Hồi quy với Biến Định tính
(Regression with Qualitative Variables)

Lê Việt Phú
Trường Chính sách Công và Quản lý Fulbright

Ngày 31 tháng 12 năm 2017

Biến định tính là gì

- Còn được gọi là biến giả (dummy variable)
- Là biến mô tả trạng thái (nam/nữ, đi làm/đi học, làm nông/công chức)
- Có thể là biến nhị phân (có/không) hoặc biến nhóm (categorical variable - có nhiều hơn 2 trạng thái giá trị, ví dụ phương tiện đi lại là ô tô/xe máy/xe đạp/đi bộ)
- Đa số trường hợp các biến định tính không thể xếp được thứ bậc (ví dụ làm việc trong khu vực nhà nước/tư nhân/nước ngoài).
- Một số trường hợp biến định tính có thể xếp được thứ bậc, ví dụ bằng cấp cao nhất có được là gì, từ không có bằng cấp, bằng tiểu học, THCS,THPT, cao đẳng, đại học, thạc sỹ, tiến sỹ.

- Không nhầm lẫn với biến số đếm rời rạc, ví dụ biến số con cái trong gia đình không phải là biến định tính.
- Thống kê mô tả biến định tính khác với biến định lượng.
  - Cần xác định nhóm tham chiếu (baseline/reference group) và nhóm được tham chiếu. Ví dụ với biến giới tính thì có thể đặt nhóm tham chiếu là nữ và nhóm được tham chiếu là nam.
  - Giá trị trung bình diễn giải xác suất xảy ra một sự kiện.
  - Giá trị lớn nhất và nhỏ nhất không có ý nghĩa kinh tế.
  - Sai số chuẩn liên quan đến xác suất quan sát được sự kiện.
  - Hệ số tương quan mẫu (correlation coefficient) không có ý nghĩa.
  - Thường dùng biến định tính để phân tách và so sánh giữa các nhóm, ví dụ nhóm nam và nữ.

Sử lý biến định tính

Sử dụng lại bộ dữ liệu VHLSS 2010.

- Cần hiểu cách mã hóa biến trong bảng dữ liệu.
- Có thể gộp biến nhóm thành biến nhị phân.
- Có thể tách biến nhóm thành nhiều biến nhị phân.
- Bẫy biến giả (dummy trap): Một biến định tính có n giá trị thì có thể tách ra tối đa là $n-1$ biến giả. Nếu tách làm n biến giả đưa vào mô hình sẽ có hiện tượng đa cộng tuyến hoàn hảo.

Hồi quy với biến định tính

Ước lượng mô hình tỷ suất thu nhấp của đi học với các biến định tính là có gia đình, học trường công, làm nhà nước, làm nước ngoài, là công chức:

$$\begin{aligned} log(income) =& \beta_0 + \beta_1 yoeduc + \beta_2 yoexper + \beta_3 yoexpersq + \beta_4 married \\ &+ \beta_5 school + \beta_6 public + \beta_7 foreign + \beta_8 official + u \end{aligned}$$

Giải thích ý nghĩa của biến định tính

```
. reg lincome yoeduc yoexper yoexpersq married publicSchool public foreign official
```

Source	SS	df	MS		
Model	1753.70541	8	219.213176	Number of obs	= 7,552
Residual	4040.86526	7,543	.535710627	F(8, 7543)	= 409.20
				Prob > F	= 0.0000
				R-squared	= 0.3026
				Adj R-squared	= 0.3019
Total	5794.57067	7,551	.767391162	Root MSE	= .73192

lincome	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
yoeduc	.0926075	.0027428	33.76	0.000	.0872309	.0979841
yoexper	.061687	.0025081	24.60	0.000	.0567705	.0666035
yoexpersq	-.0012002	.0000488	-24.58	0.000	-.0012959	-.0011044
married	.0352395	.0221221	1.59	0.111	-.0081259	.078605
publicSchool	-.1145887	.0423549	-2.71	0.007	-.1976161	-.0315613
public	-.1042541	.0329488	-3.16	0.002	-.1688429	-.0396652
foreign	.4499482	.0363715	12.37	0.000	.37865	.5212464
official	.2705426	.0359373	7.53	0.000	.2000956	.3409897
_cons	8.493551	.0474837	178.87	0.000	8.40047	8.586633

.

Diễn giải ý nghĩa của tham số ước lượng đối với biến định tính

- Nếu biến phụ thuộc là **thu nhập** thì tham số ước lượng là tác động tăng thêm của nhóm được tham chiếu so với nhóm tham chiếu.

- Nếu biến phụ thuộc là **log của thu nhập** thì diễn giải tham số ước lượng tùy thuộc vào biến giải thích là biến liên tục hay biến rời rạc.
  - Với **biến liên tục**, ví dụ số năm đi học *yoeduc*, hệ số ước lượng là % tăng thêm của thu nhập. Ví dụ 1 năm đi học làm tăng thu nhập 9.26%.

- Với **biến rời rạc**, ví dụ các biến định tính, hoặc nếu có biến số con trong gia đình, thì:
  - Nếu β nhỏ, β có thể coi là phần trăm tăng thêm của biến phụ thuộc.
  - Công thức tính chính xác đối với tác động của biến rời rạc lên biến phụ thuộc **log(Y)** là:

$$\frac{Y_1 - Y_0}{Y_0} = e^{\beta} - 1$$

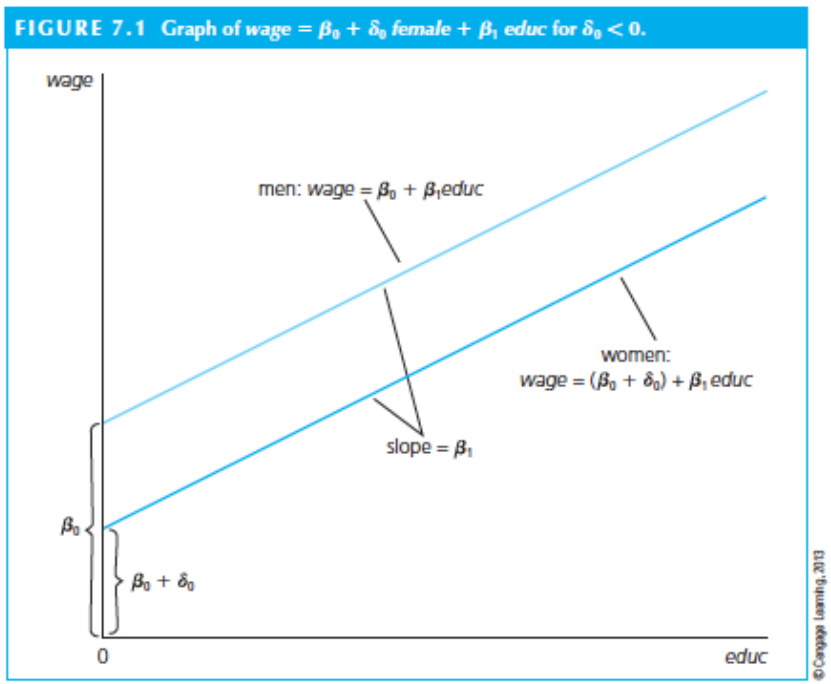
- Trong ví dụ trên:
  - Làm việc trong khu vực nước ngoài thu nhập cao hơn khu vực tư là: $2.718^{.45} - 1 = .5682$ hay 56.82% (chứ không phải là 45%).
  - Làm việc trong khu vực nhà nước thu nhập thấp hơn khu vực tư là: $2.718^{-.1043} - 1 = -.099$ hay 9.9%.
  - Nếu coi *yoeduc* là biến rời rạc thì với mỗi năm học tăng thêm thu nhập là $2.718^{.0926} - 1 = .097$ hay 9.7%.

Tung độ gốc và hệ số góc trong mô hình hồi quy

Với biến giới tính *sex* trong mô hình:

$$log(income) = \beta_0 + \beta_1 yoeduc + \beta_2 yoexper + \sigma_0 \mathbf{sex} + ... + u$$

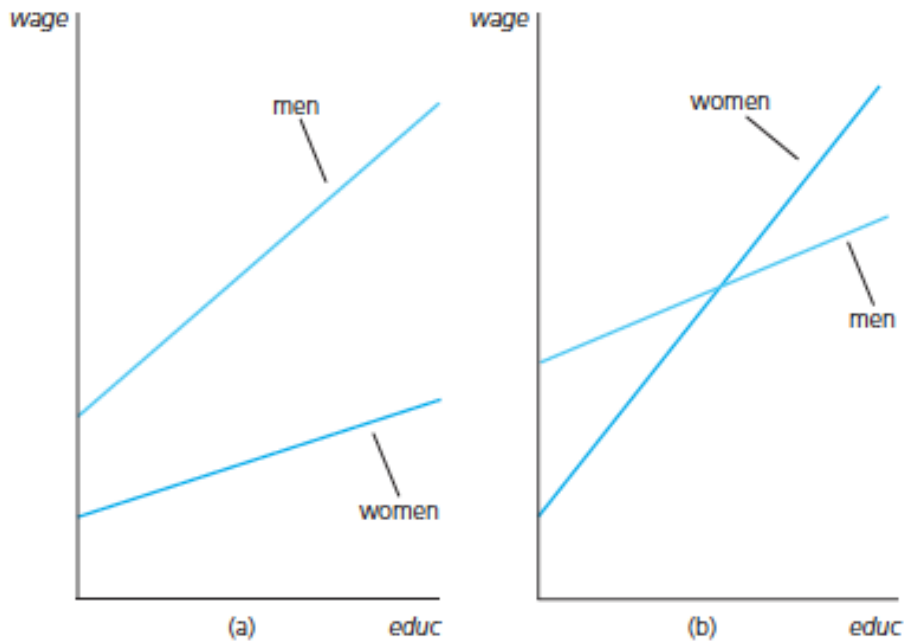
- Tung độ gốc là β_0 với nhóm nữ, và $\beta_0 + \sigma_0$ với nhóm nam
- Hệ số góc là β_1 giống nhau với cả hai nhóm (đường hồi quy song song)
- Nếu $\sigma_0 = 0$ thì hai đường hồi quy trùng nhau



Tung độ gốc và hệ số góc trong mô hình hồi quy với biến tương tác

$$log(income) = \beta_0 + \beta_1 yoeduc + \beta_2 yoexper + \sigma_0 \mathbf{sex} + \sigma_1 \mathbf{sex} * \mathbf{yoeduc} ... + u$$

- Tung độ gốc là β_0 với nhóm nữ, và $\beta_0 + \sigma_0$ với nhóm nam
- Hệ số góc là β_1 với nhóm nữ, và $\beta_1 + \sigma_1$ với nhóm nam.
- Hai đường hồi quy chỉ trùng nhau khi σ_0 và σ_1 đồng thời bằng 0.



Kiểm định khác biệt theo nhóm

- Tung độ gốc khác nhau $\Rightarrow$ t-test nếu $\sigma_0 = 0$
- Tung độ gốc và hệ số góc khác nhau $\Rightarrow$ F-test nếu $\sigma_0 = \sigma_1 = 0$
- Tất cả các tham số của hai nhóm khác nhau $\Rightarrow$ Chow test

Ôn tập các loại kiểm định

- Kiểm định đơn: $H_0 : \sigma_0 = 0$

$$t_{\hat{\sigma}_0} \sim t_{n-k-1}$$

- Kiểm định bội: $H_0 : \sigma_0 = \sigma_1 = 0$

$$F = \frac{(SSR_R - SSR_U)/q}{SSR_U/(n-k-1)} \sim F_{q,n-k-1}$$

- Kiểm định khác biệt nhóm: $H_0 : \sigma_0 = \sigma_1 = ... = \sigma_k = 0$

$$F = \frac{[SSR_p - (SSR_1 + SSR_2)]/(k+1)}{[SSR_1 + SSR_2]/(n-2(k+1))} \sim F_{k+1,n-2(k+1)}$$