PHƯƠNG PHÁP SỬ DỤNG BIẾN PHẠM TRÙ TRONG PHÂN TÍCH HỒI QUI

Lê Đỗ Mạch(*)

Hồi quy là một phương pháp phổ dụng trong thống kê. Trước đây phương pháp này chủ yếu áp dụng để phân tích mối quan hệ giữa các biến định lượng. Bài viết giới thiệu một số phương pháp hồi quy cho các biến phạm trù.

1. Hồi qui với các biến phạm trù có tương tác

1.1. Sử dụng lệnh xi

Cũng thực hiện phân tích như trên đối với hai biến hhsizecate và urban98 nhưng đưa thêm vào sự tương tác giữa chúng. Chúng ta dùng lệnh xi

```
. xi: regress lricexpd i.hhsizecate*urban98
```

```
    lricexpd |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
_Ihhsizeca~1 |   43.69716     7.3998     5.91   0.000     29.19089    58.20343
_Ihhsizeca~2 |   24.37061   5.023043     4.85   0.000     14.52363    34.21758
     urban98 |  -96.34056   6.253713   -15.41   0.000    -108.6001   -84.08103
_IhhsXurba~1 |   49.12839   13.46977     3.65   0.000      22.7228    75.53398
_IhhsXurba~2 |   8.610493    9.16319     0.94   0.347    -9.352657    26.57364
       _cons |   495.2452   3.019393   164.02   0.000     489.3261    501.1643
------------------------------------------------------------------------------
```

Chúng ta có thể kiểm tra toàn bộ ảnh hưởng bằng lệnh test, ảnh hưởng tương tác này là có ý nghĩa.

```
. test _IhhsXurban_1 _IhhsXurban_2

( 1)  _IhhsXurban_1 = 0
( 2)  _IhhsXurban_2 = 0

       F(  2,  5993) =    6.68
            Prob > F =    0.0013
```

Điều quan trọng là phải chú ý đến ý nghĩa của sự thay đổi hệ số như thế nào với việc có mặt của sự tương tác. Với mô hình trên, sự khác nhau đó được thể hiện như sau.

```
                hhsizecate =1         hhsizecate =2         hhsizecate =3
            ------------------------------------------------
 urban98=0   _cons                  _cons                 _cons
            +BIhhsizecat1          +BIhhsizecat2
            ------------------------------------------------
 urban98=1   _cons                  _cons                 _cons
            +Burban98              +Burban98             +Burban98
            +BIhhsizecat1          +BIhhsizecat2
            +B_IhhsXurban_1        +B_IhhsXurban_2
```

(*) Viện Khoa học Thống kê

Kiểm định urban98 khi hhsizecate =1

```
. test _b[urban98]+_b[_IhhsXurban_1]==0

( 1)  urban98 + _IhhsXurban_1 = 0

       F(  1,  5993) =   15.66
            Prob > F =    0.0001
```

Kiểm định này là có ý nghĩa, cho biết ảnh hưởng của khu vực nông thôn thành thị (urban98) là có ý nghĩa với nhóm cỡ hộ hhsizecate = 1. Kiểm định cũng có thể thực hiện bằng lệnh anova.

1.2. Sử dụng lệnh anova

Việc xây dựng tương tác có thể dễ hơn khi sử dụng lệnh anova. Như chúng ta đã biết, lệnh anova cho một kiểm định toàn bộ tác động chính và các tương tác mà không cần thực hiện các lệnh test tiếp sau.

```
. anova lricexpd urban98 hhsizecate urban98*hhsizecate
                  Source |  Partial SS    df       MS           F     Prob > F
      -------------------+----------------------------------------------------
                   Model |  10917603.4     5  2183520.69      99.05     0.0000
                         |
                urban98 |   5210956.4     1   5210956.4     236.39     0.0000
              hhsizecate |  2492129.36     2  1246064.68      56.53     0.0000
      urban98*hhsizecate |  294395.104     2  147197.552       6.68     0.0013
                         |
                Residual |   132111278  5993  22044.2647
      -------------------+----------------------------------------------------

                   Total |   143028882  5998  23846.0957
```

2. Hồi qui với các biến liên tục và phạm trù

2.1 Sử dụng hồi quy

Chúng ta tiến hành phân tích đối với cả hai biến phạm trù và liên tục, đó là các biến urban98 và lhhexp.

```
. regress lricexpd urban98 lhhexp
------------------------------------------------------------------------------
    lricexpd |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     urban98 |  -120.6797   5.060426   -23.85   0.000    -130.5999   -110.7594
      lhhexp |   50.30079   3.680352    13.67   0.000     43.08598    57.5156
       _cons |   120.2196   28.46238     4.22   0.000     64.42306    176.0161
------------------------------------------------------------------------------
```

Chúng ta có thể tạo ra giá trị dự đoán sử dụng lệnh predict.

```
. predict yhat
(option xb assumed; fitted values)
```

Có thể vẽ đồ thị giá trị dự đoán đối với biến lhhexp bằng lệnh:

```
. scatter yhat lhhexp
```

Hệ số của **lhhexp** cho biết với mỗi đơn vị tăng lên của **lhhexp**, **lricexpd** được dự đoán tăng lên 50.3 đơn vị. Đây là độ nghiêng của các đường thẳng trong đồ thị. Đồ thị có hai đường, đường phía trên cho khu vực nông thôn và một đường phía dưới cho khu vực thành thị. Hệ số của **urban98** là -120.6797, cho biết nếu là khu vực thành thị (**urban98=1)**, thì **lricexpd** giảm đi 121 đơn vị.

2.2. Sử dụng lệnh anova

Lệnh anova giả thiết rằng các biến là phạm trù, vì vậy trong trường hợp này cần phải sử dụng chọn lựa continuous (),viết tắt là contv ())để khai báo là lhhexp là một biến liên tụcđ

```
. anova lricexpd urban98 lhhexp, cont(lhhexp)
```

```
      Source |  Partial SS    df       MS           F     Prob > F
-----------+----------------------------------------------------
       Model |  12395772.9     2  6197886.46     284.48     0.0000
             |
    urban98 |  12390407.6     1  12390407.6     568.71     0.0000
      lhhexp |  4069702.19     1  4069702.19     186.80     0.0000
             |
    Residual |   130633109  5996  21786.7093
-----------+----------------------------------------------------
       Total |   143028882  5998  23846.0957
```

Nếu bình phương các giá trị t từ lệnh regress ở trên sẽ thấy chúng bằng các giá trị F của lệnh anova. Tác động của các biến urban98 và lhhexp là có ý nghĩa.

2.3. Sự tương tác của các biến liên tục với các biến 0/1

Ở trên đã phân tích mối quan hệ giữa lricexpd và lhhexp với sự có mặt của biến urban98 trong mô hình. Chúng ta cũng đã vẽ đồ thị về mối quan hệ giữa lricexpd và lhhexp, chú ý là đồ thị có hai đường hồi quy, một đường cao hơn và một đường thấp hơn nhưng cùng độ dốc. Một mô hình như vậy giả thiết cùng độ dốc cho hai nhóm. Nhưng có thể độ dốc là khác nhau cho các nhóm. Chúng ta sẽ chạy hồi quy riêng rẽ cho hai nhóm này bắt đầu với tiêu dùng gạo ở nông thôn.

```
. regress lricexpd lhhexp if urban98==0
```

```
------------------------------------------------------------------------------
    lricexpd |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lhhexp |   79.37483   4.582882    17.32   0.000        70.39    88.35966
       _cons |  -103.9182   35.39845    -2.94   0.003    -173.3176   -34.51883
------------------------------------------------------------------------------
```

Dự đoán giá trị hồi qui

```
. predict yhat0 if urban98==0
```

```
(option xb assumed; fitted values)
(1730 missing values generated)
```

Cũng như vậy chúng ta sẽ quan sát tiêu dùng gạo cho các hộ ở thành thị.

```
. regress lricexpd lhhexp if urban98==1
------------------------------------------------------------------------------
    lricexpd |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lhhexp |   5.123454   6.123901     0.84   0.403    -6.887584    17.13449
       _cons |   382.3473   52.02188     7.35   0.000     280.3148    484.3798
------------------------------------------------------------------------------
```

Dự đoán giá trị hồi qui

```
. predict yhat1 if urban98==1
(option xb assumed; fitted values)
(4269 missing values generated)
```

2.3.1. Tạo tương tác thủ công

Chúng ta sẽ bắt đầu tính toán tương tác giữa lhhexp và urban98 .

```
. gen  urbXlhhe=urban98*lhhexp
```

Hãy tiến hành hồi quy để kiểm tra xem về mặt ý nghĩa có sự khác nhau giữa các hộ ở nông thôn và các hộ ở thành thị đối với hệ số của lhhexp. Dĩ nhiên tương tác urbXlhhe là có ý nghĩa.

```
. regress lricexpd lhhexp urban98 urbXlhhe
------------------------------------------------------------------------------
    lricexpd |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lhhexp |   79.37483   4.680334    16.96   0.000     70.19969    88.54997
     urban98 |   486.2655   61.34515     7.93   0.000     366.0069     606.524
    urbXlhhe |  -74.25138   7.479567    -9.93   0.000    -88.91402   -59.58874
       _cons |  -103.9182   36.15118    -2.87   0.004    -174.7875   -33.04888
------------------------------------------------------------------------------
```

Tiến hành tạo ra hai đường hồi quy để mô tả sự khác nhau giữa chúng, đầu tiên dự đoán giá trị hồi quy.

```
. predict yhata
(option xb assumed; fitted values)
```

Tiếp theo tạo ra hai biến riêng cho hai khu vực, yhata0 cho các hộ ở nông thôn và yhata1 cho các hộ ở thành thị.

```
. separate yhata, by(urban98)
              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
yhata0          float  %9.0g                  yhata, urban98 == Rural
yhata1          float  %9.0g                  yhata, urban98 == Urban
```

Chú ý rằng hệ số của lhhexp trong phân tích kết hợp cũng giống như hệ số của lhhexp trong phân tích riêng với các hộ ở nông thôn. Đó là do các hộ ở nông thôn là nhóm tham khảo. Tiếp theo hệ số của biến tương tác urbXlhhe trong phân tích kết hợp là -74.25 chính là

hiệu số của phép trừ hệ số của biến lhhexp đối với các hộ ở thành thị trừ đi hệ số của biến này đối với các hộ ở nông thôn. Tương tác này chính là sự khác nhau về độ dốc của hai loại đường hồi qui, và vì thế điều này là có ích để kiểm định hai đường hồi qui của hai loại hộ có độ dốc bằng nhau không. Nếu cả hai loại hộ có cùng hệ số hồi qui cho lhhexp, thì hệ số cho tương tác urbXlhhe sẽ bằng 0.

Vì vậy nếu nhìn đồ thị của hai đường hồi qui, chúng ta sẽ thấy sự khác nhau về độ dốc của chúng. Sự khác nhau giữa hai độ dốc này là 74.25, đó là hệ số của urbXlhhe.

2.3.2. Tạo tương tác bằng lệnh xi

Sử dụng lệnh xi tương tự như trước kia.

```
. xi: regress lricexpd i.urban98*lhhexp
```

```
------------------------------------------------------------------------------
    lricexpd |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 _Iurban98_1 |   486.2655   61.34515     7.93   0.000     366.0069     606.524
      lhhexp |   79.37483   4.680334    16.96   0.000     70.19969    88.54997
_IurbXlhhe~1 |  -74.25138   7.479567    -9.93   0.000    -88.91402   -59.58874
       _cons |  -103.9182   36.15118    -2.87   0.004    -174.7875   -33.04888
------------------------------------------------------------------------------
```

Số hạng i. urban98 *lhhexp tạo ra 3 biến: lhhexp đã có trong số liệu, _Iurban98_1 mô tả khu vực nông thôn thành thị và _IurbXlhhe_1 biểu diễn sự tương tác của urban98 với lhhexp. Cũng như trên có thể tạo ra các giá trị dự đoán và các đồ thị mô tả hai đường hồi qui của hai nhóm hộ.

2.3.3. Tạo tương tác bằng lệnh anova

```
. anova lricexpd urban98 lhhexp urban98*lhhexp, contin(lhhexp)
                  Source |  Partial SS    df       MS           F     Prob > F
              -----------+----------------------------------------------------
                   Model |  14508478.8     3   4836159.6     225.59     0.0000
                         |
                 urban98 |   1347005.3     1   1347005.3      62.83     0.0000
                  lhhexp |   2736061.2     1   2736061.2     127.63     0.0000
          urban98*lhhexp |  2112705.87     1  2112705.87      98.55     0.0000
                         |
                Residual |   128520403  5995  21437.9321
              -----------+----------------------------------------------------
                   Total |   143028882  5998  23846.0957
```

Như minh hoạ ở trên chúng ta cũng có thể tính toán các giá trị dự đoán và vẽ đồ thị các đường hồi qui riêng biệt.

Tài liệu tham khảo

Lê Đỗ Mạch - Báo cáo kết quả nghiên cứu đề tài xây dựng quy trình và phương pháp thực hành phân tích hồi quy dựa trên phần mềm stata, Hà Nội 2005