

Chương 5

Đa cộng tuyến

Multicollinearity

Các giả thiết của mô hình CLRM (nhắc lại)

1. Mô hình là tuyến tính $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$
2. Kỳ vọng U_i bằng 0: $E(u_i | X_{2i}, X_{3i}) = 0$
3. Các U_i thuần nhất: $\text{var}(u_i) = \sigma^2$
4. Không có sự tương quan giữa các U_i : $\text{cov}(u_i, u_j) = 0, i \neq j$
5. Không có quan hệ tuyến tính giữa các biến giải thích.
 $\lambda_1 \mathbf{1} + \lambda_2 X_{2i} + \lambda_3 X_{3i} \neq \mathbf{0},$
 $\forall \lambda_1, \lambda_2, \lambda_3 \neq (0, 0, 0)$

Xét 3 giả thiết

- Chúng ta sẽ xét các vấn đề sau:
 - Đa cộng tuyến
 - Phương sai sai số thay đổi
 - Tự tương quan (tương quan chuỗi)
- Các chươngng có cùng cấu trúc
 1. Xác định bản chất của vấn đề
 2. Hậu quả của nó
 3. Nêu cách phát hiện
 4. Các phương pháp khắc phục

5.1. Bản chất của đa cộng tuyến

- Đa cộng tuyến hoàn hảo

$$\lambda_1 \cdot 1 + \lambda_2 X_2 + \lambda_3 X_3 = 0 \text{ với } (\lambda_1, \lambda_2, \lambda_3) \neq (0, 0, 0)$$

- Nghĩa rộng hơn (không hoàn hảo)

$$\lambda_1 \cdot 1 + \lambda_2 X_2 + \lambda_3 X_3 + v_i = 0 \text{ với } (\lambda_1, \lambda_2, \lambda_3) \neq (0, 0, 0)$$

5.2. Ước lượng khi có đa cộng tuyến hoàn hảo

- Mô hình hồi quy 3 biến có thể viết lại sau:

$$y_i = \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \hat{u}_i$$

- Tính toán trong chương 3, ta có:

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

$$\hat{\beta}_3 = \frac{(\sum y_i x_{3i})(\sum x_{2i}^2) - (\sum y_i x_{2i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2}$$

Assume that $X_{3i} = \lambda X_{2i}$, where λ is a nonzero constant

- Từ đó suy ra

$$\begin{aligned}\hat{\beta}_2 &= \frac{(\sum y_i x_{2i})(\lambda^2 \sum x_{2i}^2) - (\lambda \sum y_i x_{2i})(\lambda \sum x_{2i}^2)}{(\sum x_{2i}^2)(\lambda^2 \sum x_{2i}^2) - \lambda^2 (\sum x_{2i}^2)^2} \\ &= \frac{0}{0}\end{aligned}$$

- Tương tự, ta chỉ ra $\hat{\beta}_3$ không xác định.
- Từ chương 3, dễ thấy trong trường hợp đa cộng tuyến hoàn hảo, phương sai và sai số tiêu chuẩn của các ước lượng $\hat{\beta}_2$, $\hat{\beta}_3$ là vô hạn.

5.2. Ước lượng khi có đa cộng tuyến không hoàn hảo

- Giả thiết X_1, X_2 cộng tuyến không hoàn hảo

$$x_{3i} = \lambda x_{2i} + v_i$$

Với $\lambda \neq 0$, v_i là nhiễu ngẫu nhiên t/m $\sum x_{2i} v_i = 0$.??

Từ đó tính được:

$$\hat{\beta}_2 = \frac{\sum (y_i x_{2i}) (\lambda^2 \sum x_{2i}^2 + \sum v_i^2) - (\lambda \sum y_i x_{2i} + \sum y_i v_i) (\lambda \sum x_{2i}^2)}{\sum x_{2i}^2 (\lambda^2 \sum x_{2i}^2 + \sum v_i^2) - (\lambda \sum x_{2i}^2)^2}$$

- Tương tự tính được $\hat{\beta}_3$. Như vậy với v_i đủ nhỏ, không có lý gì để nói TH này ko ước lượng được.

5.4. Hậu quả của đa cộng tuyến

1. Phương sai và hiệp phương sai của các ƯL OLS

■ Mô hình $y_i = \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \hat{u}_i$

■ Ta có:

$$\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2 (1 - r_{23}^2)}$$

$$\text{var}(\hat{\beta}_3) = \frac{\sigma^2}{\sum x_{3i}^2 (1 - r_{23}^2)}$$

$$\text{cov}(\hat{\beta}_2, \hat{\beta}_3) = \frac{-r_{23}\sigma^2}{(1 - r_{23}^2)\sqrt{\sum x_{2i}^2 \sum x_{3i}^2}}$$

R_{23} là hệ số tương quan giữa X_2 và X_3 .

THE EFFECT OF INCREASING r_{23} ON $\text{VAR}(\hat{\beta}_2)$ AND $\text{COV}(\hat{\beta}_2, \hat{\beta}_3)$

Value of r_{23} (1)	VIF (2)	$\text{var}(\hat{\beta}_2)$ (3)*	$\frac{\text{var}(\hat{\beta}_2)(r_{23} \neq 0)}{\text{var}(\hat{\beta}_2)(r_{23} = 0)}$ (4)	$\text{cov}(\hat{\beta}_2, \hat{\beta}_3)$ (5)
0.00	1.00	$\frac{\sigma^2}{\sum x_{2i}^2} = A$	—	0
0.50	1.33	$1.33 \times A$	1.33	$0.67 \times B$
0.70	1.96	$1.96 \times A$	1.96	$1.37 \times B$
0.80	2.78	$2.78 \times A$	2.78	$2.22 \times B$
0.90	5.76	$5.26 \times A$	5.26	$4.73 \times B$
0.95	10.26	$10.26 \times A$	10.26	$9.74 \times B$
0.97	16.92	$16.92 \times A$	16.92	$16.41 \times B$
0.99	50.25	$50.25 \times A$	50.25	$49.75 \times B$
0.995	100.00	$100.00 \times A$	100.00	$99.50 \times B$
0.999	500.00	$500.00 \times A$	500.00	$499.50 \times B$

Note: $A = \frac{\sigma^2}{\sum x_{2i}^2}$

$$B = \frac{-\sigma^2}{\sqrt{\sum x_{2i}^2 \sum x_{3i}^2}}$$

\times = times

2. Khoảng tin cậy rộng hơn

Value of r_{23}	95% confidence interval for β_2
0.00	$\hat{\beta}_2 \pm 1.96 \sqrt{\frac{\sigma^2}{\sum x_{2i}^2}}$
0.50	$\hat{\beta}_2 \pm 1.96 \sqrt{(1.33)} \sqrt{\frac{\sigma^2}{\sum x_{2i}^2}}$
0.95	$\hat{\beta}_2 \pm 1.96 \sqrt{(10.26)} \sqrt{\frac{\sigma^2}{\sum x_{2i}^2}}$
0.995	$\hat{\beta}_2 \pm 1.96 \sqrt{(100)} \sqrt{\frac{\sigma^2}{\sum x_{2i}^2}}$
0.999	$\hat{\beta}_2 \pm 1.96 \sqrt{(500)} \sqrt{\frac{\sigma^2}{\sum x_{2i}^2}}$

Vậy xác suất chấp nhận giả thiết sai tăng lên.

3. Tỷ số t mất ý nghĩa

- Trong kiểm định $H_0: \beta_2 = 0$
- Ta sử dụng $T_{qs} = \hat{\beta}_2 / se(\hat{\beta}_2)$ so sánh với T_α . Khi có đa cộng tuyến gần hoàn hảo thì sai số tiêu chuẩn rất cao nên tỷ số T_{qs} nhỏ đi. Hậu quả là làm tăng khả năng chấp nhận H_0 .

4. R^2 cao nhưng tỷ số t ít ý nghĩa

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

- Nếu đa cộng tuyến cao thì có thể chỉ ra một vài hệ số góc không có ý nghĩa về mặt thống kê, mặc dù R^2 cao (và giá trị F có ý nghĩa).

5. Các ước lượng OLS và sai số tiêu chuẩn của chúng trở nên rất nhạy cảm đối với những thay đổi nhỏ trong số liệu. (Xem tr355 Gujarati)
6. Dấu của các ước lượng của hệ số hồi quy có thể sai.
- Ví dụ: Lý thuyết kinh tế cho biết cầu hàng hóa phụ thuộc (+) vào thu nhập, nhưng khi có đa cộng tuyến cao thì ước lượng hệ số của biến thu nhập có thể âm.
7. Thêm vào hay bớt đi các biến cộng tuyến với các biến khác, mô hình sẽ thay đổi về độ lớn của các ước lượng hoặc dấu của chúng.

5.5. Phát hiện ra sự tồn tại của đa cộng tuyến.

1. R^2 cao (>0.8) nhưng tỷ số t thấp.
2. Tương quan cặp giữa các biến giải thích cao (nhưng nếu tương quan cặp thấp thì chưa kết luận được là ko có đa cộng tuyến).
3. Xem xét tương quan riêng: Giả sử hồi quy Y với X_2, X_3, X_4 . Nếu nhận thấy $R^2_{1.234}$ cao, trong khi $r^2_{12.34}, r^2_{13.24}$, and $r^2_{14.23}$ thấp thì điều đó gợi ý các biến X_1, X_2, X_3 tương quan cao và ít nhất một trong các biến này là thừa.

4. Hồi quy phụ: là hồi quy biến X_i theo các biến giải thích còn lại, thu được R_i^2 .

$$X_i = \alpha_1 + \alpha_2 X_1 + \dots + \alpha_{k-1} X_{k-1}$$

$$H_0: R_i^2 = 0 \quad (\alpha_2 = \dots = \alpha_{k-1} = 0)$$

$$H_1: R_i^2 \neq 0$$

$$F_i = \frac{R_i^2 / (k - 2)}{(1 - R_i^2) / (n - k + 1)} \sim F(k-2, n-k+1)$$

- Nếu $F_i > F_{\alpha}(k-2, n-k+1)$: Bác bỏ H_0 . Kết luận X_i có liên hệ tuyến tính với các biến khác.

5.6. Biện pháp khắc phục

1. Sử dụng thông tin tiên nghiệm.
2. Thu thập thêm số liệu hoặc lấy thêm mẫu mới.
3. Bỏ biến.
4. Sử dụng sai phân cấp một
5. Giảm tương quan trong hồi quy đa thức
6.