Fulbright | TRƯỜNG CHÍNH SÁCH CÔNG VÀ QUẢN LÝ FULBRIGHT

MÔ HÌNH HỒI QUY XÁC SUẤT

Hoàng Văn Thắng

MPP2020-PA, 7/3/2019

1

Nội dung thực hành

- Ước lượng mô hình
- Tính toán tác động biên và giải thích ý nghĩa của hệ số hồi quy ước lượng
- Khả năng dự báo của mô hình
- Kiểm định các vấn đề của mô hình

2

Ước lượng mô hình logit

- Thông tin dữ liệu
  - lfp: tình trạng làm việc của phụ nữ (1 = Yes | 0 = No)
  - k5: số con dưới 6 tuổi trong hộ
  - k618: số con từ 6 – 18 tuổi trong hộ
  - age: tuổi của người phụ nữ
  - wc: phụ nữ có bằng cấp hay không (1 = Yes | 0 = No)
  - hc: người chồng có bằng cấp hay không (1 = Yes | 0 = No)
  - lwg: thu nhập kỳ vọng của người phụ nữ nếu đi làm
  - inc: thu nhập của hộ gia đình *(không bao gồm thu nhập của người phụ nữ)*

3

Ước lượng mô hình logit

- Ký hiệu và khái niệm liên quan
  - Đặt p = prob(lfp=1) = xác suất phụ nữ đi làm
  - Vậy q = 1 – p = prob(lfp=0) = xác suất phụ nữ không đi làm

  - Odds of success = p/q [= odds ratio] ➔ logistic Yi Xi
  - Odds of failure = q/p

p	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
odds	0,11	0,25	0,43	0,67	1,00	1,50	2,33	4,00	9,00
ln(odds)	(2,20)	(1,39)	(0,85)	(0,41)	-	0,41	0,85	1,39	2,20

4

Mô hình Logit

- Đặt z = $\beta_0 + \beta_i * X_i$
- Các cách viết khác nhau để tính xác suất

$$\hat{p} = \frac{1}{1+e^{-z}} = \hat{p} = \frac{1}{1+e^{-(\beta_0+\beta_i*X_i)}}$$

$$\hat{p} = \frac{e^Z}{1+e^Z} = \frac{e^{(\beta_0+\beta_i*X_i)}}{1+e^{(\beta_0+\beta_i*X_i)}}$$

$$\frac{\hat{p}}{1-\hat{p}} = e^Z$$

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = z = (\beta_0 + \beta_i * X_i)$$

5

Tác động biên của mô hình Logit

$\hat{p} = \frac{e^Z}{1+e^Z}$ và $\frac{\hat{p}}{1-\hat{p}} = e^Z$

$$\frac{dp_i}{dX_i} = \frac{dp_i}{dZ_i} * \frac{dZ_i}{dX_i}$$

$$\frac{dp_i}{dZ_i} = \frac{(e^Z)' * (1+e^Z) - (1+e^Z)' * e^Z}{(1+e^Z)^2} = \frac{e^Z}{(1+e^Z)^2} = \frac{e^Z}{(1+e^Z)} * \frac{1}{(1+e^Z)} = p_i * \frac{1}{1+\frac{p_i}{1-p_i}} = p_i * (1-p_i) \quad (1)$$

$$\frac{dZ_i}{dX_i} = (\beta_1 + \beta_2 * X_{2i} + \beta_3 * X_{3i} + \cdots + \beta_k * X_{ki})' = \beta_k \quad (2)$$

Từ (1) và (2) ➔ $\frac{dp_i}{dX_i} = \beta_k * p_i * (1-p_i)$ ➔ $\frac{dp_i}{dX_i} \epsilon\, p_i$

6

Tác động biên của mô hình Logit [Manual]

- Bước 1: Hồi quy mô hình logit (≠ logistic)
  - ✓ Logit ➔ Estimated coefficients [$\hat{\beta}_i$]
  - ✓ Logistic ➔ Odds ratio
- Bước 2: Xác định nhóm đối tượng muốn quan sát (chọn X_i) do xác suất khác nhau tại mỗi nhóm khác nhau.
- Bước 3: Tính $\hat{p}$ = prob(Yi = 1) từ [$\hat{\beta}_i$] và X_i
- Bước 4: Tính tác động biên theo công thức

$$\frac{dp_i}{dX_i} = \hat{\beta}_i * \hat{p} * (1 - \hat{p})$$

7

Tác động biên của mô hình Logit [Stata]

- Bước 1: Hồi quy mô hình logit (≠ logistic)
  - ✓ Logit ➔ Estimated coefficients [$\hat{\beta}_i$]
  - ✓ Logistic ➔ Odds ratio
- Bước 2: Xác định nhóm đối tượng muốn quan sát (chọn X_i) do xác suất khác nhau tại mỗi nhóm khác nhau.
- Bước 3: Tính tác động biên bằng cấu trúc lệnh:

mfx, at (X_1 = … X_2 = … X_k = ….)

  - ✓ X = Mean với các X_i không khai báo
  - ✓ Với các X_i là biến dummy, tính mfx trong từng trường hợp và so sánh p_i trong mỗi trường hợp

8

Tác động biên của mô hình Logit [Stata]

```
. mfx

Marginal effects after logit
      y  = Pr(lfp) (predict)
         =  .57388476
```

variable	dy/dx	Std. Err.	z	P>|z|	[95% C.I.]		X
k5	-.3482605	.0484	-7.20	0.000	-.443121	-.2534	.237716
k618	-.0189813	.01649	-1.15	0.250	-.0513	.013338	1.35325
age	-.0151834	.00309	-4.91	0.000	-.021247	-.00912	42.5378
hc*	.1058854	.04373	2.42	0.015	.020167	.191604	.391766
lwg	.177781	.03653	4.87	0.000	.106181	.249381	1.09714
inc	-.0076498	.00196	-3.89	0.000	-.0115	-.0038	20.1293

(*) dy/dx is for discrete change of dummy variable from 0 to 1

9

Tác động biên của mô hình Probit [Manual]

- Bước 1: Hồi quy mô hình probit (≠ logistic)
  - ✓ Probit ➔ Estimated coefficients [$\hat{\beta}_i$]
- Bước 2: Xác định nhóm đối tượng quan sát (chọn X_i) do xác suất khác nhau tại mỗi nhóm khác nhau.
- Bước 3: Tính $\hat{p}$ = prob(Yi = 1) từ [$\hat{\beta}_i$] và X_i
- Bước 4: Tính tác động biên theo công thức

$$\frac{dp_i}{dX_i} = \frac{1}{\sqrt{2\pi}} * e^{-\left(\frac{I_0^2}{2}\right)} * \hat{\beta}_i = prob(I < I_0) * \hat{\beta}_i = normsdist(I_0) * \hat{\beta}_i$$

$$với\ I_0 = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}$$

- Với Stata, cách tính tương tự như Logit

10

Tác động biên của mô hình Probit [Stata]

```
. probit lfp k5 k618 age hc lwg inc

Iteration 0:   log likelihood =  -514.8732
Iteration 1:   log likelihood = -459.42824
Iteration 2:   log likelihood = -459.30315
Iteration 3:   log likelihood = -459.30314

Probit regression                                 Number of obs   =        753
                                                  LR chi2(6)      =     111.14
                                                  Prob > chi2     =     0.0000
Log likelihood = -459.30314                       Pseudo R2       =     0.1079
```

lfp	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
k5	-.8479583	.1131212	-7.50	0.000	-1.069672	-.6262449
k618	-.0461519	.0404131	-1.14	0.253	-.1253602	.0330563
age	-.0373581	.0075675	-4.94	0.000	-.05219	-.0225261
hc	.2569849	.1103053	2.33	0.020	.0407905	.4731793
lwg	.429156	.0850771	5.04	0.000	.2624079	.595904
inc	-.0185569	.0046903	-3.96	0.000	-.0277496	-.0093641
_cons	1.847385	.3785987	4.88	0.000	1.105345	2.589424

11

Khả năng dự báo của mô hình

- Với LPM: khả năng dự báo là R_2 hoặc R^2_{adj}
- Với Logit hay Probit, ý tưởng tiếp cận như sau:
  - ✓ Bước 1: Ước lượng mô hình
  - ✓ Bước 2: Dự báo, tạo biến $\hat{p}_i$ (xác suất ứng với các Xi từ dữ liệu trong mẫu)
  - ✓ Tạo biến forecast nhận giá trị bằng 1 nếu $\hat{p}_i$ > 0,5
    [Giá trị 0,5 có thể thay đổi tùy theo mức độ chấp nhận rủi ro của người làm dự báo]
  - ✓ Tạo biến test = forecast = Y_i từ dữ liệu ban đầu
    [Test nhận giá trị 1 nếu forecast = Y_i]
- Với STATA, gõ lệnh estat classification sau hồi quy để ra kết quả.

12

Khả năng dự báo của mô hình logit [Stata]

Logistic model for lfp

Classified	True D	True ~D	Total
+	342	168	510
-	86	157	243
Total	428	325	753

Correctly classified 66.27%

13

Khả năng dự báo của mô hình probit [Stata]

Probit model for lfp

Classified	True D	True ~D	Total
+	344	170	514
-	84	155	239
Total	428	325	753

Correctly classified 66.27%

14

Các kiểm định liên quan

- Kiểm định hệ số hồi quy [Estimated Coefficients]: kiểm định Z thay vì kiểm định t
- Kiểm định LM để lựa chọn mô hình thêm biến, loại biến [lý thuyết vẫn quan trọng nhất quyết định biến đưa vào]
- Các kiểm định Robust, Reset Ramsey không thể thực hiện trực tiếp từ kết quả hồi quy Logit/Probit nhưng vẫn có thể thực hiện thông qua các kiểm định từ phần dư của mô hình hồi quy.
- Đa cộng tuyến [Tương quan mạnh giữa các biến X_i]

15