Chương 5

ĐA CỘNG TUYẾN

1. Đa cộng tuyến là gì ?

Theo giả thiết của phương pháp OLS thì các biến độc lập không có mối quan hệ tuyến tính.

Nếu quy tắc này bị vi phạm thì sẽ có hiện tượng đa cộng tuyến,

Như vậy , "đa cộng tuyến "là hiện tượng các biến độc lập trong mô hình phụ thuộc tuyến tính lẫn nhau và thể hiện được dưới dạng hàm số

1. Đa cộng tuyến là gì ?

Xét mô hình hồi quy tuyến tính k biến với hàm PRF :

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + U_i$$

Đa cộng tuyến hoàn hảo xảy ra khi giữa các biến độc lập có mối quan hệ chính xác theo dạng

$$a_2 X_2 + a_3 X_3 + ... + a_k X_k = 0$$

Đa cộng tuyến không hoàn hảo xảy ra khi giữa các biến độc lập có mối quan hệ theo dạng

$$a_2 X_2 + a_3 X_3 + ... + a_k X_k + V = 0$$

1. Đa cộng tuyến là gì ?

Ví dụ • Đa cộng tuyến hoàn hảo:

X_2	X_3	X_4
10	50	52
15	75	78
18	90	97
24	120	129
11	55	63

X_2 và X_3 có mối quan hệ tuyến tính chính xác:

$X_3 = 5X_2$ => Trường hợp này có đa cộng tuyến hoàn hảo

1. Đa cộng tuyến là gì ?

- ***Điều gì xảy ra khi có đa cộng tuyến hoàn hảo ?***

Xét ví dụ hàm hồi quy tuyến tính 3 biến

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + U_i$$

Và giả sử có đa cộng tuyến hoàn hảo : $X_{3i}=aX_{2i}$

Ta có :

$$\hat{\beta}_2 = \frac{\left(\sum y_i x_{2i}\right)\left(\sum x_{3i}^2\right) - \left(\sum x_{2i} x_{3i}\right)\left(\sum y_i x_{3i}\right)}{\left(\sum x_{2i}^2\right)\left(\sum x_{3i}^2\right) - \left(\sum x_{2i} x_{3i}\right)^2}$$

Vì : $X_{3i}=aX_{2i}$

$$\hat{\beta}_2 = \frac{\left(\sum y_i x_{2i}\right)\left(a^2 \sum x_{2i}^2\right) - \left(a \sum x_{2i} x_{2i}\right)\left(a \sum y_i x_{2i}\right)}{\left(\sum x_{2i}^2\right)\left(a^2 \sum x_{2i}^2\right) - \left(a \sum x_{2i} x_{2i}\right)^2} = \frac{0}{0}$$

1. Đa cộng tuyến là gì ?

Đây là dạng vô định => Vậy không xác định được $\hat{\beta}_2$

Tương tự => Vậy không xác định được $\hat{\beta}_3$

Tổng quát : ma trận (X^TX) suy biến, không có ma trận nghịch đảo

Như vậy trong trường hợp đa cộng tuyến hoàn hảo thì sẽ không xây dựng được mô hình hồi quy

1. Đa cộng tuyến là gì ?

- ***Điều gì xảy ra khi có đa cộng tuyến không hoàn hảo ?***

Chúng ta vẫn ước lượng được các tham số và xây dựng được mô hình hồi quy nhưng hãy xét đến hậu quả của đa cộng tuyến không hoàn hảo trong các phần tiếp theo

2. Hệ quả của đa cộng tuyến

Khi gặp đa cộng tuyến hoàn hảo, *chúng ta không thể ước lượng được mô hình*

Hệ quả khi có đa cộng tuyến không hoàn hảo

1. Khi dùng phương pháp ước lượng OLS, phương sai vẫn là nhỏ nhất nhưng giá trị lại khá lớn so với giá trị ước lượng
2. Sai số chuẩn của các hệ số hồi qui sẽ lớn

Do đó:
- *Khoảng tin cậy lớn và việc kiểm định ít có ý nghĩa.*
- *Giả thiết H_0 dễ dàng được chấp nhận*

2. Hậu quả của đa cộng tuyến

3. R^2 cao nhưng tỷ số t ít có ý nghĩa

   Dễ dàng bác bỏ giả thuyết "không" của thống kê F và cho rằng mô hình ước lượng có gía trị.

2. Hậu quả của đa cộng tuyến

4. Các ước lượng và sai số chuẩn của ước lượng rất nhạy cảm với sự thay đổi của dữ liệu

   Chỉ cần một sự thay đổi nhỏ trong mẫu dữ liệu sẽ kéo theo sự thay đổi lớn các hệ số ước lượng.

2. Hậu quả của đa cộng tuyến

Ví dụ
- Xem kết quả ước lượng hàm tiêu dùng:
  - $Y = 24.77 + 0.94X_2 - 0.04X_3$
  - $R^2=0.96$, $F = 92.40$
  - X_2 : thu nhập
  - X_3 : của cải
  - R^2 rất cao giải thích 96% biến đổi của hàm tiêu dùng.

Sai sót :
✓Có một biến sai dấu.
✓Biến thu nhập và của cải tương quan rất mạnh với nhau do đó không thể nào ước lượng được tác động biên chính xác cho thu nhập hoặc của cải lên tiêu dùng

3. Nguồn gốc của đa cộng tuyến

➢Do phương pháp thu thập dữ liệu

Các giá trị của các biến độc lập phụ thuộc lẫn nhau trong mẫu, nhưng không phụ thuộc lẫn nhau trong tổng thể

Ví dụ: *người có thu nhập cao hơn khuynh hướng sẽ có nhiều của cải hơn. Điều này có thể đúng với mẫu mà không đúng với tổng thể . Cụ thể , trong tổng thể sẽ có các quan sát về các cá nhân có thu nhập cao nhưng không có nhiều của cải và ngược lại.*

3. Nguồn gốc của đa cộng tuyến

➢Dạng hàm mô hình:

Ví dụ: - hồi qui dạng hàm đa thức
- hồi quy mà số biến độc lập nhiều hơn số quan sát

➢Các biến độc lập được quan sát theo chuỗi thời gian có cùng chiều hướng biến động

Ví dụ: giá cả các mặt hàng tăng theo thời gian

4. Nhận biết đa cộng tuyến

➢ R^2 cao và thống kê t thấp.

Dấu hiệu này thể hiện nghịch lý gì ?

Nhược điểm : chỉ thể hiện rõ khi có đa cộng tuyến ở mức cao

4. Nhận biết đa cộng tuyến

➢Hệ số tương quan giữa các biến độc lập cao.

Công thức tính hệ số tương quan giữa X_2 và X_3

$$r_{23} = \frac{\sum (X_{2i} - \bar{X}_2)(X_{3i} - \bar{X}_3)}{\sqrt{\sum (X_{2i} - \bar{X}_2)^2 \sum (X_{3i} - \bar{X}_3)^2}}$$

Hệ số tương quan có ý nghĩa như thế nào ?

Nhược điểm của phương pháp này là gì ?

4. Nhận biết đa cộng tuyến

➢Thực hiện hồi qui phụ

Hồi qui giữa một biến độc lập nào đó theo các biến độc lập còn lại với nhau và quan sát hệ số R^2 của các hồi qui phụ

Hồi quy chính : $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + U_i$

Hồi quy phụ : $X_{4i} = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + V_i$

- Nhược điểm của việc hồi quy phụ là gì ?

4. Nhận biết đa cộng tuyến

➢Dùng nhân tử phóng đại phương sai

$$VIF_j = \frac{1}{1 - R_j^2}$$

R_j^2 là hệ số xác định của mô hình hồi qui phụ X_j theo các biến độc lập khác.

Nếu có đa cộng tuyến thì VIF lớn.

***$VIF_j > 10$** thì X_j có đa cộng tuyến cao với các biến khác.*

5. Khắc phục đa cộng tuyến

a) Bỏ qua đa cộng tuyến nếu $|t| > 2$

b) Bỏ qua đa cộng tuyến nếu R^2 của mô hình cao hơn R^2 của mô hình hồi qui phụ.

c) Bỏ qua đa cộng tuyến nếu mục tiêu xây dựng mô hình sử dụng để dự báo chứ không phải kiểm định.

5. Khắc phục đa cộng tuyến

d) Bỏ bớt biến độc lập.

Ví dụ: bỏ biến của cải ra khỏi mô hình hàm tiêu dùng.

e) Bổ sung dữ liệu hoặc tìm dữ liệu mới

f) Thay đổi dạng mô hình:

Ví dụ minh hoạ

Khảo sát chi tiêu cho tiêu dùng (Y), thu nhập (X_2) và quy mô tài sản (X_3) ta có số liệu sau :

Y	**70**	**65**	**90**	**95**	**110**	**115**	**120**	**140**	**155**	**150**
X_2	**80**	**100**	**120**	**140**	**160**	**180**	**200**	**220**	**240**	**260**
X_3	**810**	**1009**	**1273**	**1425**	**1633**	**1876**	**2052**	**2201**	**2435**	**2686**

Equation: UNTITLED Workfile: VIDUDACONGTUYEN::Untitl...

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: Y
Method: Least Squares
Date: 11/16/08 Time: 22:35
Sample: 1 10
Included observations: 10

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	24.77473	6.752500	3.668972	0.0080
X2	0.941537	0.822898	1.144172	0.2902
X3	-0.042435	0.080664	-0.526062	0.6151

R-squared	0.963504	Mean dependent var	111.0000
Adjusted R-squared	0.953077	S.D. dependent var	31.42893
S.E. of regression	6.808041	Akaike info criterion	6.917411
Sum squared resid	324.4459	Schwarz criterion	7.008186
Log likelihood	-31.58705	F-statistic	92.40196
Durbin-Watson stat	2.890614	Prob(F-statistic)	0.000009

Equation: UNTITLED Workfile: VIDUDACONGTUYEN::Untitl...

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: X2
Method: Least Squares
Date: 11/16/08 Time: 22:37
Sample: 1 10
Included observations: 10

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.386271	2.897956	-0.133291	0.8973
X3	0.097923	0.001578	62.04047	0.0000

R-squared	0.997926	Mean dependent var	170.0000
Adjusted R-squared	0.997667	S.D. dependent var	60.55301
S.E. of regression	2.925035	Akaike info criterion	5.161346
Sum squared resid	68.44662	Schwarz criterion	5.221863
Log likelihood	-23.80673	F-statistic	3849.020
Durbin-Watson stat	2.068509	Prob(F-statistic)	0.000000

Equation: UNTITLED Workfile: VIDUDACONGTUYEN::Untitl...

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: Y
Method: Least Squares
Date: 11/16/08 Time: 22:39
Sample: 1 10
Included observations: 10

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	24.45455	6.413817	3.812791	0.0051
X2	0.509091	0.035743	14.24317	0.0000

R-squared	0.962062	Mean dependent var	111.0000
Adjusted R-squared	0.957319	S.D. dependent var	31.42893
S.E. of regression	6.493003	Akaike info criterion	6.756184
Sum squared resid	337.2727	Schwarz criterion	6.816701
Log likelihood	-31.78092	F-statistic	202.8679
Durbin-Watson stat	2.680127	Prob(F-statistic)	0.000001

Equation: UNTITLED Workfile: VIDUDACONGTUYEN::Untitl...

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: Y
Method: Least Squares
Date: 11/16/08 Time: 22:40
Sample: 1 10
Included observations: 10

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	24.41104	6.874097	3.551164	0.0075
X3	0.049764	0.003744	13.29166	0.0000

R-squared	0.956679	Mean dependent var	111.0000
Adjusted R-squared	0.951264	S.D. dependent var	31.42893
S.E. of regression	6.938330	Akaike info criterion	6.888856
Sum squared resid	385.1233	Schwarz criterion	6.949373
Log likelihood	-32.44428	F-statistic	176.6681
Durbin-Watson stat	2.417419	Prob(F-statistic)	0.000001