Chương 6

Biến giả trong phân tích hồi quy

TS. Đinh Thị Thanh Bình

Khoa Kinh Tế Quốc Tế- Đại học Ngoại thương

6.1 KHÁI NIỆM

- Biến định lượng: các giá trị quan sát được thể hệ bằng con số
- Biến định tính: thể hiện một số tính chất nào đó
- Để đưa những thuộc tính của biến định tính vào mô hình hồi quy, cần lượng hóa chúng => sử dụng biến giả (binary, zero-one, dummy variables)

6.1 Chỉ có một biến giả trong mô hình

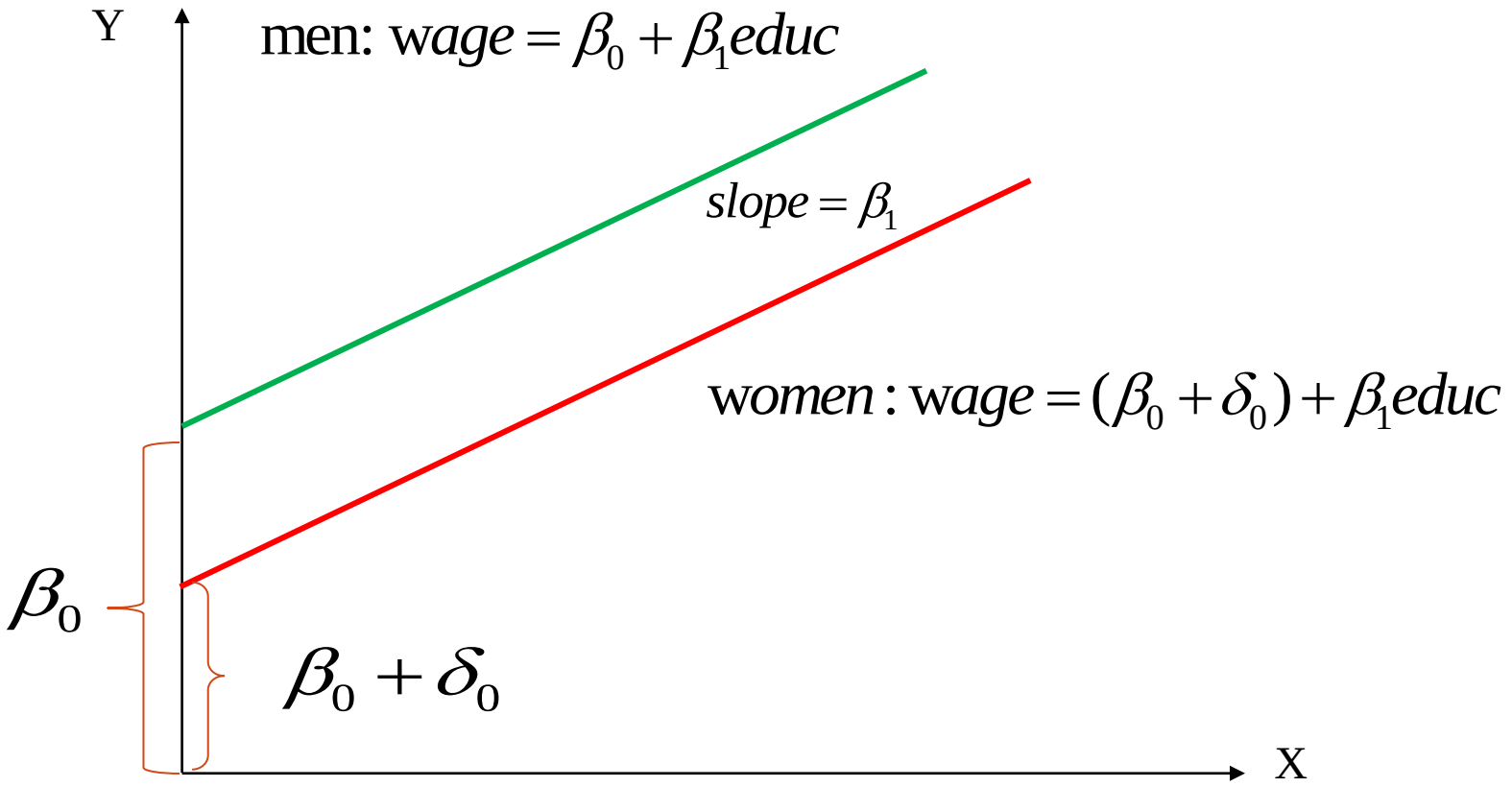
$$wage = \beta_0 + \delta_0 female + \beta_1 educ + u \quad \textbf{(1)}$$

$$\delta_0 = E(wage \mid female = 1, educ) - E(wage \mid female = 0, educ)$$

Female = 1 tương ứng với nữ giới, female = 0 tương ứng với nam

$$\delta_0 = E(wage \mid female, educ) - E(wage \mid male, educ)$$

Nghĩa là: với trình độ học vấn như nhau, sự khác biệt về lương, δ_0 , là do sự khác biệt về giới tính.



Hình 6.1: Đồ thị của $wage = \beta_0 + \delta_0\, female + \beta_1 educ + u; \delta_0 < 0$

- Độ dốc như nhau do không phụ thuộc vào educ.
- Hệ số tự do khác nhau (intercept)

Chú ý: Một chỉ tiêu chất lượng có **n** phạm trù (thuộc tính) khác nhau thì dùng **n-1** biến giả

Ví dụ: giới tính có 2 phạm trù (male, female) → dùng 1 biến giả

- Ở ví dụ trên, male được gọi là **phạm trù cơ sở** (base group)

- Nếu male là phạm trù cơ sở thì có mô hình như sau:

$$wage = \alpha_0 + \lambda_0 female + \beta_1 educ + u$$

- Các phương pháp kiểm định giả thuyết thống kê với biến giả giống như với biến định lượng.

6.2. Sử dụng nhiều biến giả trong mô hình

-Chúng ta có thể đưa nhiều hơn 1 biến giả vào phương trình hồi qui:

$$\text{w}age = \beta_0 + \delta_0 female + \delta_1 \text{m}arried + \beta_1 educ + u \qquad (2)$$

-Tuy nhiên, một hạn chế của phtr này là: ảnh hưởng của biến giả "married" được giả định là giống nhau cho cả nam và nữ.

- Chúng ta sẽ khắc phục hạn chế này bằng cách cho phép có sự khác biệt về lương giữa 4 nhóm: married man, married woman, single man, single woman

-Nếu chọn phạm trù cơ sở là single men, khi đó phtr hồi qui mẫu:

$$wage = \beta_0 + \delta_0 marrmale + \delta_1 marrfemale + \delta_2 \sin gfem + \beta_1 educ + u \quad \textbf{(3)}$$

Chú ý: chúng ta phải bỏ biến female, married ra khỏi mô hình trên

Thực hành với file WAGE1

- Ví dụ: từ file WAGE1

$$\log((\text{w}age) = 0.321 + 0.213marrmale - 0.198marrfem$$

$$-0.110\sin gfem + \beta_4 educ$$

Chú ý:

-Hệ số ở các biến giả trên đo sự khác biệt về thu nhập tương đối so với nhóm cơ sở - single male.

- Nam giới có gia đình được dự đoán có thu nhập cao hơn nam giới độc thân là 21.3%, ceteris paribus.

- Ảnh hưởng của nhóm cơ sở - single male- được thể hiện ở hệ số tự do (0.321).

-Nữ giới độc thân có thu nhập cao hơn nữ giới kết hôn là 8.8% ← (=-0.110-(-0.198) = 0.088)

-Tuy nhiên chúng ta không thể kiểm định sự khác biệt này có ý nghĩa thống kê hay không. Nếu muốn kiểm định, chúng ta phải chạy lại mô hình với một trong hai nhóm trên là nhóm cơ sở.

- Ví dụ: chọn married woman làm nhóm cơ sở

$$\log((\text{w}age) = 0.123 + 0.411marrmale + 0.198\sin gmale$$
$$+0.088\sin gfem + ...$$

- Trường hợp sử dụng biến giả đối với thông tin được sắp xếp theo thứ tự (ordinal information)

- Ví dụ: loại hình sở hữu doanh nghiệp

6.3. Biến tương tác liên quan đến 2 biến giả

- Ở phần trên, chúng ta chỉ ra 4 phạm trù dựa trên tình trạng hôn nhân và giới tính.

$$wage = \beta_0 + \delta_0 marrmale + \delta_1 marrfemale + \delta_2 \sin gfem + \beta_1 educ + u \quad (3)$$

- Tuy nhiên, mô hình trên có thể viết lại bằng cách cho biến tương tác giữa female và married vào mô hình:

$$wage = \beta_0 + \delta_0\, female + \delta_1 married + \delta_2\, female.married + ... + u \quad (4)$$

- Mô hình (4) cho biết ảnh hưởng của tình trạng hôn nhân khác nhau đối với nam và nữ, giống mô hình (3)

- Ví dụ: 4 phạm trù dựa trên tình trạng hôn nhân và giới tính.

$$\lg(wage) = 0.321 - 0.110\,female + 0.213married - 0.301\,female.married + \beta_4 educ$$

- Nếu female = 0 và married = 0 ➔ tương ứng với nhóm single male (nhóm cơ sở) ➔ mức độ ảnh hưởng của nhóm này là 0.321

- female = 0 và married = 1 ➔ tương ứng với nhóm married man ➔ mức độ ảnh hưởng của nhóm này là : 0.321 + 0.213

➔ Nam giới có gia đình thu nhập cao hơn nam giới độc thân 21,3%.

6.4. Biến tương tác liên quan đến 1 biến giả và 1 biến định lượng

- Xem xét liệu ảnh hưởng của giáo dục đến thu nhập có giống nhau đối với nam và nữ.

$$wage = \beta_0 + \delta_0\, female + \beta_1 educ + \delta_1\, female.educ + u$$

➔ $$wage = (\beta_0 + \delta_0\, female) + (\beta_1 + \delta_1\, female) educ + u \qquad (5)$$

-Nếu female = 0, hệ số tự do của male là β_0 và độ dốc là β_1

-Nếu female = 1, hệ số tự do của female là $\beta_0 + \delta_0$ và độ dốc là $\beta_1 + \delta_1$

. δ_0 miêu tả sự khác nhau giữa hệ số tự do giữa male và female

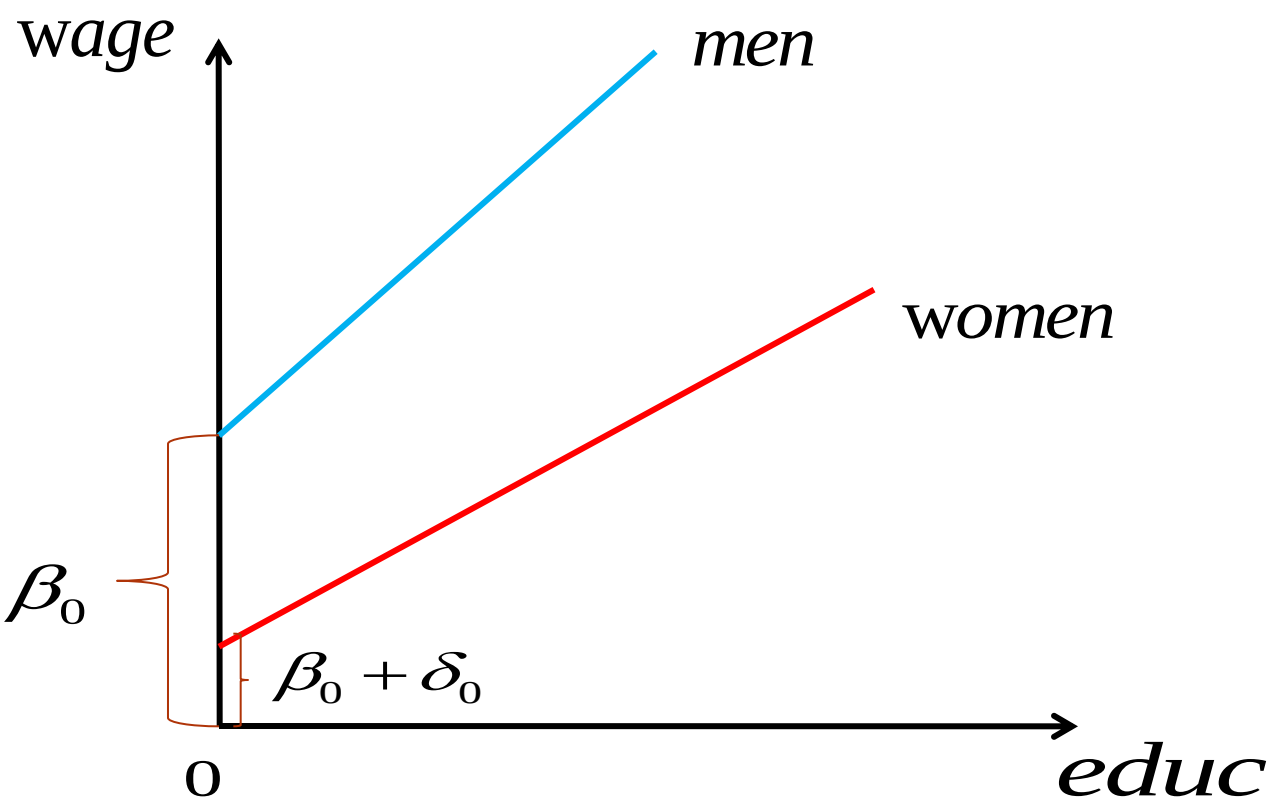
. δ_1 miêu tả sự khác nhau về ảnh hưởng của giáo dục đến thu nhập giữa male và female

TH1: $wage = (\beta_0 + \delta_0\, female) + (\beta_1 + \delta_1\, female)educ + u$

$$\delta_0 < 0, \delta_1 < 0$$

Nữ thu nhập thấp hơn nam ở tất cả các trình độ học vấn và khoảng cách này tăng khi trình độ học vấn càng cao.

Hình 6.2



TH2: $$wage = (\beta_0 + \delta_0\, female) + (\beta_1 + \delta_1\, female)educ + u$$

$$\delta_0 < 0, \delta_1 > 0$$

-Hệ số tự do của nữ thấp hơn của nam giới nhưng độ dốc của trình độ học vấn cho nữ lại lớn hơn nam. Nghĩa là:

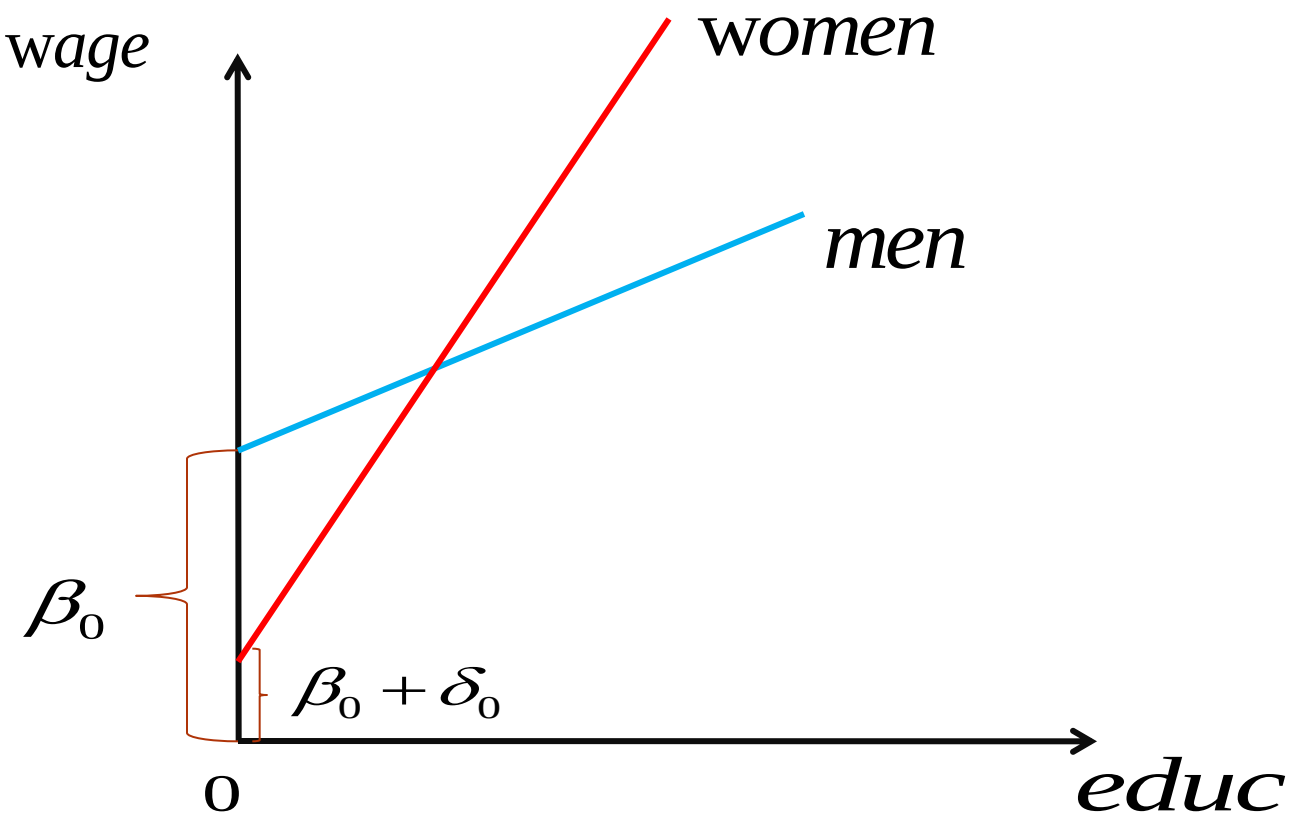
- Nữ thu nhập thấp hơn nam ở trình độ học vấn thấp, nhưng khoảng cách hẹp dần khi trình độ học vấn tăng.

- Ở một điểm nào đó, nữ giới thu nhập cao hơn nam giới với trình độ học vấn như nhau.

$$wage = (\beta_0 + \delta_0\, female) + (\beta_1 + \delta_1\, female)educ + u$$

$$\delta_0 < 0, \delta_1 > 0$$

Hình 6.3



Xây dựng giả thuyết thống kê:

Giả thuyết 1: ảnh hưởng của trình độ học vấn (return to education) đến thu nhập là như nhau đối với cả nam và nữ

$$H_0 : \delta_1 = 0$$

- Không có ràng buộc nào với δ_0 , nghĩa là dưới giả thuyết này sự khác nhau về thu nhập giữa nam và nữ là có thể, nhưng sự ảnh hưởng của trình độ học vấn là như nhau. (Hình 6.1)

- Sử dụng t-test

Giả thuyết 2: mức lương trung bình là như nhau cho cả nam và nữ với trình độ học vấn như nhau.

$$H_0 : \delta_0 = 0, \delta_1 = 0$$

- Sử dụng F-test

6.5 Ví dụ về ứng dụng sử dụng biến giả

Số liệu tiết kiệm và thu nhập cá nhân ở nước Anh từ 1946-63 (triệu pounds)

TK I	Tiết kiệm	Thu nhập	TK II	Tiết kiệm	Thu nhập
1946	0.36	8.8	1955	0.59	15.5
1947	0.21	9.4	1956	0.9	16.7
1948	0.08	10	1957	0.95	17.7
1949	0.2	10.6	1958	0.82	18.6
1950	0.1	11	1959	1.04	19.7
1951	0.12	11.9	1960	1.53	21.1
1952	0.41	12.7	1961	1.94	22.8
1953	0.5	13.5	1962	1.75	23.9
1954	0.43	14.3	1963	1.99	25.2

Mục tiêu: Kiểm tra hàm tiết kiệm có thay đổi cấu trúc giữa 2 thời kỳ hay không.

Cách 1: Lập hai mô hình tiết kiệm ở 2 thời kỳ

- Thời kỳ tái thiết: 1946-54: $Y_i = \alpha_1 + \alpha_2 X_i + u_{1i}$

- Thời kỳ hậu tái thiết: 1955-63: $Y_i = \lambda_1 + \lambda_2 X_i + u_{2i}$

- Và kiểm định các trường hợp sau

$$\begin{cases} \alpha_1 = \lambda_1 \\ \alpha_2 = \lambda_2 \end{cases} \quad \begin{cases} \alpha_1 = \lambda_1 \\ \alpha_2 \neq \lambda_2 \end{cases} \quad \begin{cases} \alpha_1 \neq \lambda_1 \\ \alpha_2 = \lambda_2 \end{cases} \quad \begin{cases} \alpha_1 \neq \lambda_1 \\ \alpha_2 \neq \lambda_2 \end{cases}$$

Cách 2: Sử dụng biến giả

B1. Lập hàm tiết kiệm tổng quát của cả 2 thời kỳ

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{\beta}_3 Z_i + \hat{\beta}_4 X_i Z_i + u_i$$

Với $n = n_1 + n_2$

Z = 1 quan sát thuộc thời kỳ tái thiết

Z = 0 quan sát thuộc thời kỳ hậu tái thiết

B2. Kiểm định giả thuyết H_0: $\beta_3=0$

Nếu chấp nhận H_0: loại bỏ Z ra khỏi mô hình

B3. Kiểm định giả thuyết H_0: $\beta_4=0$

Nếu chấp nhận H_0: loại bỏ Z_iX_i ra khỏi mô hình

Kết quả hồi quy theo mô hình như sau

$$Y_i = -1,75 + 0,15045X_i + 1,4839Z_i - 0,1034X_iZ_i + u_i$$

t =	(-5,27)	(9,238)	(3,155)	(-3,109)
p =	(0,000)	(0,000)	(0,007)	(0,008)

$$Y_i = (-1,75 + 1,4839Z_i) + (0,15045 - 0,1034Z_i)X_i + u_i$$

Nhận xét

- Tung độ gốc chênh lệch và hệ số góc chênh lệch có ý nghĩa thống kê
- Các hồi quy trong hai thời kỳ là khác nhau

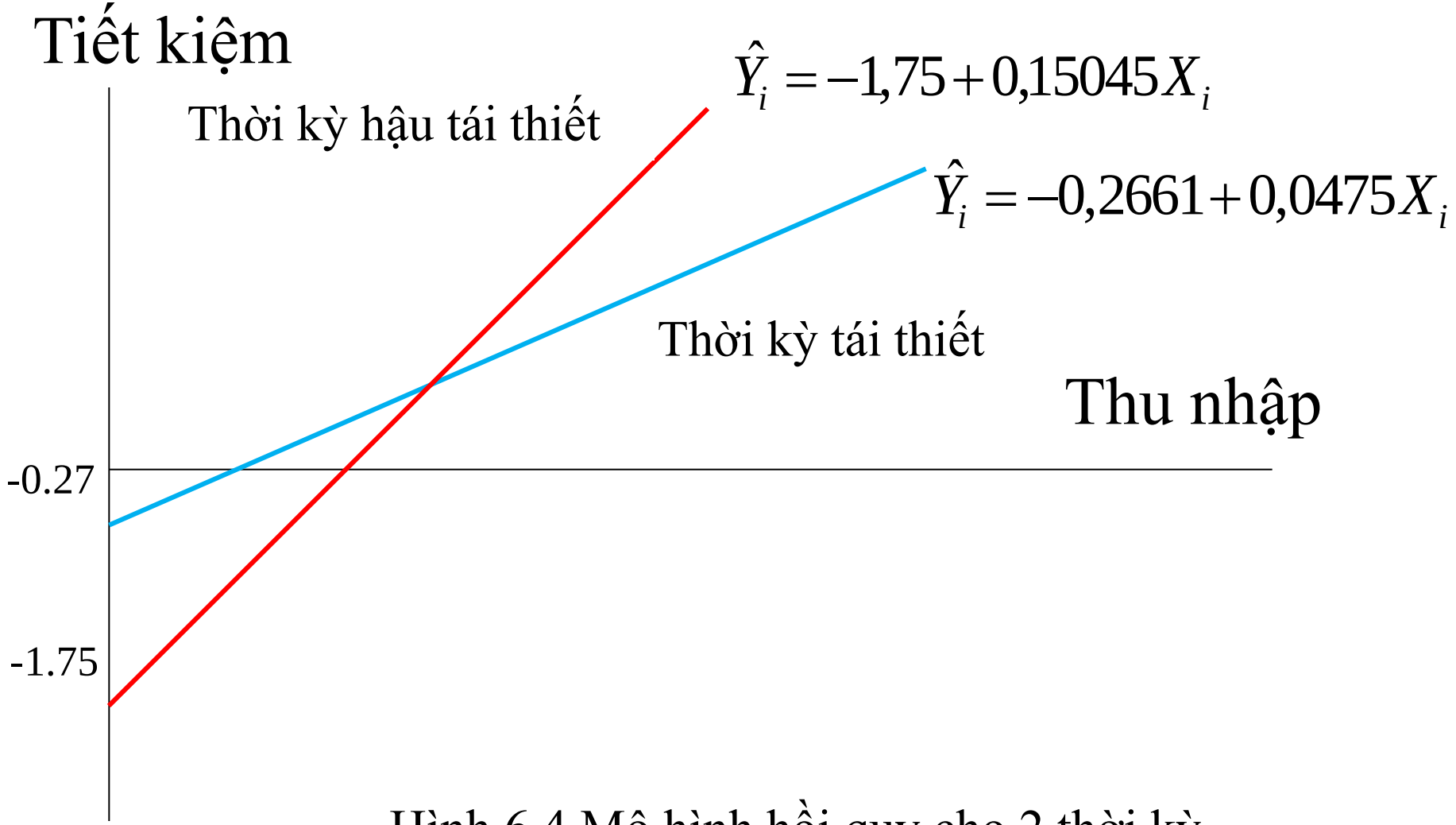
Thời kỳ tái thiết: Z = 1

$$\hat{Y}_i = -1{,}75 + 0{,}15045 X_i + 1{,}4839 - 0{,}1034 X_i$$

$$\hat{Y}_i = -0{,}2661 + 0{,}0475 X_i$$

Thời kỳ hậu tái thiết: Z = 0

$$\hat{Y}_i = -1{,}75 + 0{,}15045 X_i$$



Hình 6.4 Mô hình hồi quy cho 2 thời kỳ