Fulbright | TRƯỜNG CHÍNH SÁCH CÔNG VÀ QUẢN LÝ FULBRIGHT

# MÔ HÌNH HỒI QUY XÁC SUẤT

Hoàng Văn Thắng

*MPP2020-PA, 5/3/2019*

1

## Nội dung

- Phân loại biến
  - Biến liên tục và biến rời rạc
  - Biến định tính và biến định lượng
  - Biến bị giới hạn và biến không giới hạn
- OLS – các giả định và phương pháp ước lượng
  - Tối thiểu hóa tổng bình phương phần dư
- Mô hình hồi quy khi biến phụ thuộc là giới hạn
  - Mô hình xác suất tuyến tính (LPM)
  - Mô hình logit, probit (MLE)
- Ý nghĩa từ kết quả hồi quy và tác động biên
  - Khi $X_i$ là biến định lượng
  - Khi $X_i$ là biến định tính

2

## Phân loại biến

- Biến liên tục và biến rời rạc
  - Tốc độ tăng trưởng, thu nhập, xác suất xảy ra một sự kiện
  - Số người phụ thuộc, số lần thay đổi công việc, số sản phẩm lỗi
- Biến định tính và biến định lượng
  - Có đi làm hay không đi làm; có vay tín dụng hay không
  - Số ngày làm việc trong năm; dư nợ tín dụng hiện tại là bao nhiêu
- Biến không bị giới hạn và bị giới hạn
  - Tiền lãi/lỗ từ hoạt động kinh doanh
  - Thu nhập từ đi làm trong năm
  - Số nhân viên trong tổ chức, số ngày nghỉ chế độ trong năm

3

## Hồi quy OLS và phương pháp ước lượng

- Xem xét mô hình SLR đơn giản

$$Y_i = \beta_0 + \beta_1 * X_i + u_i$$

  - Các giả định đi kèm mô hình là gì?
  - $\beta_1$ được ước lượng như thế nào?
  - Kết quả dự báo của $Y_i$ có bị giới hạn hay không?

4

## Mô hình hồi quy khi biến phụ thuộc bị giới hạn

- Mô hình nào phù hợp khi biến phụ thuộc là:
  - Đi làm hay không đi làm, mua nhà hay không mua nhà,…?
  - Làm trong khu vực nhà nước, khu vực FDI hay doanh nghiệp tư nhân trong nước?
  - Số ngày đi khám sức khỏe trong năm?
  - Mức độ hài lòng với dịch vụ hành chính công tại địa phương: không hài lòng, đáp ứng cơ bản, hài lòng?
  - Thu nhập kiếm được trong năm?

5

## Các mô hình có thể tiếp cận

- Mô hình xác suất tuyến tính (LPM)
- 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)

6

## Tình huống đặt ra

- Bạn đang xem xét tác động của giá thuốc lá lên hành vi hút thuốc của người tiêu dùng. Việc tăng giá thuốc lá có làm thay đổi hành vi hút thuốc của người tiêu dùng?

$$SMOKING_i = \beta_0 + \beta_1 * PRICE_i + u_i \qquad [1]$$

- ✓ *SMOKING = 1: nếu người được hỏi có hút thuốc*
- ✓ *SMOKING = 0: nếu người được hỏi không hút thuốc*
- ✓ *PRICE là giá bán lẻ của thuốc lá*

➔ Ước lượng OLS thông thường có phù hợp?

7

## Mô hình xác suất tuyến tính
## *(Linear Probability Model – LPM)*

$$SMOKING_i = \beta_0 + \beta_1 * PRICE_i + u_i \qquad [1]$$

*Vì* $E(u_i|PRICE_i) = 0$ *khi ước lượng theo OLS nên*

- $E(SMOKING = 1|PRICE_i) = \beta_0 + \beta_1 * PRICE_i$
- $E(SMOKING) = 1 * P(SMOKING = 1) + 0 * P(SMOKING = 0)$

➔ $P(SMOKING = 1|PRICE_i) = \beta_0 + \beta_1 * PRICE_i \qquad$ [LPM]

- reg SMOKE sex price [Stata Code]
- Giải thích kết quả mô hình:

P(SMOKE = 1) = 0.5461 – 0.0027*PRICE + 0.005*SEX

8

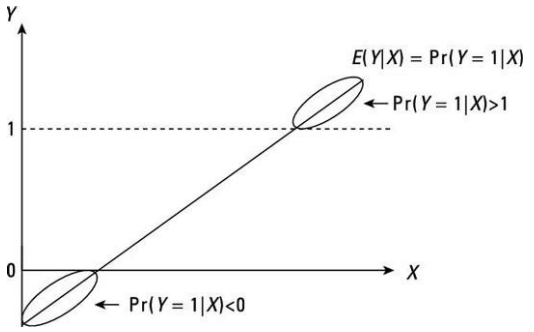
## Mô hình LPM và vấn đề không phù hợp

Gọi $P_i$ là xác xuất để $Y_i = 1$ và $(1-P_i)$ là xác suất để $Y_i = 0$.

Như vậy $Y_i$ có phân phối xác xuất Bernoulli

➔ $E(Y_i) = 0*(1 - P_i) + 1*P_i = P_i$.

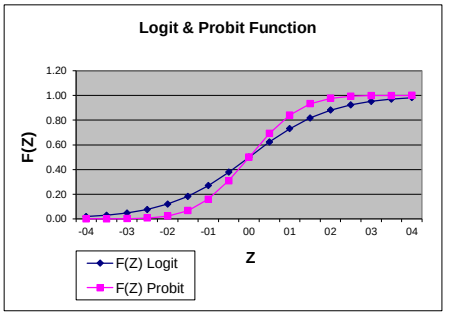
➔ $var(u_i) = P_i (1 - P_i) \neq const$

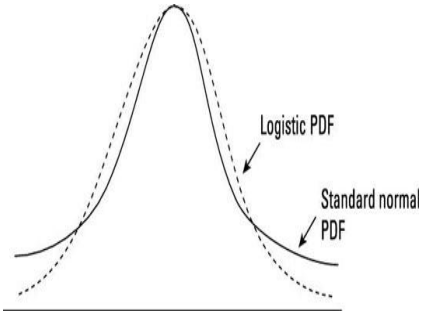



9

## Phương pháp Maximum Likelihood Estimation (MLE) và cách khắc phục hạn chế từ LPM

- Mục tiêu của OLS là tối thiểu hóa phần dư
- Mục tiêu của MLE là tối đa hóa việc quan sát được mẫu với các thuộc tính cho trước




Nguồn: Cao Hào Thi

10

## Mô hình Logit

- Đặt z = $\beta_0 + \beta_1 * PRICE_i$
- Các cách viết khác nhau để tính xác suất

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

$$\hat{p} = \frac{e^z}{1 + e^z}$$

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

$$\ln\left(\frac{\hat{p}}{1 - \hat{p}}\right) = z$$

11

## Mô hình Logit

- logit SMOKE sex price [Stata Code]

```
Logistic regression                             Number of obs     =        807
                                                LR chi2(2)        =       0.60
                                                Prob > chi2       =     0.7414
Log likelihood = -537.20635                     Pseudo R2         =     0.0006
```

| SMOKE | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| sex | .0213603 | .2226991 | 0.10 | 0.924 | -.4151219 | .4578425 |
| price | -.0115975 | .0152353 | -0.76 | 0.447 | -.0414581 | .018263 |
| _cons | .2082038 | .9529955 | 0.22 | 0.827 | -1.659633 | 2.076041 |

12

## Mô hình Logit

- Đặt $z_0 = \beta_0 + \beta_1 * PRICE_i$

$$\hat{p} = p(z < z_0) = \text{normsdist}(z_0);$$

- Vì $z_0$ tuần theo quy luật của hàm phân phối xác suất tích lũy
- probit SMOKE sex price [Stata Code ]

```
Probit regression                                 Number of obs     =        807
                                                  LR chi2(2)        =       0.60
                                                  Prob > chi2       =     0.7419
Log likelihood = -537.20698                       Pseudo R2         =     0.0006
```

| SMOKE | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| sex | .0132215 | .1377357 | 0.10 | 0.924 | -.2567356 | .2831785 |
| price | -.0071758 | .0094458 | -0.76 | 0.447 | -.0256892 | .0113376 |
| _cons | .1262947 | .5909921 | 0.21 | 0.831 | -1.032029 | 1.284618 |

13

## So sánh 3 mô hình

Regression Results

| | LPM<br>b/se | Logit<br>b/se | Probit<br>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 |

14

## Yêu cầu thực hành

- Sử dụng dữ liệu đã cung cấp và ước lượng 2 mô hình LPM và Logit. Trong đó biến phụ thuộc là SMOKE và biến độc lập chỉ bao gồm biến PRICE.
- Tính giá trị dự báo của p(SMOKE = 1)ứng với mô hình LPM khi PRICE thay đổi
- Tính giá trị dự báo p(SMOKE = 1) ứng với mô hình Logit khi PRICE thay đổi
- Vẽ đồ thị biểu diễn sự thay đổi các giá trị dự báo từ 2 mô hình trên theo các mức PRICE khác nhau.
- Mô tả và bình luận kết quả

15

## Nội dung buổi giảng tiếp theo

- Tính toán và giải thích tác động biên từ mô hình Logit
- Dự báo từ mô hình Logit
- Các kiểm định trước và sau hồi quy Logit

16