Biến phân loại

Bởi:

Phạm Trí Cao

Biến phân loại (Biến giả-Dummy variable)

Trong các mô hình hồi quy mà chúng ta đã khảo sát từ đầu chương 3 đến đây đều dựa trên biến độc lập và biến phụ thuộc đều là biến định lượng. Thực ra mô hình hồi quy cho phép sử dụng biến độc lập và cả biến phụ thuộc là biến định tính. Trong giới hạn chương trình chúng ta chỉ xét biến phụ thuộc là biến định lượng. Trong phần này chúng ta khảo sát mô hình hồi quy có biến định tính.

Đối với biến định tính chỉ có thể phân lớp, một quan sát chỉ có thể rơi vào một lớp. Một số biến định tính có hai lớp như:

Biến định tính	Lớp 1	Lớp 2
Giới tính	Nữ	Nam
Vùng	Thành thị	Nông thôn
Tôn giáo	Có	Không
Tốt nghiệp đại học	Đã	Chưa

Bảng 4.1. Biến nhị phân

Người ta thường gán giá trị 1 cho một lớp và giá trị 0 cho lớp còn lại. Ví dụ ta ký hiệu S là giới tính với S =1 nếu là nữ và S = 0 nếu là nam.

Các biến định tính được gán giá trị 0 và 1 như trên được gọi là biến giả(dummy variable), biến nhị phân, biến phân loại hay biến định tính.

Hồi quy với một biến định lượng và một biến phân loại

Ví dụ 4.1. Ở ví dụ này chúng ta hồi quy tiêu dùng cho gạo theo quy mô hộ có xem xét hộ đó ở thành thị hay nông thôn.

Mô hình kinh tế lượng như sau:

$Y_i = \beta_1 + \beta_2 X_i + \beta_3 D_i + \beta i$(4.19)Y: Chi tiêu cho gạo, ngàn đồng/năm

X : Quy mô hộ gia đình, người

D: Biến phân loại, D = 1 nếu hộ ở thành thị, bằng D = 0 nếu hộ ở nông thôn.

Chúng ta muốn xem xét xem có sự khác biệt trong tiêu dùng gạo giữa thành thị và nông thôn hay không ứng với một quy mô hộ gia đình Xi xác định.

Đối với hộ ở nông thôn

$$E[Y_i|X_i, D_i = 0] = \beta_1 + \beta_2 X_i$$

(4.20)

Đối với hộ ở thành thị

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

(4.21)

Vậy sự chênh lệch trong tiêu dùng gạo giữa thành thị và nông thôn như sau

$$E[Y_i|X_i, D_i = 1] - E[Y_i|X_i, D_i = 0] = \beta_3$$

(4.22)

Sự khác biệt trong tiêu dùng gạo giữa thành thị và nông thôn chỉ có ý nghĩa thống kê khi β_3 khác không có ý nghĩa thống kê.

Chúng ta đã có phương trình hồi quy như sau

Y = 187 + 508*X - 557*D (4.23)

t-stat [0,5] [6,4] [-2,2]

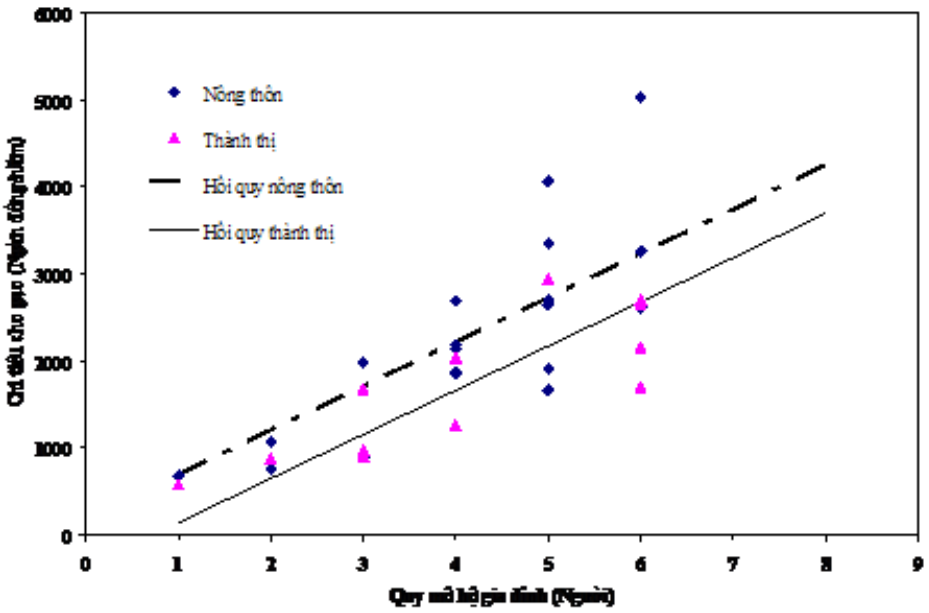
R^2 hiệu chỉnh = 0,61

Hệ số hồi quy

$$\hat{\beta}_3 = -557$$

khác không với độ tin cậy 95%. Vậy chúng ta không thể bác bỏ được sự khác biệt trong tiêu dùng gạo giữa thành thị và nông thôn.

Chúng ta sẽ thấy tác động của làm cho tung độ gốc của phương trình hồi quy của thành thị và nông thôn sai biệt nhau một khoảng β_3 = -557 ngàn đồng/năm. Cụ thể ứng với một quy mô hộ gia đình thì hộ ở thành thị tiêu dùng gạo ít hơn hộ ở nông thôn 557 ngàn đồng/năm.Chúng ta sẽ thấy điều này một cách trực quan qua đồ thị sau:



Hình 4.1. Hồi quy với một biến định lượng và một biến phân loại.

Hồi quy với một biến định lượng và một biến phân loại có nhiều hơn hai phân lớp

Ví dụ 4.2. Giả sử chúng ta muốn ước lượng tiền lương được quyết định bởi số năm kinh nghiệm công tác và trình độ học vấn như thế nào.

Gọi Y : Tiền lương

X : Số năm kinh nghiệm

D: Học vấn. Giả sử chúng ta phân loại học vấn như sau : chưa tốt nghiệp đại học, đại học và sau đại học.

Phương án 1:

Di = 0 nếu chưa tốt nghiệp đại học

Di = 1 nếu tốt nghiệp đại học

Di =2 nếu có trình độ sau đại học

Cách đặt biến này đưa ra giả định quá mạnh là phần đóng góp của học vấn vào tiền lương của người có trình độ sau đại học lớn gấp hai lần đóng góp của học vấn đối với người có trình độ đại học. Mục tiêu của chúng ta khi đưa ra biến D chỉ là phân loại nên ta không chọn phương án này.

Phương án 2: Đặt bộ biến giả

D_{1i} D_{2i} Học vấn

00 Chưa đại học

10 Đại học

01 Sau đại học

Mô hình hồi quy

$$Y_i = \beta_1 + \beta_2 X + \beta_3 D_{1i} + \beta_4 D_{2i} + \beta_i (4.24)$$

Khai triển của mô hình (4.24) như sau

Đối với người chưa tốt nghiệp đại học

$$E(Y_i) = \beta_1 + \beta_2 X \ (4.25)$$

Đối với người có trình độ đại học

$$E(Y_i) = (\beta_1 + \beta_3) + \beta_2 X_3 (4.26)$$

Đối với người có trình độ sau đại học

$$E(Y_i) = (\beta_1 + \beta_3 + \beta_4) + \beta_2 X \ (4.27)$$

Cái bẩy của biến giả

Số lớp của biến phân loại Số biến giả

Trong ví dụ 4.1. 21

Trong ví dụ 4.232

Điều gì xảy ra nếu chúng ta xây dựng số biến giả đúng bằng số phân lớp?

Ví dụ 4.3. Xét lại ví dụ 4.1.

Giả sử chúng ta đặt biến giả như sau

D_{1i} D_{2i} Vùng

10 Thành thị

01 Nông thôn

Mô hình hồi quy là

$Y_i = \beta_1 + \beta_2 X_i + \beta_3 D_{1i} + \beta_4 D_{2i} + \beta_i$ (4.28)

Chúng ta hãy xem kết quả hồi quy bằng Excel

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	2235,533	0	65535	#NUM!
X	508,1297	80,36980143	6,322396	1,08E-06
D1	-2605,52	0	65535	#NUM!
D2	-2048	0	65535	#NUM!

Kết quả hồi quy rất bất thường và hoàn toàn không có ý nghĩa kinh tế.

Lý do là có sự đa cộng tuyến hoàn hảo giữa D1, D2 và một biến hằng X2 =-1.

$D_{1i} + D_{2i} + X_2 = 0 \ \forall \ i.$

Hiện tượng đa cộng tuyến hoàn hảo này làm cho hệ phương trình chuẩn không có lời giải. Thực tế sai số chuẩn tiến đến vô cùng chứ không phải tiến đến 0 như kết quả tính toán của Excel. Hiện tượng này được gọi là cái bẫy của biến giả.

Quy tắc: Nếu một biến phân loại có k lớp thì chỉ sử dụng (k-1) biến giả.

Hồi quy với nhiều biến phân loại

Ví dụ 4.4. Tiếp tục ví dụ 4.2. Chúng ta muốn khảo sát thêm có sự phân biệt đối xử trong mức lương giữa nam và nữ hay không.

Đặt thêm biến và đặt lại tên biến

GT_i: Giới tính, 0 cho nữ và 1 cho nam.

TL : Tiền lương

KN: Số năm kinh nghiệm làm việc

ĐH: Bằng 1 nếu tốt nghiệp đại học và 0 cho chưa tốt nghiệp đại học

SĐH: Bằng 1 nếu có trình độ sau đại học và 0 cho chưa.

Mô hình hồi quy $TL_i = \beta_1 + \beta_2 KN_i + \beta_3 ĐH_i + \beta_4 SĐH_i + \beta_5 GT_i + \beta_i$(4.29)

Chúng ta xét tiền lương của nữ có trình độ sau đại học

$E(TL_i / SĐH=1 \cap GT=0) = (\beta_1 + \beta_4) + v_2 KN_i$

Biến tương tác

Xét lại ví dụ 4.1. Xét quan hệ giữa tiêu dùng gạo và quy mô hộ gia đình.Để cho đơn giản trong trình bày chúng ta sử dụng hàm toán như sau.

Nông thôn: $Y = \beta_1 + \beta_1 X$

Thành thị: $Y = \beta_2 + \beta_2 X$

D : Biến phân loại, bằng 1 nếu hộ ở thành thị và bằng 0 nếu hộ ở nông thôn.

Có bốn trường hợp có thể xảy ra như sau

$\beta_1 = \beta_2$ và $\beta_1 = \beta_2$, hay không có sự khác biệt trong tiêu dùng gạo giữa thành thị và nông thôn.

Mô hình : $Y = a + b X$

Trong đó $\beta_1 = \beta_2 = a$ và $\beta_1 = \beta_2 = b$.

$\beta_1 \neq \beta_2$ và $\beta_1 = \beta_2$, hay có sự khác biệt về tung độ gốc

Mô hình: $Y = a + bX + cD$

Trong đó $\beta_1 = a$, $\beta_2 = a + c$ và $\beta_1 = \beta_2 = b$.

$\beta_1 = \beta_2$ và $\beta_1 \neq \beta_2$, hay có sự khác biệt về độ dốc

Mô hình: $Y = a + bX + c(DX)$

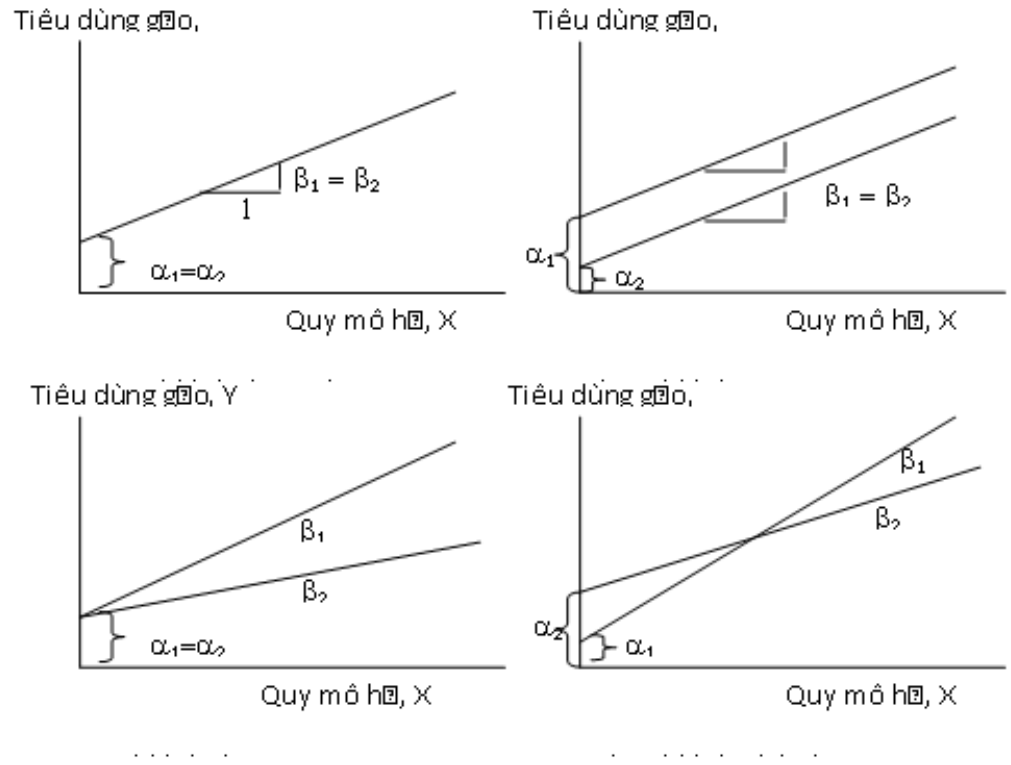
Trong đó $DX = X$ nếu nếu $D = 1$ và $DX = 0$ nếu $D = 0$

$\beta_1 = \beta_2 = a$, $\beta_1 = b$ và $\beta_2 = b + c$.

$\beta_1 \neq \beta_2$ và $\beta_1 \neq \beta_2$, hay có sự khác biệt hoàn toàn về cả tung độ gốc và độ dốc.

Mô hình: $Y = a + bX + cD + d(DX)$

$\beta_1 = a$, $\beta_2 = a + c$, $\beta_1 = b$ và $\beta_2 = b + d$.



Hình 4.2. Các mô hình hồi quy

Biến DX được xây dựng như trên được gọi là biến tương tác. Tổng quát nếu X_p là một biến định lượng và D_q là một biến giả thì X_pD_q là một biến tương tác. Một mô hình hồi quy tuyến tổng quát có thể có nhiều biến định lượng, nhiều biến định tính và một số biến tương tác.