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

Lê Việt Phú
Trường Chính sách Công và Quản lý Fulbright

07/1/2020

Khái niệm 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

- 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.

- 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 phương pháp OLS cùng các giả định của mô hình CLRM vào dữ liệu có biến phụ thuộc bị giới hạn?

Xem xét mô hình:

$$Smoking_i = \beta_0 + \beta_1 * Price_i + u_i \tag{1}$$

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.

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

- 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 sai số bằng 0, $E[u|Price] = 0$:

$$E[Smoking|Price] = \beta_0 + \beta_1 * Price \tag{2}$$

- Đồng thời:

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

$$\Rightarrow$$

$$P(Smoking = 1|Price) = E[Smoking|Price] = \beta_0 + \beta_1 * Price$$

- 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 $\beta_1 = -0.1$, nếu giá bán tăng 1 đơn vị thì xác suất vị thành niên hút thuốc sẽ giảm 10%.

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.[1]

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

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

Do $Var(u_i|Price_i)$ phụ thuộc vào $Price_i$, hay nói cách khác, phương sai của sai số trong mô hình LPM thay đổi.

[1] Biến 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 luỹ $G(.)$:

$$P(Smoking_i = 1|Price_i) = G(\beta_0 + \beta_1 * Price_i) \qquad (3)$$

  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 tích lũy $G(.)$ dựa vào giả định hoặc các lý thuyết kinh tế để giải thích.

Các hàm phân phối xác suất thông dụng

- Nếu $G(.)$ có phân phối tích luỹ Logistic, khi đó ta có hồi quy "Logit":

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

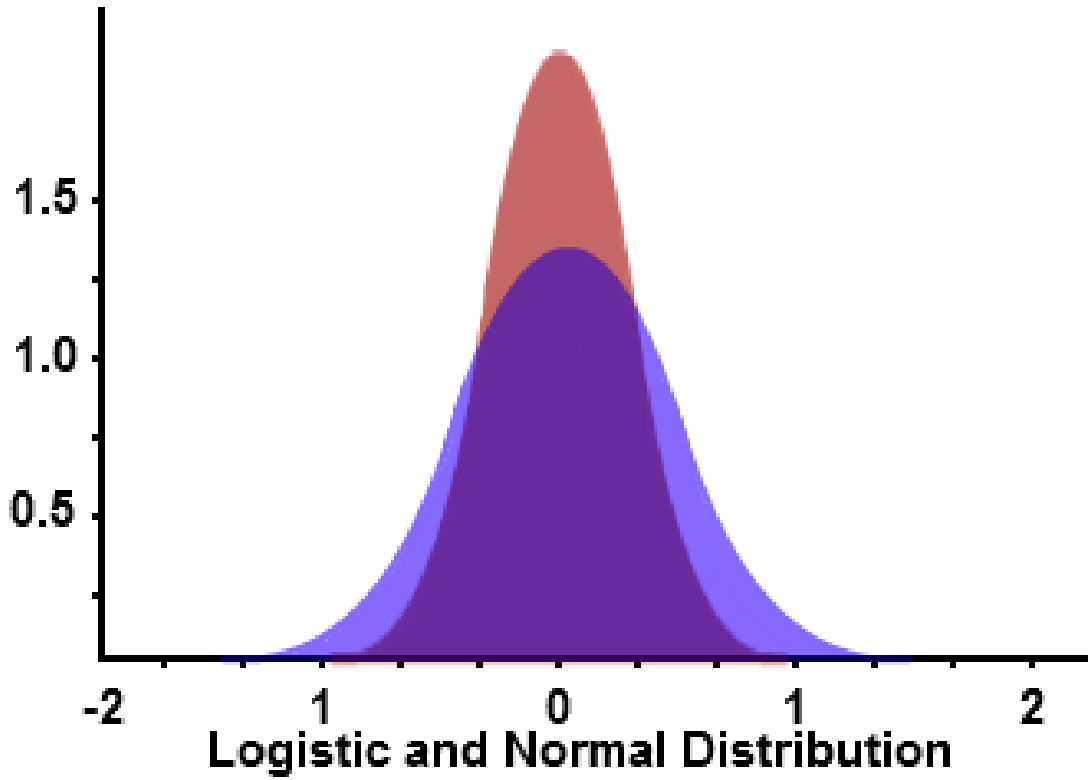
với hàm mật độ phân phối Logistic $g(z) = G'(z) = \frac{e^z}{(1+e^z)^2}$

- Nếu $G(.)$ có phân phối tích luỹ chuẩn $\Rightarrow$ hồi quy Probit:

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

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

Đồ thị hàm mật độ phân phối Logistic (màu tím) và phân phối chuẩn (màu cam)



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

Ước lượng mô hình hồi quy Logit và Probit

- Khác với phương pháp sai số bình phương 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 sai số của mô hình, còn hàm mục tiêu của phương pháp MLE là tối đa xác suất quan sát được mẫu với thuộc tính cho trước.

- 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 $P(Smoking_i|Price_i) = 1 - G(.)$

- Phương pháp MLE ước lượng các tham số của hàm xác suất $G(.)$ bằng cách tối đa hóa tích của xác suất quan sát được một mẫu có những người hút thuốc và không hút thuốc:

$$\begin{aligned} Max\ \mathbf{P}_{MLE} &= \prod_{\mathbf{i=1}}^{\mathbf{N}} P(Smoking_i|Price_i) \quad (5) \\ &= \prod_{\mathbf{i=1}}^{\mathbf{N}} [G(.)]^{Smoking_i} \times [1 - G(.)]^{1-Smoking_i} \end{aligned}$$

- Do G(.) là hàm đơn điệu (hàm phân phối xác suất tích luỹ chỉ tăng hoặc giảm theo biến giải thích), chúng ta có thể đơn giản hàm tối ưu tích (5) sang hàm tối ưu tổng bằng cách lấy logarithm và tối đa giá trị log-likelihood **L**:

$$Max\ \mathbf{L}_{MLE} = \sum_{\mathbf{i=1}}^{\mathbf{N}} \underbrace{\left\{ S_i * ln[G(.)] + [1 - S_i] * ln[1 - G(.)] \right\}}_{\ell_i}$$

$$= \sum_{\mathbf{i=1}}^{\mathbf{N}} \ell_i \tag{6}$$

với S_i là tình trạng hút thuốc, $Smoking_i$, và $G(.)$ là hàm phân phối xác suất tích luỹ $G(\beta_0 + \beta_1 * Price_i)$.

- Để ước lượng tham số β_0 và β_1 nhằm tối đa giá trị **L**, sử dụng điều kiện đạo hàm bậc nhất (first-order condition):

$$\frac{\partial \mathbf{L}}{\partial \beta_0} = \sum_{\mathbf{i=1}}^{\mathbf{N}} \frac{\partial \ell_i}{\partial \beta_0} = 0$$

$$\frac{\partial \mathbf{L}}{\partial \beta_1} = \sum_{\mathbf{i=1}}^{\mathbf{N}} \frac{\partial \ell_i}{\partial \beta_1} = 0$$

Lưu ý quy tắc chuỗi (chain-rule) khi lấy đạo hàm của hàm hợp:

$$\frac{\partial G(\beta_0 + \beta_1 * X_i)}{\partial \beta_0} = g(\beta_0 + \beta_1 * X_i)$$
$$\frac{\partial G(\beta_0 + \beta_1 * X_i)}{\partial \beta_1} = g(\beta_0 + \beta_1 * X_i) * X_i$$
$$\frac{\partial ln[G(.)]}{\partial \beta_0} = \frac{1}{G(.)} * g(.)$$
$$\frac{\partial ln[G(.)]}{\partial \beta_1} = \frac{1}{G(.)} * g(.) * X_i$$

$$\frac{\partial \mathbf{L}}{\partial \beta_0} = \sum_{\mathbf{i}} \left\{ S_i * g(.) - [1 - S_i] * \frac{1}{1 - G(.)} * g(.) \right\} = 0 \qquad (7)$$

$$\frac{\partial \mathbf{L}}{\partial \beta_1} = \sum_{\mathbf{i}} \left\{ \frac{S_i}{G(.)} * g(.) * X_i - [1 - S_i] * \frac{1}{1 - G(.)} * g(.) * X_i \right\} = 0 \qquad (8)$$

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 đơn giản hóa thành:

$$\frac{\partial \mathbf{L}}{\partial \beta_0} = \sum_{\mathbf{i}} S_i - \sum_{\mathbf{i}} \frac{e^{\beta_0+\beta_1*X_i}}{1+e^{\beta_0+\beta_1*X_i}} = 0 \tag{9}$$

$$\frac{\partial \mathbf{L}}{\partial \beta_1} = \sum_{\mathbf{i}} S_i * X_i - \sum_{\mathbf{i}} \frac{e^{\beta_0+\beta_1*X_i}}{1+e^{\beta_0+\beta_1*X_i}} * X_i = 0 \tag{10}$$

- 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 $\hat{\beta}_0$ và $\hat{\beta}_1$ phải sử dụng phương pháp số (numerical solution) bằng 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

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

$$P(Smoking_i = 1|Price_i) = G(\beta_0 + \beta_1 * Price_i) \qquad (11)$$

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:

$$\frac{\partial P(Smoking)}{\partial Price} = g(\beta_0 + \beta_1 * Price_i) * \beta_1 \qquad (12)$$

với $g(\beta_0 + \beta_1 * Price_i)$ là hàm mật độ phân phối xác suất, tính tại giá trị $Price_i$.

- Trong phương pháp MLE, tác động biên của giá lên hành vi hút thuốc thay đổi tuỳ 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 xác suất tuyến tính LPM!

- Thông thường chúng ta tính tác động biên tại mức giá trung bình, tại các tứ phân vị, tại các giá trị tối đa/tối thiểu.
- Nếu biến giải thích là biến rời rạc (ví dụ có thêm biến giới tính hay số con trong gia đình 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 từ công thức (11):

$$\begin{aligned}\Delta P &= P(Smoking|Male) - P(Smoking|Female) \\ &= G(\beta_0 + \beta_1 * Price + D) - G(\beta_0 + \beta_1 * Price) \quad (13)\end{aligned}$$

với D là tham số của biến giới tính.

So sánh giữa LPM, Logit và Probit

Sử dụng bộ dữ liệu mô phỏng SMOKE.dta

Regression Results

	LPM b/se	Logit b/se	Probit b/se
main			
sex	0.0050	0.0214	0.0132
	(0.0526)	(0.2227)	(0.1377)
price	-0.0028	-0.0116	-0.0072
	(0.0036)	(0.0152)	(0.0094)
Constant	0.5461*	0.2082	0.1263
	(0.2270)	(0.9530)	(0.5910)
N	807	807	807

* p<0.05, ** p<0.01, *** p<0.001

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

- **LPM:**

$$\widehat{Smoke_i} = .5461 + .0050 * sex_i - .0028 * price_i$$

- Đối tượng là nam giới có xác suất hút thuốc lá cao hơn nữ giới là 0.5%. Giá thuốc lá tăng 1 đơn vị (cent) làm giảm xác suất hút thuốc 0.28%.
- Tác động biên là hằng số, không phụ thuộc vào giá bán gốc.

- **Hồi quy Logit:**

$$G(z) = \frac{e^z}{1+e^z} \Rightarrow \frac{G(z)}{1-G(z)} = e^z$$
$$\Rightarrow log\left(\frac{G(z)}{1-G(z)}\right) = z$$

Phương trình hồi quy logit được viết dưới dạng log của tỷ lệ thành công (odds ratio - OR[2]):

$$log\left(\frac{\widehat{Smoke_i}}{1-\widehat{Smoke_i}}\right) = .2082 + .0214 * sex_i - .0116 * price_i$$

- Cách 1: Diễn giải tham số qua OR. Đối tượng là nam giới làm tăng OR hành vi hút thuốc thêm 0.0214. Giá tăng một đơn vị làm giảm OR hành vi hút thuốc là 0.0116.

[2]OR là tỷ số của xác suất xảy ra một sự kiện cho xác suất không xảy ra sự kiện đó. $OR = 1$ khi biến cố xảy ra hay không có xác suất như nhau. $OR > 1$ khi xác suất xảy ra cao hơn khả năng không xảy ra, và ngược lại.

Cách 2: Diễn giải thành tác động lên xác suất hút thuốc, áp dụng công thức (12) và (13) cho biến số là liên tục hay rời rạc.

- Đối với biến giá là biến liên tục. Giả định chúng ta muốn ước lượng tác động biên của việc tăng giá lên đối tượng là nam giới ($sex = 1$), tại mức giá trung bình ($price = 60.03$):

$$\begin{aligned}\frac{\partial P(SMOKE)}{\partial price} &= g(.) * \beta_{price} = \frac{e^z}{(1+e^z)^2} * \beta_{price} \\ &= \frac{e^{(.0214-.0116*60.03+.2082)}}{(1+e^{(.0214-.0116*60.03+.2082)})^2} * (-.0116) \\ &= -.0027451\end{aligned}$$

$\Rightarrow$ tăng giá thuốc lá 1 cent/bao từ mức giá trung bình làm giảm xác suất hút thuốc là 0.27% với đối tượng là nam.

Homework: Nếu mức giá gốc lần lượt là 44 cent/bao và 70 cent/bao thì tác động biên là bao nhiêu?

- Đối với biến giới tính là biến rời rạc. Giả định chúng ta muốn ước lượng sự khác biệt của xác suất hút thuốc giữa hai nhóm nam và nữ, tại mức giá trung bình ($price = 60.03$):

$$\begin{aligned}\Delta P &= P(Smoking|Male) - P(Smoking|Female) \\ &= G(\beta_0 + \beta_1 * Price + D) - G(\beta_0 + \beta_1 * Price)\end{aligned}$$

- Hàm phân phối tích lũy Logit là $G(z) = \frac{e^z}{1+e^z}$,

$$\Delta P = .0050433 \approx 0.5\%$$

- **Hồi quy Probit:**

$$\widehat{Smoke_i} = \widehat{\Phi(z)} = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{(0.1263+0.0132*sex_i-0.0072*price_i)^2}{2}} dx$$

- Khó diễn giải trực tiếp tham số của mô hình Probit!!
- Có thể sử dụng công thức (12) hoặc (13) để ước tính tác động của các biến số tại các giá trị cho trước.

Ước lượng tác động biên đối với mô hình Probit, với đối tượng là nam giới, tại mức giá trung bình:

$$\begin{aligned}\frac{\partial P(SMOKE)}{\partial price} &= g(.) * \beta_{price} = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} * \beta_{price} \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{(0.0132-.0072*60.03+.1263)^2}{2}} * (-.0072) \\ &= -.0027423\end{aligned}$$

$\Rightarrow$ tăng giá thuốc lá 1 cent/bao từ mức giá trung bình làm giảm xác suất hút thuốc là .27% với đối tượng là nam.

Homework: Nếu mức giá gốc lần lượt là 44 cent/bao và 70 cent/bao thì tác động biên là bao nhiêu?

- Khác biệt về xác suất hút thuốc giữa nhóm nam và nữ như thế nào, tại mức giá trung bình?

$$\begin{aligned}\Delta P &= P(Smoking|Male) - P(Smoking|Female) \\ &= G(\beta_0 + \beta_1 * Price + D) - G(\beta_0 + \beta_1 * Price)\end{aligned}$$

- Hàm phân phối chuẩn (Probit) là $G(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$,

$$\Delta P = .0050428 \approx 0.5\%$$

Đánh giá mô hình xác suất

- Khả năng dự báo của mô hình: thể hiện xác suất mô hình dự đoán đúng thực tế, bao gồm cả dự báo đúng hành vi hút thuốc và không hút thuốc.
- Một dự báo được coi là đúng nếu xác suất hút thuốc ước lượng được > 0.5 đối với người có hút thuốc, và xác suất không hút thuốc ước lượng được < 0.5 đối với người không hút thuốc.

Ma trận dự báo

		Dự báo		Tổng số
		Có	**Không**	
Thực tế	**Có**	X1 **[Đúng]**	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 $\gamma = \frac{X1+X4}{X1+X2+X3+X4}$

Ma trận dự báo

		Thực tế		Tổng số
		Có	**Không**	
Dự báo	**Có**	0	0	0
	Không	310	497	807
Tổng số		310	497	807

- $\gamma = \frac{(0+497)}{807} = 61.59\%$
- Do dữ liệu tự mô phỏng dẫn đến mô hình này dự đoán sai hoàn toàn đối với những người hút thuốc!

Thực hiện trong Stata:

```
. estat classification

Logistic model for SMOKE
```

	True		
Classified	D	~D	Total
+	0	0	0
-	310	497	807
Total	310	497	807

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:
  - $H_0 : \ \beta_i = ... = \beta_k = 0$
  - H_1 : Ít nhất một $\beta_j \neq 0$

Các bước kiểm định LR

1. Ước lượng hai mô hình riêng biệt: mô hình không giới hạn (unrestricted, u) với đầy đủ biến giải thích, và mô hình giới hạn (restricted, r) không có biến giải thích $X_1, ..., X_k$.
2. Tính trị kiểm định $LR = 2 * (\mathbf{L_u} - \mathbf{L_r})$, với $\mathbf{L_u}$ và $\mathbf{L_r}$ là giá trị log-likelihood từ công thức (6) và tương ứng với mô hình không giới hạn và mô hình giới hạn.[3]
3. LR có phân phối χ^2_k với số bậc tự do k.
4. Bác bỏ giả thuyết $H_o \Rightarrow$ ít nhất một trong các tham số kiểm định $\beta_j \neq 0$.

Thực hành trên Stata với bộ dữ liệu SMOKE.dta.

[3] Cơ chế của kiểm định đa biến LR trong phương pháp MLE tương đồng với kiểm định F trong phương pháp OLS. Khi áp ràng buộc vào mô hình thì xác suất tối ưu bị giảm, khiến $\mathbf{L_r} < \mathbf{L_u}$, hoặc tổng bình phương sai số của mô hình tăng $SSR_r > SSR_u$.