Chương 4
Hồi quy với biến giả

4.1. Bản chất của biến giả

- Trong nhiều mô hình hồi quy, chúng ta cần xét biến giải thích (thậm chí biến phụ thuộc) là biến chất lượng (biến định tính).
- Ví dụ biến về:
  - Vùng địa lý, tôn giáo, giới tính, loai hình đào tạo, loại hình công việc, mùa, …
- Loại thông tin này có tính chất tự nhiên như là biến chỉ dẫn.
- Trong kinh tế lượng, các biến như thế gọi là biến giả.

Ví dụ: Lương giáo viên phổ thông

- Chúng ta có số liệu về lương của giáo viên 51 địa điểm.
- Chia ra ba loại
  - Phía bắc (21 điểm)
  - Nam (17 điểm)
  - Trung (13 điểm)
- Làm thế nào để đặt các biến giả này?

Ví dụ: Lương giáo viên phổ thông (tiếp)

- Đặt 3 biến giả
  - D1 = 1 nếu là vùng miền Trung; =0 nếu ngược lại.
  - D2 = 1 nếu là vùng miền Bắc; =0 nếu ngược lại.
  - D3 = 1 nếu là vùng miền Nam; =0 nếu ngược lại.
- Câu hỏi: Lương trung bình của các giáo viên các miền có bằng nhau không?
- Mô hình: ANOVA

Mô hình là:

$$Y_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + u_i$$

Ta có:

$$E(Y_i \mid D_{2i} = 0, D_{3i} = 0) = \beta_1$$

$$E(Y_i \mid D_{2i} = 1, D_{3i} = 0) = \beta_1 + \beta_2$$

$$E(Y_i \mid D_{2i} = 0, D_{3i} = 1) = \beta_1 + \beta_3$$

Một biểu diễn thay thế

$$Y_i = \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 D_{3i} + u_i$$

Chúng ta có:

$$E(Y_i \mid D_{1i} = 1, D_{2i} = 0, D_{3i} = 0) = \beta_1$$

$$E(Y_i \mid D_{1i} = 0, D_{2i} = 1, D_{3i} = 0) = \beta_2$$

$$E(Y_i \mid D_{1i} = 0, D_{2i} = 0, D_{3i} = 1) = \beta_3$$

$D_1+D_2+D_3=1$ nên có ĐCT.

4.2. Hồi quy với một biến lượng và một biến chất.

- Biến chất có hai phạm trù $Y_i=\beta_1+ \beta_2D_i+ \beta_3X_i+U_i$

Y_i: Tiền lương hàng tháng của 1 công nhân i.

X_i: bậc thợ của công nhân i.

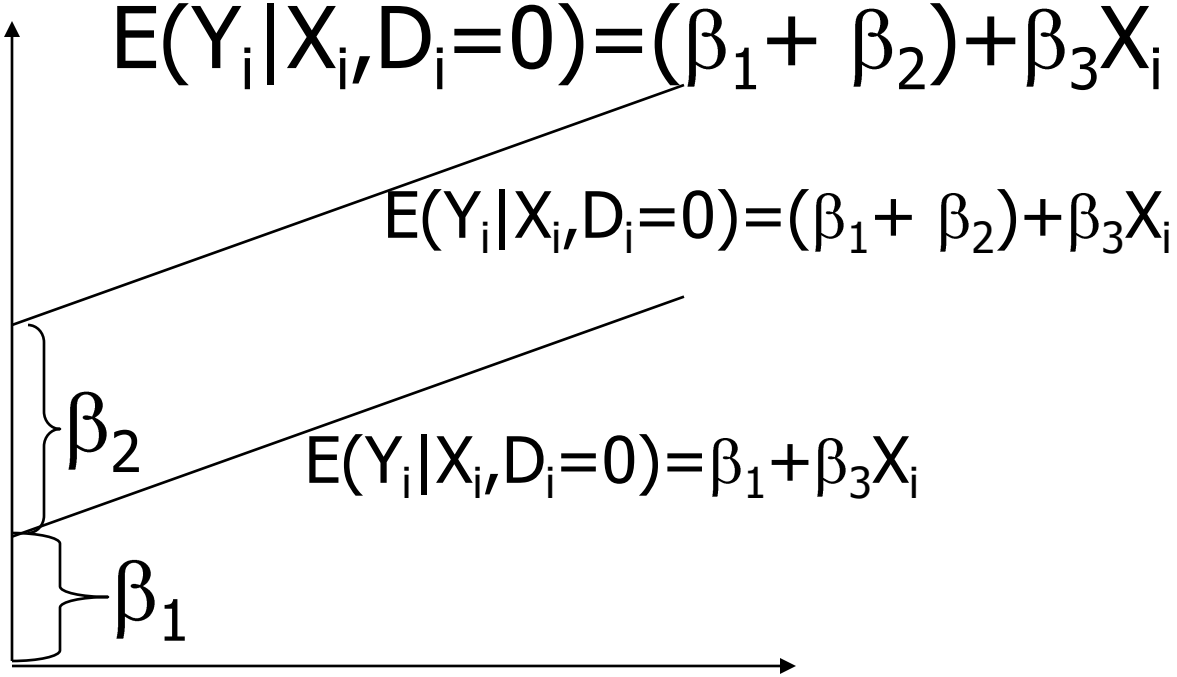
$D_i =$ { 1 nếu công nhân i làm việc KV tư nhân.
0 nếu công nhân i làm việc KV nhà nước.

- Với giả thiết $E(U_i)=0$, Thì:

-Tiền lương trung bình của công nhân cơ khí làm việc trong KV nhà nước:

$$E(Y_i|X_i,D_i=0)=\beta_1+ \beta_3X_i$$

-Tiền lương trung bình của công nhân cơ khí làm việc trong KV tư nhân:



Vậy:

- Tốc độ tăng lương trong cả hai TH như nhau β_3
- Nếu $\beta_2 \neq 0$ thì tiền lương CN hai KV khác nhau.

Nếu $\beta_2=0$ thì tiền lương CN hai KV như nhau.

- Biến chất có nhiều hơn hai phạm trù
- Nếu MH có n phạm trù thì đưa vào MH n-1 biến giả làm biến giải thích.
- Xét mô hình $Y_i=\beta_1+ \beta_2D_{2i} + \beta_3D_{3i} + \beta_4X_i+U_i$
- Y: Thu nhập hàng năm của một GV đại học.
- X: Tuổi nghề của giáo viên.

$D_1=$ { 1 nếu GV thuộc trường ĐH miền Bắc.
0 với các trường hợp khác.

$D_2=$ { 1 nếu GV thuộc trường ĐH miền Nam.
0 với các trường hợp khác.

- Giảng viên thuộc trường đại học miền trung là phạm trù cơ sở ứng với $D_1=0$, $D_2=0$.

Thu nhập trung bình của giảng viên đại học các miền là:

- Miền trung:

$$E(Y_i|D_1=0,D_2=0,X_i)=\beta_1+\beta_4X_i$$

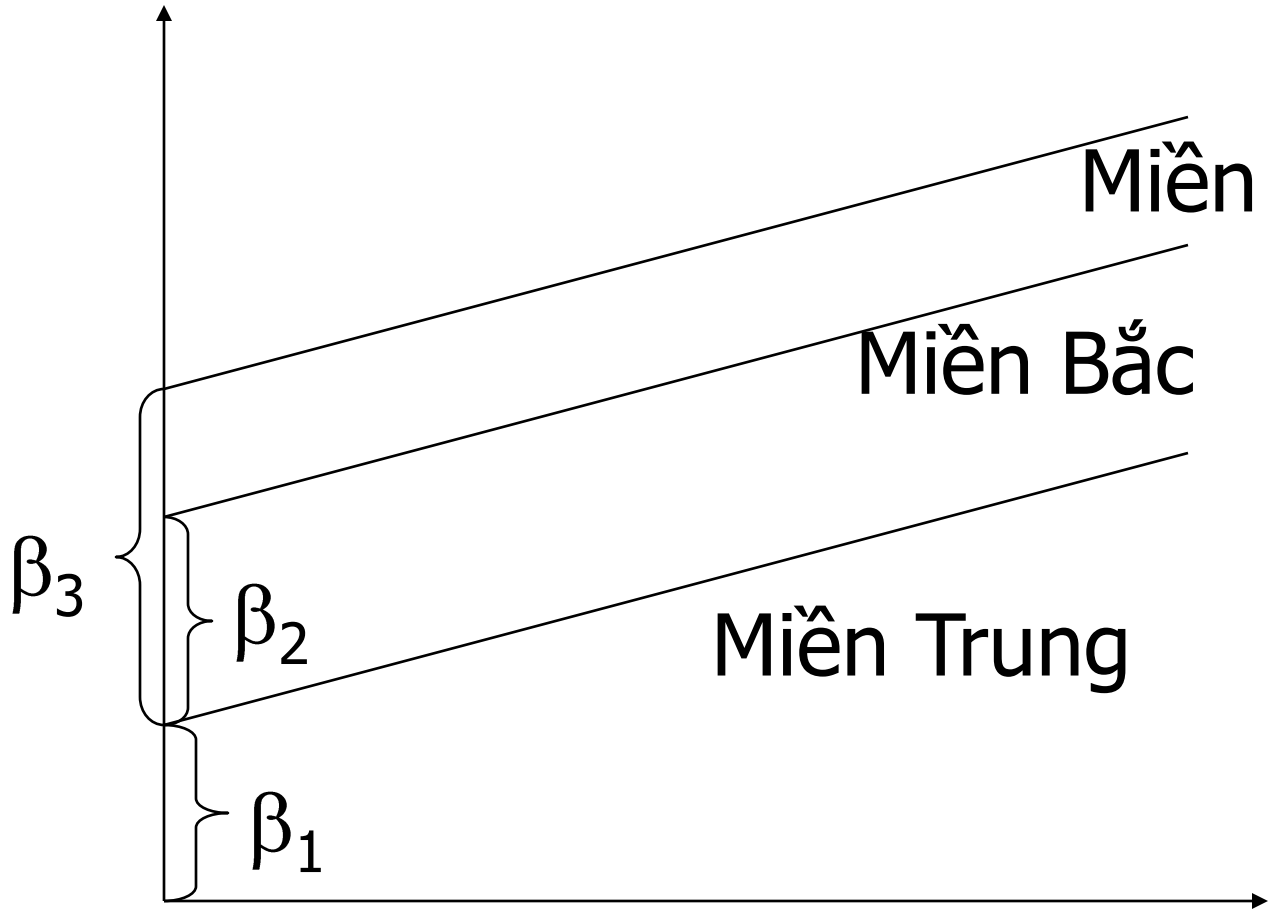
- Miền Bắc

$$E(Y_i|D_1=1,D_2=0,X_i)=(\beta_1+\beta_2)+\beta_4X_i$$

- Miền Nam

$$E(Y_i|D_1=0,D_2=1,X_i)=(\beta_1+\beta_3)+\beta_4X_i$$

Nếu $\beta_1 >0$, ta có minh họa sau:



Kđ gì?

- TN miền Bắc và Trung như nhau ko?
- TN miền Nam và Trung như nhau ko?
- TN miền Nam và Bắc như nhau ko?

β_2 cho biết thu nhập trung bình của GV thay đổi khi chuyển từ miền trung tới miền Bắc.

β_3 cho biết thu nhập trung bình của GV thay đổi khi chuyển từ miền trung tới miền Nam.

4.2. Hồi quy với một biến lượng và hai biến chất.

- Xét MH: $Y_i=\beta_1+ \beta_2D_{2i} + \beta_3D_{3i} + \beta_4D_{4i} + \beta_5X_i+U_i$

Y, X là thu nhập và tuổi nghề của giảng viên.

D_{1i}=1 nếu i là GV miền Bắc, =0 nếu ngược lại.

D_{2i}=1 nếu i là GV miền Nam, =0 nếu ngược lại.

D_{3i}=1 nếu i là GV Nam, =0 nếu ngược lại.

(Vùng có 3 phạm trù, giới tính có 2 phạm trù)

- G/s $E(U_i)=0$, ta có thể thu được kết quả sau:

- Thu nhập trung bình của giảng viên Nam:

- miền Trung: $E(Y_i|D_1=0, D_2=0, D_3=1,X_i)$
- miền Bắc: $E(Y_i|D_1=1, D_2=0, D_3=1,X_i)$
- miền Nam: $E(Y_i|D_1=0, D_2=1, D_3=1,X_i)$

- Thu nhập trung bình của giảng viên Nữ:

- miền Trung: $E(Y_i|D_1=0, D_2=0, D_3=0,X_i)$
- miền Bắc: $E(Y_i|D_1=1, D_2=0, D_3=0,X_i)$
- miền Nam: $E(Y_i|D_1=0, D_2=1, D_3=0,X_i)$

- Các câu hỏi mở kđ.

4.4. So sánh hai hồi quy

1. Tư tưởng cơ bản

Ví dụ: xét quan hệ tiết kiệm và thu nhập trước và sau chuyển đổi kinh tế

- Hồi quy thời kỳ trước cải tổ

$$Y_i=\lambda_1+ \lambda_2X_i+u_{1i} \qquad \text{(có } n_1 \text{ quan sát)}$$

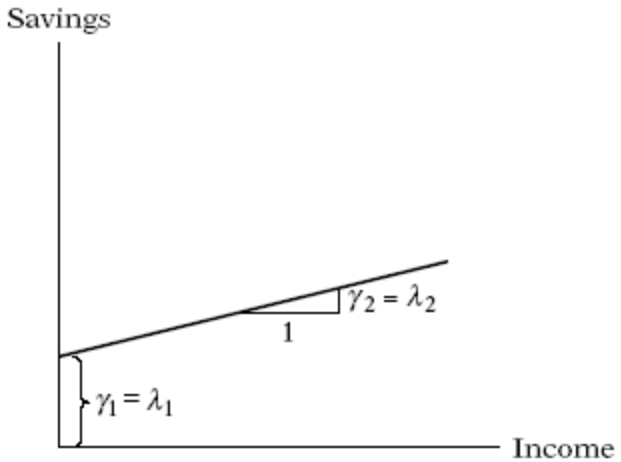
- Hồi quy thời kỳ sau cải tổ

$$Y_j=\gamma_1+ \gamma_2X_j+u_{1i} \qquad \text{(có } n_2 \text{ quan sát)}$$

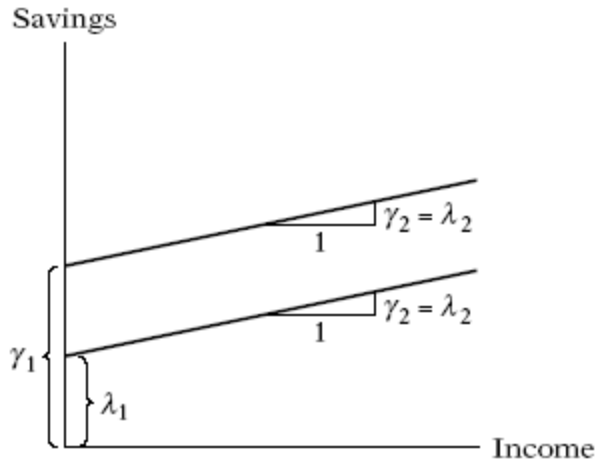
Hai hồi quy đó có khác nhau không? Sự khác nhau ở hệ số chặn, hệ số góc hay cả hai?

- Có 4 khả năng xảy ra như sau:

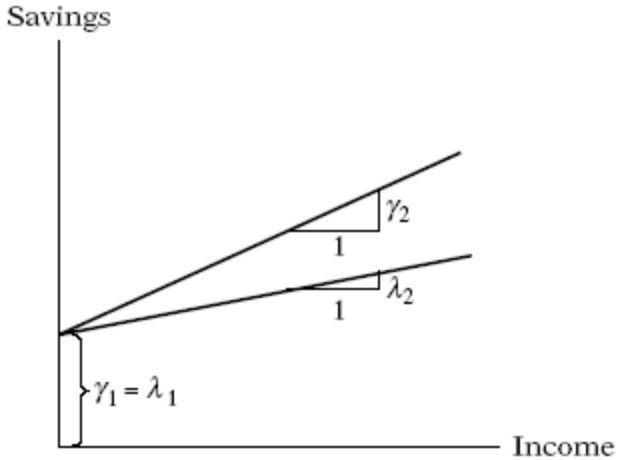
Interpretation of the possible regressions



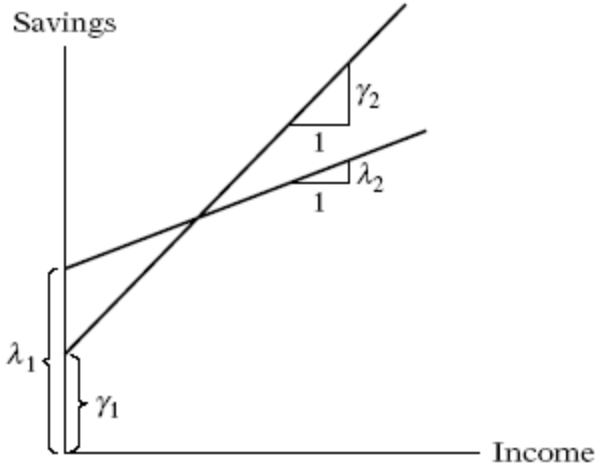
(*a*) Coincident regressions



(*b*) Parallel regressions



(*c*) Concurrent regressions



(*d*) Dissimilar regressions

2. So sánh hai hồi quy - Kiểm định Chow

- Giả thiết u_{1i}, u_{2i} có phân phối chuẩn.

  u_{1i}, u_{2i} có phân phối độc lập.

- Thủ tục:

- Bước 1: Kết hợp số liệu ca hai thời kỳ, có $n=n_1+n_2$ quan sát. Hồi quy mô hình:

$$Y_i=\beta_1+\beta_2X_{2i}+u_i \qquad (1)$$

Thu được RSS

$df=n-k=n_1+n_2-k$ (*k là tham số (1)*)

- Bước 2: Ước lượng riêng từng hồi quy

+ Hồi quy thời kỳ trước cải tổ

$$Y_i=\lambda_1+\lambda_2 X_i+u_{1i}$$ (có n_1 quan sát)

+ Hồi quy thời kỳ sau cải tổ

$$Y_j=\gamma_1+\gamma_2 X_j+u_{1i}$$ (có n_2 quan sát)

Thu được RSS_1 có df=n_1-k, RSS_2 có df=n_2-k.

- Đặt $\overline{RSS} = RSS_1 + RSS_2$ với df=n_1+n_2-2k

- Bước 3: Tính

$$F = \frac{(\overline{RSS} - RSS)/k}{RSS/(n_1+n_2-2k)} \sim F(k, n_1+n_2-2k)$$

- Nếu F >F_α thì bác bỏ giả thiết H_0.
- (H_0: Hai hồi quy là như nhau)

3. So sánh hai hồi quy - thủ tục biến giả

- Thục tục biến giả có thể gộp n_1, n_2 quan sát lại với nhau và ước lượng:

$$Y_i=\beta_1+\beta_2 D_i+\beta_3 X_i+\beta_4(D_i X_i)+u_i$$

- $D_i = \begin{cases} 1 \text{ nếu i trước thời kỳ chuyển đổi.} \\ 0 \text{ nếu ngược lại.} \end{cases}$
- Với giả thiết $E(u_i)=0$.

- Tiết kiệm trung bình trước chuyển đổi:

$$E(Y_i|D_i=0,X_i) =\beta_1+ \beta_3 X_i$$

- Tiết kiệm trung bình sau chuyển đổi:

$$E(Y_i|D_i=1,X_i) =(\beta_1+ \beta_2)+ (\beta_3+\beta_4)X_i$$

4.5. Ảnh hưởng của tương tác giữa các biến giả

- Xét $Y_i=\beta_1+\beta_2D_{2i}+\beta_3D_{3i}+\beta_4X_i+u_i$ (1)

  Y, X là chi tiêu quần áo và thu nhập.

  D_{2i}=1 nếu i là nữ, =0 nếu ngược lại.

  D_{3i}=1 nếu i là sinh viên, =0 nếu ngược lại.

- MH (1): ảnh hưởng chênh lệch của nữ sinh viên và nam sinh viên luôn bằng ảnh hưởng chênh lệch của nữ công nhân viên và nam công nhân viên. Trong thực tế, nhiều khi điều đó không đúng, tức là có sự tương tác giữa hai biến giả.

4.5. Ảnh hưởng của tương tác giữa các biến giả (tiếp)

- $Y_i=\beta_1+\beta_2D_{2i}+\beta_3D_{3i}+\beta_4(D_{2i}D_{3i})+\beta_5X_i+u_i$ (2)
- Chi tiêu trung bình quần áo của nữ sinh là:

$E(Y|D_2=1,D_3=1,X_i) =(\beta_1+ \beta_2+\beta_3+\beta_4)+ \beta_5X_i$

- Chi tiêu trung bình quần áo của nam sinh là:

$E(Y|D_2=0,D_3=1,X_i) =(\beta_1+\beta_3)+ \beta_5X_i$

- Chi tiêu trung bình quần áo của nữ công nhân:

$E(Y|D_2=1,D_3=0,X_i) =(\beta_1+ \beta_2)+ \beta_5X_i$

- Chi tiêu trung bình quần áo của nam côngnhân:

$E(Y|D_2=0,D_3=0,X_i) =\beta_1+ \beta_5X_i$

4.6. Sử dụng biến giả trong phân tích mùa

- Mô hình:

$$Y_i=\beta_1+\beta_2D_{2i}+\beta_3D_{3i}+\beta_4D_{4i}+\beta_5X_i+$$
$$+\beta_6(D_{2i}X_i)+\beta_7(D_{3i}X_i)+\beta_8(D_{4i}X_i)+u_i$$

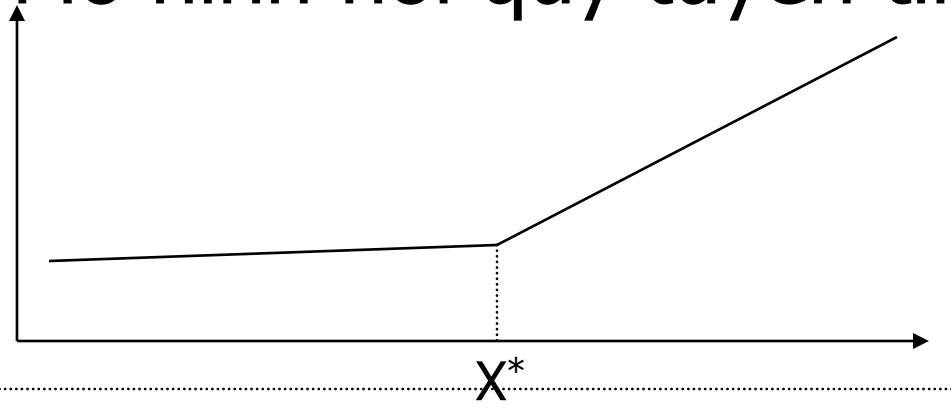
D_{2i}=1 nếu i là mùa xuân, =0 nếu ngược lại.

D_{3i}=1 nếu i là mùa hạ, =0 nếu ngược lại.

D_{4i}=1 nếu i là mùa đông, =0 nếu ngược lại.

4.7. Hồi quy tuyến tính từng khúc

- Mô hình hồi quy tuyến tính từng khúc



- $Y_t=\beta_1+ \beta_2X_t+ \beta_3(X_t-X_{t0})D_t+u_t$

Y_t: tiêu dùng, X_t: thu nhập.

T_0: năm bắt đầu chuyển từ cơ chế có kế hoạch sang cơ chế thị trường.

D_t= 1 nếu $t>t_0$, =0 nếu ngược lại.

Next time

- Introduction to extensions of the classical linear regression model
- Multicollinearity (Chapter 10)

Equation: UNTITLED Workfile: CH4BT1

View | Procs | Objects | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: Q
Method: Least Squares
Date: 02/01/07 Time: 09:39
Sample: 97M1 to 99M3
Included observations: 27

Variable	Coefficient	Std. Error	t-Statistic	Prob.
PG	-7.0673	.20832	33.9252	.000
D	106.0104	98.5409	1.0758	.293
DPG	.278299	.078845	2.6307	.012
C	2403.548	564.049	4.2609	.000

R-squared	.99252	Mean dependent var	1831.4
Adjusted R-squared	.99154	S.D. dependent var	451.9370
S.E. of regression	41.5680	Akaike info criterion	
Sum squared resid	39741.7	Schwarz criterion	
Log likelihood	-136.78	F-statistic F(3,23)	1016.8
Durbin-Watson stat	1.9506	Prob(F-statistic)	0.000