6	.0146162	.01819	0.80	0.422	-.021032	.050265		1.35325

- ▶ Tại giá trị trung bình của các biến giải thích trong **Bảng mô tả dữ liệu**, nếu số năm đi học tăng một năm, xác suất tham gia lao động tăng 5.4%.
- ▶ Có thể tính tác động biên tại các giá trị khác nhau của biến độc lập bằng lệnh:

```
mfx, at(mean kidslt6=0)
```

```
mfx, at(mean kidslt6=1)
```

```
mfx, at(mean kidslt6=2)
```

Giải thích mô hình Logit (2)

Có thể kiểm tra tác động biên của biến giáo dục bằng công thức:

$$\left. \frac{\partial P}{\partial X} \right|_{educ} = g(\beta_0 + \beta_1 * \overline{X_1} + \dots + \beta_6 * \overline{X_6}) * \beta_{educ} \quad (14)$$

$$\left. \frac{\partial P}{\partial X} \right|_{educ} = g(-.0213*20.13 + .221*12.29 + .206*10.63 - .00315*10.63^2$$
$$-.088 * 42.54 - 1.443 * .24 + .0601 * 1.35 + .425) * .221 \quad (15)$$

$$= g(.537) * .221$$

với hàm mật độ phân phối Logistic $g(.) = \frac{e^z}{(1+e^z)^2}$:

$$\left. \frac{\partial P}{\partial X} \right|_{educ} = \frac{2.718^{.537}}{(1 + 2.718^{.537})^2} * .221 = .0515 \quad (16)$$

⇒ Kết quả tương đồng với kết luận ở trên.

Giải thích mô hình Logit (3)

Có thể tính tác động biên bằng các lệnh sau trong STATA:

```
logit inlf nwifeinc educ exper exper2 age kidslt6 kidsge6
qui sum nwifeinc
scalar _x1 = r(mean)
qui sum educ
scalar _x2 = r(mean)
qui sum exper
scalar _x3 = r(mean)
qui sum age
scalar _x4 = r(mean)
qui sum kidslt6
scalar _x5 = r(mean)
qui sum kidsge6
scalar _x6 = r(mean)
scalar xb = (_b[_cons] + _b[nwifeinc]*_x1 + _b[educ]*_x2 +
_b[exper]*_x3 + _b[exper2]*_x3 ^ 2 + _b[age]*_x4 +
_b[kidslt6]*_x5 + _b[kidsge6]*_x6)
di 'Tac dong bien cua mot nam hoc =' exp(xb)/(1+exp(xb)) ^ 2 *
_b[educ]
```

Giải thích mô hình Logit (4)

- ▶ Đối với biến số con nhỏ dưới 6 tuổi, con số ước lượng 35% chỉ đúng nếu đây là biến liên tục.
- ▶ Do số con là biến rời rạc, phải sử dụng công thức (13) để tính tác động biên của việc có thêm một con.
- ▶ Ví dụ xác suất tham gia lao động thay đổi thế nào đối với phụ nữ chưa có con nhỏ dưới 6 tuổi chuyển sang có một con nhỏ dưới 6 tuổi?

$$\Delta P = \hat{G}(\beta_0 + \beta_1 * \overline{X_1} + \dots + \beta_{kidslt6}) - \hat{G}(\beta_0 + \beta_1 * \overline{X_1} + \dots) \quad (17)$$

với $G(.)$ là hàm phân phối tích lũy Logistic, $G(z) = \frac{e^z}{1+e^z}$:

$$\Delta P = \hat{G}(-.559) - \hat{G}(.884) = .364 - .708 = -.344$$

$\Rightarrow$ xác suất tham gia lao động giảm 34.4%.

Khả năng dự báo của mô hình xác suất (1)

- ▶ Khả năng dự báo của mô hình: thể hiện xác suất mô hình dự đoán đúng thực tế, bao gồm cả dự báo đúng việc tham gia và không tham gia.
- ▶ Một dự báo được coi là đúng nếu xác suất tham gia ước lượng được > 0.5 đối với người tham gia trên thực tế, và xác suất tham gia ước lượng được < 0.5 đối với người không tham gia trên thực tế.

Ma trận dự báo

	Dự báo		Tổng số
	Có	Không	
Thực tế	Có X1 [Đúng]	Không X2 [Sai]	X1 + X2
	Không X3 [Sai]	X4 [Đúng]	X3 + X4
Tổng số	X1 + X3	X2 + X4	$\sum X_i$

- ▶ Khả năng dự báo của mô hình $\gamma = \frac{X1+X4}{X1+X2+X3+X4}$

Khả năng dự báo của mô hình xác suất (2)

```
logit inlf nwifeinc educ exper exper2 age kidslt6 kidsge6
predict p, pr
di 'tính xác suất du bao dung voi phu nu di lam'
count if p>0.5 & inlf==1
count if inlf == 1
di 'tính xác suất du bao dung voi phu nu không di lam'
count if p<0.5 & inlf==0
count if inlf == 0
```

		<u>Ma trận dự báo</u>		Tổng số
		Dự báo		
Thực tế	Có	Có 347	Không 81	428
	Không	118	207	325
Tổng số		465	288	753

► $\gamma = \frac{(347+207)}{753} = 73.57\%$

Khả năng dự báo của mô hình xác suất (3)

Có thể làm đơn giản hơn bằng lệnh:

```
estat classification
```

Logistic model for inlf

Classified	----- True -----		Total
	D	~D	
+	347	118	465
-	81	207	288
Total	428	325	753

Kiểm định hồi quy Logit

- ▶ Đối với kiểm định đơn biến, sử dụng z-test.
- ▶ Đối với kiểm định đa biến, sử dụng kiểm định Likelihood Ratio (LR). Ví dụ kiểm định k tham số ước lượng đồng thời không có ý nghĩa thống kê:

$H_0 : \beta_{i_1} = \dots = \beta_{i_k} = 0$ với $H_1 : \text{Ít nhất một } \beta_{i_s} \neq 0$

- ▶ Ước lượng hai mô hình riêng biệt: mô hình không giới hạn (unrestricted, ur) với đầy đủ biến giải thích, và mô hình giới hạn (restricted, r) không có biến giải thích $X_{i_1}, \dots, X_{i_k}$.
- ▶ Tính trị kiểm định $LR = 2 * (\mathbf{L}_{ur} - \mathbf{L}_r)$, với $\mathbf{L}_{ur}$ và $\mathbf{L}_r$ là giá trị log-likelihood từ công thức (7) và tương ứng với mô hình không giới hạn và mô hình giới hạn.
- ▶ LR có phân phối χ_k^2 với số bậc tự do k .

Kiểm định hồi quy Logit

Ví dụ kiểm định hai biến *kidslt6* và *kidsge6* đồng thời không có ý nghĩa thống kê. Có hai cách kiểm định LR:

- ▶

```
logit inlf nwifeinc educ exper exper2 age kidslt6 kidsge6  
scalar _lur = e(11)  
logit inlf nwifeinc educ exper exper2 age  
scalar _lr = e(11)  
di 2*(_lur-_lr)
```
- ▶

```
logit inlf nwifeinc educ exper exper2 age kidslt6 kidsge6  
eststo U  
logit inlf nwifeinc educ exper exper2 age  
eststo R  
lrtest U R
```

Likelihood-ratio test LR $\chi^2(2) = 62.02$

(Assumption: R nested in U) Prob > $\chi^2 = 0.0000$

⇒ bác bỏ H_0 . Ít nhất một trong hai biến giải thích, hoặc cả hai đều có ý nghĩa thống kê trong mô hình.

Tự thực hành

- ▶ Học viên ước tính tác động biên sử dụng mô hình Probit ở những giá trị khác nhau của biến giải thích: giá trị tối đa/tối thiểu, các tứ phân vị. So sánh các tác động biên này với ước lượng từ mô hình LPM.
- ▶ Học viên so sánh tác động biên và khả năng dự báo của mô hình Logit và Probit.

Tài liệu tham khảo

1. Verbeek, Marno 2000. A Guide to Modern Econometrics, John Wiley & Sons (Chapter 6 & 7).
2. Wooldridge, Jeffrey M. 2002. Introductory Econometrics: A Modern Approach, South-Western (2nd edition, Chapter 17).