Chương 5
Hồi qui với biến giả

I. Bản chất của biến giả- Mô hình trong đó các biến độc lập đều là biến giả

Biến định tính thường biểu thị các mức độ khác nhau của một tiêu thức thuộc tính nào đó.

Ví dụ : …

Để lượng hoá được biến định tính, trong phân tích hồi qui người ta sử dụng kỷ thuật biến giả.

Ví dụ 1 : Một cty sử dụng 2 công nghệ (CN) sản xuất (A, B). Năng suất của mỗi CN là đại lượng ngẫu nhiên phân phối chuẩn có phương sai bằng nhau, kỳ vọng khác nhau. Hãy lập mô hình mô tả quan hệ giữa năng suất của cty với việc sử dụng CN sản xuất.

Mô hình : $Y_i = \beta_1 + \beta_2 Z_i + U_i$

Trong đó : Y : năng suất, Z : biến giả

$$Z_i = \begin{cases} 1 \text{ nếu sử dụng CN A} \\ 0 \text{ nếu sử dụng CN B} \end{cases}$$

Ta có :

$E(Y_i/Z_i = 0) = \beta_1$: năng suất trung bình của CN A.

$E(Y_i/Z_i = 1) = \beta_1 + \beta_2$: năng suất trung bình của CN B.

$\Rightarrow \beta_2$: chênh lệch năng suất giữa CN B và A.

Giả thiết $H_0 : \beta_2 = 0$ ($\Leftrightarrow$ giữa CN A và CN B không có khác biệt về năng suất).

* Giả sử tiến hành khảo sát năng suất của CN A và CN B trong vòng 10 ngày, người ta thu được số liệu sau :

CN sử dụng	**B**	**A**	**A**	**B**	**B**	**A**	**B**	**A**	**A**	**B**
Năng suất	**28**	**32**	**35**	**27**	**25**	**37**	**29**	**34**	**33**	**30**

Năng suất (đvt : Tấn/ ngày)

Dùng mẫu số liệu trên, hồi qui mô hình đang xét, ta có : $\hat{Y}_i = 27{,}8 + 6{,}4 Z_i$

Ví dụ 2 : Tương tự ví dụ 1, nhưng công ty có 3 CN sản suất (A, B, C).

Mô hình : $Y_i = \beta_1 + \beta_2 Z_{1i} + \beta_3 Z_{2i} + U_i$

Trong đó : Y - năng suất, Z_1, Z_2 : biến giả

$$Z_{1i} = \begin{cases} 1 : \text{sử dụng CN A} \\ 0 : \text{không sử dụng CN A} \end{cases}$$

$$Z_{2i} = \begin{cases} 1 : \text{sử dụng CN B} \\ 0 : \text{không sử dụng CN B} \end{cases}$$

Ta có :

$E(Y_i/Z_{1i} = 1, Z_{2i} = 0) = \beta_1 + \beta_2$: năng suất trung bình của CN A.

$E(Y_i/ Z_{1i} = 0, Z_{2i} = 1) = \beta_1 + \beta_3$: năng suất trung bình của CN B.

$E(Y_i/ Z_{1i} = 0, Z_{2i} = 0) = \beta_1$: năng suất trung bình của CN C.

$\Rightarrow \beta_2$: chênh lệch năng suất giữa CN A và C.

$\Rightarrow \beta_3$: chênh lệch năng suất giữa CN B và C.

- Chú ý :

- Một biến định tính có m mức độ (m phạm trù) thì cần sử dụng (m-1) biến giả đại diện cho nó.
- Phạm trù được gán giá trị 0 được xem là phạm trù cơ sở (việc so sánh được tiến hành với phạm trù này).

II. Hồi qui với biến định lượng và biến định tính

Ví dụ 3 : Hãy lập mô hình mô tả quan hệ giữa thu nhập của giáo viên với thâm niên giảng dạy và vùng giảng dạy (thành phố, tỉnh đồng bằng, miền núi).

Gọi Y : thu nhập (triệu đồng/năm)

X : thâm niên giảng dạy (năm)

Z_1, Z_2 : biến giả.

$Z_{1i} =$ 1 : thành phố $Z_{2i} =$ 1 : tỉnh

0 : nơi khác 0 : nơi khác

Ta có mô hình :

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 Z_{1i} + \beta_4 Z_{2i} + U_i$$

Ý nghĩa của $\beta_2, \beta_3, \beta_4$: …

Ví dụ 4 : Hãy lập mô hình mô tả quan hệ giữa thu nhập của giáo viên với thâm niên giảng dạy, vùng giảng dạy (thành phố, tỉnh đồng bằng, miền núi) và giới tính của giáo viên.

Mô hình :

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 Z_{1i} + \beta_4 Z_{2i} + \beta_5 D_i + U_i$$

Trong đó : Y, X, Z_{1i}, Z_{2i} giống ví dụ 3.

D_i (biến giả) = 1 : nam giới

0 : nữ giới

Ý nghĩa của β_5 : …

Ví dụ 5 : Lập mô hình quan hệ giữa chi tiêu cá nhân với thu nhập và giới tính của cá nhân đó.

$$Y_i = \beta_1 + \beta X_i + \beta_3 Z_i + U_i \qquad (1)$$

Y – chi tiêu (triệu/tháng)

X – thu nhập (triệu/tháng)

Z_i = 1 : nam giới

0 : nữ giới.

* Mở rộng mô hình : Với mô hình trên, khi thu nhập cá nhân tăng 1 triệu đồng thì chi tiêu tăng β triệu đồng bất kể là nam hay nữ.

Nhưng với giả thiết cho rằng nếu thu nhập tăng 1 triệu đồng thì mức chi tiêu tăng thêm của nam và nữ khác nhau thì β phải là

$$\beta = \beta_2 + \beta_4 Z_i$$

Lúc này mô hình (1) được viết :

$$Y_i = \beta_1 + (\beta_2 + \beta_4 Z_i)X_i + \beta_3 Z_i + U_i$$

Hay :

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 Z_i + \beta_4 X_i Z_i + U_i \quad (2)$$

Trong đó : X_iZ_i được gọi là biến tương tác giữa X và Z.

- Khi $Z_i = 1$: $Y_i = (\beta_1 + \beta_3) + (\beta_2 + \beta_4)X_i + U_i$
  Đây là hồi qui chi tiêu-thu nhập của nam.
- Khi $Z_i = 0$: $Y_i = \beta_1 + \beta_2 X_i + U_i$
  Đây là hồi qui chi tiêu-thu nhập của nữ.

Ý nghĩa của các hệ số :

- β_1: Khi không có thu nhập thì chi tiêu trung bình của một người nữ là β_1 triệu.
- β_2: Khi thu nhập của một người nữ tăng 1 triệu đồng thì chi tiêu của họ tăng β_2 triệu đồng.

- β_3: Khi không có thu nhập thì chi tiêu trung bình của một người nam chênh lệch so với của một người nữ là β_3 triệu (hay chênh lệch về hệ số tung độ gốc giữa hàm hồi qui cho nam và hàm hồi qui cho nữ).
- β_4: Khi thu nhập của một người nam tăng 1 triệu đồng thì chi tiêu của họ tăng nhiều hơn của nữ β_4 triệu đồng (nếu $\beta_4 > 0$) hay tăng ít hơn của nữ β_4 triệu đồng (nếu $\beta_4 < 0$) (Hay chênh lệch về hệ số độ dốc giữa hàm hồi qui cho nam và hàm hồi qui cho nữ).

Do đó :

$H_0 : \beta_3 = 0 \Leftrightarrow$ hệ số tung độ gốc giữa hồi qui cho nam và cho nữ là giống nhau.

$H_0 : \beta_4 = 0 \Leftrightarrow$ hệ số độ dốc giữa hồi qui cho nam và cho nữ là giống nhau.

$H_0 : \beta_3 = \beta_4 = 0 \Leftrightarrow$ hồi qui cho nam và cho nữ là giống hệt nhau (chi tiêu của nam và của nữ là giống nhau)

III. Sử dụng biến giả trong phân tích mùa

Có nhiều phương pháp để loại nhân tố mùa khỏi chuỗi thời gian, một trong số đó là phương pháp biến giả.

<u>Ví dụ</u> : Giả sử cần nghiên cứu quan hệ giữa lợi nhuận và doanh thu ở một công ty, người ta thu nhập mẫu số liệu theo quý và cho rằng mỗi quí có thể biểu thị mẫu theo mùa. Mô hình đề nghị :

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 Z_{2i} + \beta_4 Z_{3i} + \beta_5 Z_{4i} + U_i$$

Y- lợi nhuận (triệu đồng/quý)

X- doanh thu (triệu đồng/quý)

Z_{2i} =1: qsát ở quý 2; Z_{2i}= 0 : qsát ở quý khác

Z_{3i} =1: qsát ở quý 3; Z_{3i}= 0 : qsát ở quý khác

Z_{4i} =1: qsát ở quý 4; Z_{4i}= 0 : qsát ở quý khác

H_0: $\beta_3 = 0$ (không có mùa vụ xảy ra ở quý 2)

H_0: $\beta_4 = 0$ (không có mùa vụ xảy ra ở quý 3)

H_0: $\beta_5 = 0$ (không có mùa vụ xảy ra ở quý 4)

- Loại bỏ yếu tố mùa : Giả sử sau khi ước lượng hàm hồi qui trên, ta có hệ số của Z_2 là 1322 và khác 0 có nghĩa. Lúc này, để loại bỏ yếu tố mùa ở quý 2, ta lấy các giá trị của lợi nhuận ở quý 2 trừ đi 1322.
- Giả sử sự tương tác giữa mùa và doanh thu có ảnh hưởng lên lợi nhuận thì mô hình sẽ là :

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 Z_{2i} + \beta_4 Z_{3i} + \beta_5 Z_{4i} +$$
$$+ \beta_6 (Z_{2i} X_i) + \beta_7 (Z_{3i} X_i) + \beta_8 (Z_{4i} X_i) + U_i$$

IV. So sánh hai hồi qui - phương pháp biến giả

Ví dụ : Số liệu về tiết kiệm (Y) và thu nhập cá nhân (X) ở Anh từ năm 1946 đến 1963 chia làm hai thời kỳ :

- Thời kỳ tái thiết (1946 - 1954) → $n_1=9$
- Thời kỳ hậu tái thiết (1955-1963) → $n_2=9$

Với thời kỳ tái thiết, hàm hồi qui :

$$Y_i = \alpha_1 + \alpha_2 X_i + U_i \quad (1)$$

Với số liệu → $\hat{Y}_i = -0.266 + 0.04705\ X_i$

Với thời kỳ hậu tái thiết, hàm hồi qui :

$$Y_i = \gamma_1 + \gamma_2 X_i + U_i \quad (2)$$

Với số liệu → $\hat{Y}_i = -1.75 + 0.15045 \; X_i$

Vấn đề : Hai hàm hồi qui ứng với hai thời kỳ trên có giống nhau không ? (hay là : mối quan hệ giữa tiết kiệm và thu nhập có giống nhau ở hai thời kỳ ?)

* Phương pháp :

- Gom 2 mẫu con thành một mẫu lớn có kích thước $n = n_1 + n_2$ và hồi qui mô hình :

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 Z_i + \beta_4 X_i Z_i + U_i \quad (*)$$

Với $Z_i = 1$: nếu là thời kỳ tái thiết,

0 : nếu là thời kỳ hậu tái thiết.

$\Rightarrow \beta_3$ là chênh lệch về hệ số tung độ gốc, β_4 là chênh lệch về hệ số độ dốc giữa hai hồi qui.

Vì :

+ Nếu $Z_i = 1$: (*) trở thành :

$Y_i = (\beta_1 + \beta_3) + (\beta_2 + \beta_4)X_i + U_i$:

hàm hồi qui cho thời kỳ tái thiết

+ Nếu $Z_i = 0$: (*) trở thành :

$Y_i = \beta_1 + \beta_2 X_i + U_i$:

hàm hồi qui cho thời kỳ hậu tái thiết

- Nên các kiểm định sau so sánh được 2 hqui:

$H_0 : \beta_3 = 0$ (hai hồi qui giống nhau ở tung độ gốc).

$H_0 : \beta_4 = 0$ (hai hồi qui giống nhau ở hsố góc)

$H_0 : \beta_3 = \beta_4 = 0$ (hai hồi qui giống hệt nhau)

Ví dụ : Sau khi gom số liệu cả hai thời kỳ và hồi qui mô hình (*), ta được :

$$\hat{Y}_i = -1.75 + 0.15045\ X_i + 1.484\ Z_i - 0.1034\ X_i Z_i$$

Se = (0.33) (0.470) (0.0163) (0.0333)

t = (-5.27) (3.155) (9.238) (-3.11)

p = (0.000) (0.007) (0.000) (0.008)

Kết quả trên cho thấy hai hồi qui cho hai thời kỳ hoàn toàn khác nhau vì : …