CHƯƠNG
10

VẤN ĐỀ ĐA CỘNG TUYẾN VÀ CỠ MẪU NHỎ [1]

> Không có cụm từ nào được lạm dụng, cả trong sách kinh tế lượng lẫn trong tài liệu ứng dụng nhiều như cụm từ " vấn đề đa cộng tuyến." Sự thật là trong cuộc sống, chúng ta có những biến giải thích có tính cộng tuyến cao. Và hoàn toàn rõ ràng là có những thiết kế mang tính thực nghiệm **X**'**X** [nghĩa la, ma trận dữ liệu] thường được ưa chuộng hơn là nhiều thiết kế thực nghiệm tự nhiên đem lại cho chúng ta [đó là mẫu cụ thể]. Nhưng một phàn nàn về bản chất chưa tốt; có thể thấy rõ ràng của tự nhiên thì không hề mang tính góp ý xây dựng, và các phương cách đặc biệt cho một thiết kế không tốt, như hồi qui theo từng bước (stepwise regression) hoặc hồi qui dạng sóng (ridge regression), có thể hoàn toàn không thích hợp. Tốt hơn, chúng ta nên chấp nhận ngay sự việc phi thực nghiệm của chúng ta [nghĩa là, dữ liệu không được thu thập bằng những thực nghiệm đã được thiết kế] đôi khi không có nhiều thông tin về thông số mà ta quan tâm. [2]

Giả thiết 10 của *mô hình hồi qui tuyến tính cổ điển* (CLRM) là: không có quan hệ đa cộng tuyến giữa các biến hồi qui trong mô hình hồi qui. Giả thiết 7, số lần quan sát phải lớn hơn số biến hồi qui độc lập (vấn đề cỡ mẫu nhỏ), và Giả thiết 8, phải có đủ các trạng thái biến đổi trong giá trị của một biến hồi qui độc lập. Tất cả các giả thiết trên bổ sung cho giả thiết đa cộng tuyến. Trong chương này, chúng ta quan tâm đặc biệt đến giả thiết phi đa cộng tuyến bằng cách trả lời các câu hỏi sau:

1. Bản chất của đa cộng tuyến là gì?
2. Đa cộng tuyến có thật sự là một vấn đề cần phải xem xét hay không?
3. Đâu là những kết quả ứng dụng của vấn đề này?

[1] Thuật ngữ *micronumerosity* là do Arthur S. Goldberger và có nghĩa là "cỡ mẫu nhỏ." Xem cuốn A Course in Economics, Harvard University Press, Cambridge, Mass., 1991, trang 249.
[2] Edward E. Leamer, " Model Choice and Specification Analysis," (Chọn mô hình và phân tích đặc trưng) trong Zvi Griliches và Michael D. Intriligator, Handbook of Econometrics, (*Sổ tay kinh tế lượng),* số I, North Holland Publishing Company, Amsterdam, 1983, trang 300-301.

4. Bằng cách nào để nhận ra vấn đề đa cộng tuyến?
5. Sử dụng các biện pháp giải quyết gì để làm giảm bớt các vấn đề của đa cộng tuyến?
Chúng ta cũng sẽ xét xem Giả thiết 7 và 8 thích hợp với giả thiết phi đa cộng tuyến như thế nào.

10.1 BẢN CHẤT CỦA ĐA CỘNG TUYẾN

Thuật ngữ *đa cộng tuyến* do Ragnar Frisch đề nghị.[3] Khởi đầu nó có nghĩa là sự tồn tại mối quan hệ tuyến tính "hoàn hảo" hoặc chính xác giữa một số hoặc tất cả các biến giải thích trong một mô hình hồi qui.[4] Đối với hồi qui k biến liên quan đến các biến X_1, X_2, ..., X_k (với $X_1 = 1$ đối với mọi quan sát kể cả số hạng tung độ gốc), một quan hệ tuyến tính chính xác được cho là tồn tại khi thỏa điều kiện sau:

$$\lambda_1 X_1 + \lambda_2 X_2 + ... + \lambda_k X_k = 0 \qquad (10.1.1)$$

trong đó λ_1, λ_2, ..., λ_k là các hằng số và không đồng thời bằng 0.[5]

Tuy nhiên, ngày nay, thuật ngữ đa cộng tuyến được dùng với nghĩa rộng hơn, bao gồm trường hợp đa cộng tuyến hoàn hảo như (10.1.1) cũng như trường hợp các biến X có tương quan với nhau nhưng không hoàn hảo như dưới đây:[6]

$$\lambda_1 X_1 + \lambda_2 X_2 + ... + \lambda_k X_k + \nu_i = 0 \qquad (10.1.2)$$

với ν_i là số hạng sai số ngẫu nhiên.
Để thấy được sự khác biệt giữa đa cộng tuyến *hoàn hảo* và *chưa được hoàn hảo,* giả thiết, ví dụ, $\lambda_2 \neq 0$. Lúc đó (10.1.1) có thể viết lại như sau:

$$X_{2i} = -\frac{\lambda_1}{\lambda_2} X_{1i} - \frac{\lambda_3}{\lambda_2} X_{3i} - .... - \frac{\lambda_k}{\lambda_2} X_{ki} \qquad (10.1.3)$$

cho thấy X_2 tương quan tuyến tính một cách chính xác với các biến khác như thế nào hoặc có thể tìm được X_2 từ một tổ hợp tuyến tính của các biến khác như thế nào. Trong trường hợp này, hệ số

[3] Ragnar Frisch, Statistical Confluence Analysis by Means of Complete Regression Systems,(Phân tích sự hợp nhất thống kê bằng phương tiện của các hệ thống hồi qui toàn phần), Institute of Economics, Olso University, xuất bản lần 5, 1934.
[4] Nghiêm khắc mà nói thì *đa cộng tuyến* đề cập đến sự tồn tại của nhiều hơn một mối quan hệ tuyến tính chính xác, và *cộng tuyến* là nói đến sự tồn tại duy nhất một mối quan hệ tuyến tính. Nhưng sự phân biệt này hiếm khi tồn tại trong thực tế, và đa cộng tuyến được dùng cho cả hai trường hợp.
[5] Các dịp để có được một mẫu các giá trị trong đó các biến hồi qui độc lập liên quan đến mô hình này trong thực tế thật sự rất nhỏ trừ khi thiết kế, ví dụ khi số lần quan sát bé hơn số biến hồi qui độc lập hoặc khi "có biến giả" như trình bày trong chương 15. Xem bài tập 10.2.
[6] Nếu chỉ có hai biến giải thích, *tương quan giữa các biến* có thể được đánh giá bằng bậc không (zero-order) hoặc hệ số tương quan đơn. Nhưng nếu có hơn hai biến X, tương quan giữa các biến có thể được đánh giá bằng các hệ số tương quan riêng phần hoặc bằng hệ số tương quan đa biến R của một biến X với tất cả các biến X khác.

tương quan giữa biến X_2 và tổ hợp tuyến tính ở vế bên phải của phương trình (10.1.3) chắc chắn là 1 đơn vị.

Tương tự, nếu $\lambda_2 \neq 0$, công thức (10.1.2) có thể viết như sau:

$$X_{2i} = -\frac{\lambda_1}{\lambda_2} X_{1i} - \frac{\lambda_3}{\lambda_2} X_{3i} - .... - \frac{\lambda_k}{\lambda_2} X_{ki} - \frac{1}{\lambda_2} \nu_i \qquad (10.1.3)$$

cho thấy X_2 không phải là một tổ hợp tuyến tính chính xác của các biến X khác vì nó cũng còn được xác định bởi số hạng sai số ngẫu nhiên ν_i.

Để có một ví dụ số cụ thể, hãy xem dữ liệu có tính giả thuyết sau:

X_2	X_3	X_3^*
10	50	52
15	75	75
18	90	97
24	120	129
30	150	152

Có thể thấy rõ ràng là là $X_{3i} = 5X_{2i}$. Vì vậy, có sự cộng tuyến hoàn hảo giữa X_2 và X_3 bởi vì hệ số tương quan r_{23} là 1 đơn vị. Biến X_3^* được tạo thành từ X_3 đơn giản bằng cách cộng thêm các số sau, những số này được lấy từ bảng số ngẫu nhiên: 2, 0, 7, 9, 2. Bây giờ, không còn có sự cộng tuyến hoàn hảo giữa biến X_2 và X_3^*. Tuy nhiên, hai biến này tương quan chặt bởi vì tính toán cho thấy hệ số tương quan giữa chúng là 0.9959.

Phương pháp đại số trước đây liên quan đến đa cộng tuyến có thể được Ballentine mô tả cô đọng (nhớ lại hình 7.1). Trong hình này, các vòng tròn Y, X_2 và X_3 đại diện một cách tương ứng các biến đổi trong Y (biến độc lập) theo X_2 và X_3 (các biến giải thích). Mức độ cộng tuyến có thể được đánh giá bằng độ rộng của phần chung (vùng tô đen) của vòng tròn X_2 và X_3. Trong hình 10.1a, không có phần chung giữa X_2 và X_3, và vì vậy không có cộng tuyến. Trong các hình 10.1b - 10.1e, có các mức độ từ "thấp đến "cao" của sự cộng tuyến phần chung giữa X_2 và X_3 càng rộng (phần tô đen càng rộng), thì mức độ cộng tuyến càng cao. Ở trạng thái cực đoan, nếu X_2 và X_3 hoàn toàn trùng nhau (hoặc nếu X_2 hoàn toàn ở trong X_3, hay ngược lại), sự cộng tuyến là hoàn hảo.

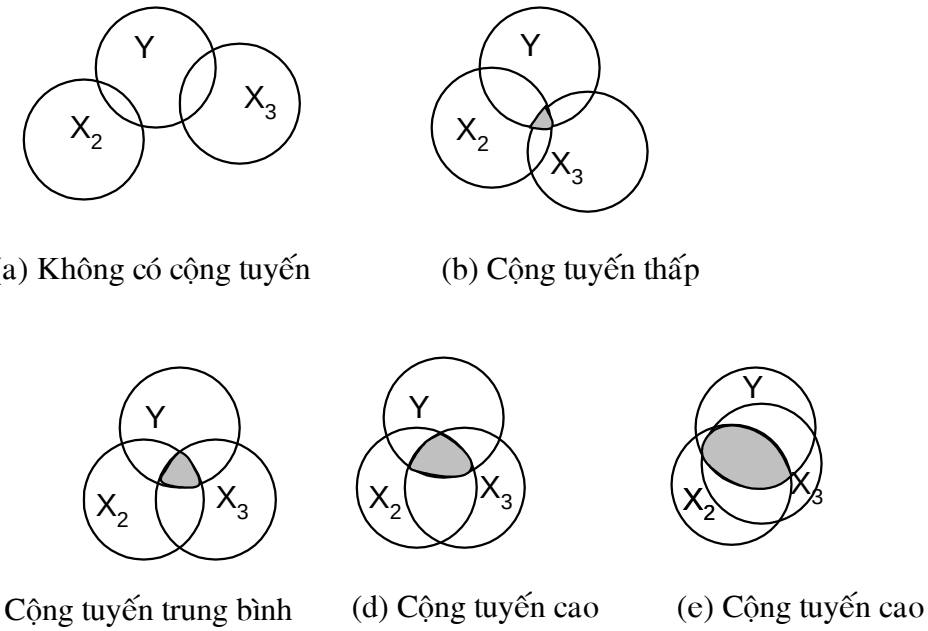
Nhân đây, lưu ý rằng đa cộng tuyến, như chúng ta đã định nghĩa, chỉ đề cập đến các quan hệ tuyến tính giữa các biến X. Nó không bỏ qua các quan hệ phi tuyến giữa các biến X. Ví dụ, xem xét mô hình hồi qui sau:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + u_i \qquad (10.1.5)$$

trong đó, Y = tổng chi phí sản xuất và X = sản lượng ra. Các biến X_i^2 (sản lượng bình phương ra) và X_i^3 (sản lượng lập phương ra) rõ ràng có quan hệ theo hàm số với X_i nhưng quan hệ này là phi tuyến. Chính xác thì những mô hình như (10.1.5) không vi phạm đến các giả định về phi đa

cộng tuyến. Tuy nhiên, trong những ứng dụng cụ thể, hệ số tương quan được đo lường một cách qui ước sẽ cho thấy X_i, X_i^2 và X_i^3 tương quan chặt, và tương quan này như chúng ta sẽ thấy, sẽ gây khó khăn cho việc ước lượng các thông số của mô hình (10.1.5) chính cao xác hơn (nghĩa là với sai số chuẩn hoá hơn).

Tại sao mô hình hồi qui tuyến tính cổ điển giả định rằng không có vấn đề đa cộng tuyến giữa các biến X? Lý do là: **Nếu đa cộng tuyến hoàn hảo theo (10.1.1), các hệ số hồi qui của các biến X là vô định và các sai số chuẩn là không xác định. Nếu đa cộng tuyến chưa hoàn hảo, như trong (10.1.2), các hệ số hồi qui, mặc dù là xác định nhưng lại có sai số chuẩn (liên quan đến bản thân các hệ số) lớn, có nghĩa là không thể ước lượng các hệ số này với độ chính xác cao.** Các phát biểu này được chứng minh trong những phần sau đây.



(a) Không có cộng tuyến (b) Cộng tuyến thấp

(c) Cộng tuyến trung bình (d) Cộng tuyến cao (e) Cộng tuyến cao

Hình 10. 1 Quan điểm của Ballentine về đa cộng tuyến

Có nhiều nguồn tạo ra đa cộng tuyến. Theo Montgomery và Peck, đa cộng tuyến có thể là do các nhân tố sau:[7]

1. *Phương pháp thu thập dữ liệu sử dụng,* ví dụ, lấy mẫu trong phạm vi các giá trị giới hạn các biến hồi qui độc lập trong tập hợp chính.

[7] Douglas Montgomery và Elizabeth Peck, *Introduction to Linear Regression Analysis (*Nhập môn phân tích hồi qui tuyến tính), John Wiley & Sons, New York, 1982, trang 289-290. Xem thêm R. L. Mason, R> L. Gunst và J. T. Webster, "Regression Analysis and Problem of Multicollinearity," (Phân tích hồi qui và vấn đề đa cộng tuyến), *Comunication in Statistics A*, quyển 4, số 3, 1975, trang 277-292; R.F. Gunst, và R. L. Manson, "Advantages of Examining Multicollinearity in Regression Analysis," (Các điều thuận lợi của việc khảo sát đa cộng tuyêén trong phân tích hồi qui), Biometrics, quyển 33, 1977, trang 249-260

2. *Các ràng buộc về mô hình hay về tổng thể được lấy mẫu.* Ví dụ, trong mô hình hồi qui của việc tiêu thụ điện theo thu nhập (X_2) và kích thước nhà ở (X_3) có một ràng buộc cụ thể về tổng thể, trong đó các gia đình có thu nhập cao hơn nói chung ở nhà rộng hơn các gia đình có thu nhập thấp hơn.

3. *Đặc trưng mô hình,* ví dụ, thêm những số hạng đa thức vào một mô hình hồi qui, đặc biệt khi khoảng giá trị của biến X nhỏ.

4. *Một mô hình xác định quá mức.* Là khi mô hình này có nhiều biến giải thích hơn số lần quan sát được. Trường hợp này thường xảy ra trong các nghiên cứu y học số bệnh nhân thì ít nhưng phải thu thập thông tin về các bệnh nhân này trên một lượng lớn các biến.

10.2 ƯỚC LƯỢNG TRONG TRƯỜNG HỢP ĐA CỘNG TUYẾN HOÀN HẢO

Như đã đề cập, trong trường hợp đa cộng tuyến hoàn hảo, các hệ số hồi qui vẫn là không xác định và các sai số chuẩn của chúng là vô hạn. Hiện tượng này có thể được giải thích dưới dạng mô hình hồi qui ba biến. Sử dụng dạng độ lệch, trong đó tất cả các biến có thể được diễn tả bằng độ lệch của chúng so với trung bình mẫu. Chúng ta có thể viết mô hình hồi qui ba biến như sau:

$$y_i = \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \hat{u}_i \qquad (10.2.1)$$

Bây giờ, theo chương 7 ta có:

$$\hat{\beta}_2 = \frac{\left(\sum y_i x_{2i}\right)\left(\sum x^2_{3i}\right) - \left(\sum y_i x_{3i}\right)\left(\sum x_{2i} x_{3i}\right)}{\left(\sum x^2_{2i}\right)\left(\sum x^2_{3i}\right) - \left(\sum x_{2i} x_{3i}\right)^2} \qquad (7.4.7)$$

$$\hat{\beta}_3 = \frac{\left(\sum y_i x_{3i}\right)\left(\sum x^2_{2i}\right) - \left(\sum y_i x_{2i}\right)\left(\sum x_{2i} x_{3i}\right)}{\left(\sum x^2_{2i}\right)\left(\sum x^2_{3i}\right) - \left(\sum x_{2i} x_{3i}\right)^2} \qquad (7.4.8)$$

Giả sử $X_{3i} = \lambda X_{2i}$, với λ là một hằng số khác 0 (ví dụ, 2, 4, 1.8. ect.). Thay vào (7.4.7) ta có

$$\hat{\beta} = \frac{\left(\sum y_i x_{2i}\right)\left(\lambda^2 \sum x^2_{2i}\right) - \left(\lambda \sum y_i x_{2i}\right)\left(\lambda \sum x^2_{2i}\right)}{\left(\sum x^2_{2i}\right)\left(\lambda^2 \sum x^2_{2i}\right) - \lambda^2 \left(\sum x^2_{2i}\right)^2} = \frac{0}{0} \qquad (10.2.2)$$

Đây là một biểu thức không xác định. Người đọc có thể kiểm tra lại là $\hat{\beta}_3$ cũng không xác định.[8]

[8] Một cách nhìn khác là: Theo định nghĩa, hệ số tương quan giữa biến X_2 và X_3 , r_{23} , là $\left(\sum x_{2i} x_{3i}\right) / \sqrt{\sum x^2_{2i} \sum x^2_{3i}}$. Nếu $r^2_{23} = 1$, đó là cộng tuyến hoàn hảo giữa X_2 và X_3 , mẫu số của (7.4.7) sẽ bằng 0, vì vậy không thể ước lượng β_2 (hoặc β_3) được.

Tại sao chúng ta có được kết quả ở biểu thức (10.2.2)? Nhớ lại ý nghĩa của $\hat{\beta}_2$:$\hat{\beta}_2$ chỉ mức độ thay đổi về giá trị trung bình của Y khi X_2 thay đổi 1 đơn vị, với điều kiện X_3 được giữ cố định. Nhưng nếu X_3 và X_2 cộng tuyến hoàn hảo thì không có cách nào để giữ cố định X_3. Khi X_2 thay đổi, thì X_3 cũng thay đổi bởi nhân tố λ. Điều đó có nghĩa là không có cách nào tách riêng các ảnh hưởng của X_2 và X_3 từ mẫu cho trước. Đối với các mục đích thực tiễn, X_2 và X_3 là không thể phân biệt được. Trong kinh tế lượng ứng dụng, vấn đề này gây thiệt hại nhiều nhất vì chủ định là tách riêng hoàn toàn các ảnh hưởng riêng phần của mỗi biến X lên biến phụ thuộc.

Để thấy được sự khác biệt này, chúng ta hãy thay $X_{3i} = \lambda X_{2i}$ vào biểu thức (10.2.1), chúng ta có biểu thức sau [xem thêm (7.1.10)]:

$$\begin{aligned} y_i &= \hat{\beta}_2 x_{2i} + \hat{\beta}_3 (\lambda x_{2i}) + \hat{u}_i \\ &= (\hat{\beta}_2 + \lambda\hat{\beta}_3) x_{2i} + \hat{u}_i \\ &= \hat{\alpha} x_{2i} + \hat{u}_i \qquad (10.2.3) \end{aligned}$$

với $$\hat{\alpha} = (\hat{\beta}_2 + \lambda\hat{\beta}_3) \qquad (10.2.4)$$

Sử dụng công thức thông dụng OLS đối với (10.2.3) ta có

$$\hat{\alpha} = (\hat{\beta}_2 + \lambda\hat{\beta}_3) = \frac{\sum x_{2i} y_i}{\sum x^2_{2i}} \qquad (10.2.5)$$

Vì vậy, mặc dù chúng ta có thể ước lượng được α, nhưng không có cách nào để ước lượng riêng β_2 và β_3; chính xác thì:

$$\hat{\alpha} = \hat{\beta}_2 + \lambda\hat{\beta}_3 \qquad (10.2.6)$$

cho chúng ta duy nhất một phương trình có hai ẩn số (lưu ý λ được cho trước) và có vô số nghiệm cho (10.2.6) ứng với các giá trị cho trước của $\hat{\alpha}$ và λ . Ví dụ với các số hạng cụ thể, $\hat{\alpha}$ = 0.8 và λ = 2. Ta có

$$0.8 = \hat{\beta}_2 + 2\hat{\beta}_3 \qquad (10.2.7)$$

hoặc

$$\hat{\beta}_2 = 0.8 - 2\hat{\beta}_3 \qquad (10.2.8)$$

Bây giờ chọn một giá trị $\hat{\beta}_3$ tùy ý, chúng ta sẽ có lời giải cho $\hat{\beta}_2$. Chọn một giá trị khác cho $\hat{\beta}_3$, chúng ta lại sẽ có một lời giải khác cho $\hat{\beta}_2$. Cho dù chúng ta cố gắng như thế nào đi nữa cũng sẽ không thể tìm được cho $\hat{\beta}_2$ một giá trị duy nhất.

Tóm lại những diều đã thảo luận ở trên là trong trường hợp đa cộng tuyến hoàn hảo, không thể có được một lời giải duy nhất cho các hệ số hồi qui riêng. Nhưng chú ý là có thể tìm được lời

giải duy nhất cho các tổ hợp tuyến tính của những hệ số này.Tổ hợp tuyến tính $(\hat{\beta}_2 + \lambda\hat{\beta}_3)$ là ước lượng duy nhất của α, với giá trị λ cho trước.[9]
Nhân đây, lưu ý rằng trong trường hợp đa cộng tuyến hoàn hảo, phương sai và sai số chuẩn của $\hat{\beta}_2$ và $\hat{\beta}_3$ không thể xác định một cách tiêng biệt được. (Xem bài tập 10.21.)

10.3 ƯỚC LƯỢNG TRONG TRƯỜNG HỢP CÓ ĐA CỘNG TUYẾN "CAO" NHƯNG "KHÔNG HOÀN HẢO"

Đa cộng tuyến hoàn hảo là một trường hợp thuộc về một thái cực. Thông thường, không tồn tại mối quan hệ tuyến tính chính xác giữa các biến X, đặc biệt là trong dữ liệu liên quan đến chuỗi thời gian kinh tế. Vì vậy, chuyển sang dùng mô hình hồi qui ba biến dưới dạng độ lệch trong (10.2.1), thay vì dùng đa cộng tuyến chính xác, chúng ta có thể có

$$x_{3i} = \lambda x_{2i} + \nu_i \qquad (10.3.1)$$

với $\lambda \neq 0$ và ν_i là số hạng sai số ngẫu nhiên do đó $\sum x_{2i}\nu_i = 0$. (Tại sao?)

Một cách ngẫu nhiên, các mô hình Ballentine trong các hình từ 10.1b đến 10.1e đại diện cho các trường hợp đa cộng tuyến không hoàn hảo.

Trong trường hợp này, các hệ số hồi qui β_2 và β_3 có thể ước lượng được. Ví dụ, thay (10.3.1) vào (7.4.5), chúng ta có

$$\hat{\beta}_2 = \frac{\sum(y_i x_{2i})\left(\lambda^2 \sum x^2_{2i} + \sum \nu^2_i\right) - \left(\lambda \sum y_i x_{2i} + \sum y_i \nu_i\right)\left(\lambda \sum x^2_{2i}\right)}{\sum x^2_{2i}\left(\lambda^2 \sum x^2_{2i} + \sum \nu^2_i\right) - \left(\lambda \sum x^2_{2i}\right)^2} \qquad (10.3.2)$$

với $\sum x_{2i}\nu_i = 0$. Có thể thiết lập một biểu thức tương tự cho $\hat{\beta}_3$.

Bây giờ, khác với (10.3.2), không có lý do gì để tin rằng (10.3.2) không thể ước lượng được. Dĩ nhiên, nếu ν_i không đủ nhỏ, hay nói cách khác không gần bằng 0, (10.3.1) sẽ mô tả sự cộng tuyến gần như hoàn hảo và chúng ta sẽ quay lại trường hợp không xác định (10.2.2)

10.4 ĐA CỘNG TUYẾN: KHÔNG CÓ CHUYỆN GÌ CẢ MÀ CŨNG LÀM RỐI LÊN? HỆ QUẢN LÝ THUYẾT CỦA ĐA CỘNG TUYẾN

Hãy nhớ lại nếu thỏa các giả định của mô hình cổ điển, các ước lượng OLS của ước lượng hồi qui là BLUE (hoặc BUE, nếu có thêm giả định chuẩn). Bây giờ có thể thấy rằng ngay cả khi đa cộng tuyến chặt, như trong trường hợp *gần đa cộng tuyến (near multicollinearity)*, các ước lượng

[9] trong tài liệu kinh tế lượng, một hàm số như $(\hat{\beta}_2 + \lambda\hat{\beta}_3)$ được gọi là **hàm có thể ước lượng được** (estimable function).

QLS vẫn có tính chất của BLUE.[10] Vậy vấn đề đa cộng tuyến làm ầm lên về chuyện gì? Như Christopher Achen nhận xét (lưu ý thêm điều Leamer đã đề cập đến trong phần mở đầu của chương này):

> Những sinh viên khi bắt đầu học phương pháp luận đôi khi lo lắng rằng các biến độc lập của họ có tương quan với nhau cái gọi là vấn đề đa cộng tuyến. Nhưng vấn đề đa cộng tuyến không vi phạm các giả định. Các ước lượng nhất quán không thiên lệch chắc chắn sẽ xảy ra và các sai số chuẩn của chúng cũng sẽ được ước lượng một cách chính xác. Ảnh hưởng duy nhất của đa cộng tuyến là gây khó khăn cho việc đạt được các ước lượng hệ số với sai số chuẩn nhỏ. Nhưng số lần quan sát ít cũng gây nên tác động đến biến độc lập với phương sai nhỏ. (Nói tóm lại, ở mức độ lý thuyết, đa cộng tuyến, số lần quan sát bé, và phương sai nhỏ trên các biến độc lập đều là một vấn đề giống nhau.) Vì vậy câu hỏi " Tôi nên làm gì với đa cộng tuyến?" thì giống như câu hỏi "Tôi nên làm gì nếu tôi có số lần quan sát ít?". Không có một câu trả lời thống kê nào cho vấn đề này.[11]

Quay lại với tầm quan trọng của cỡ mẫu, Goldberger đã đặt ra thuật ngữ **cỡ mẫu nhỏ** (micronumerosity), để đối lại từ đa âm tiết ngoại lai *multicollinearity (đa cộng tuyến)*. Theo Goldberger, **cỡ mẫu nhỏ chính xác** (exact micronumerosity) (tương ứng của đa cộng tuyến chính xác) xảy ra khi *n,* kích thước mẫu , bằng 0, trong trường hợp đó, mọi ước lượng là không thể được. C*ỡ mẫu gần như nhỏ (near micronumerosity)*, giống như gần như đa cộng tuyến hoàn hảo, xảy ra khi số lần quan sát vừa đủ vượt quá số thông số được ước lượng.

Leamer, Achen và Goldberger đã đúng khi họ tiếc là đã thiếu quan tâm đến vấn đề cỡ mẫu mà lại quan tâm quá mức đến vấn đề đa cộng tuyến. Đáng tiếc thay, trong khi ứng dụng các dữ liệu thứ cấp (đó là các dữ liệu được một số tổ chức thu thập, như là dữ liệu về GNP do chính phủ thu thập), một nhà nghiên cứu tư nhân có lẽ không thể quan tâm nhiều đến kích thước của dữ liệu mẫu và có lẽ phải đối phó với " các vấn đề về ước lượng đủ quan trọng để biện hộ cho việc chúng ta xử lý vấn đề này [vấn đề đa cộng tuyến] như một sự vi phạm mô hình CLR [mô hình hồi qui cổ điển]". [12]

Thứ nhất, đúng là ngay cả trong trường hợp gần như đa cộng tuyến các hàm ước lượng OLS cũng không thiên lệch. Nhưng sự không thiên lệch là một tính chất của mẫu bội hoặc là việc lấy mẫu lập lại. Điều này có nghĩa là, giữ cố định các giá trị của biến X, nếu có được các mẫu lập lại và tính các hàm ước lượng OLS cho những mẫu này, thì trung bình của các giá trị mẫu sẽ hội tụ về các giá trị thực của tổng thể của các ước lượng khi số lượng mẫu tăng. Nhưng điều này không nói lên điều gì về các tính chất của các hàm ước lượng trong một mẫu cho trước bất kỳ.

---

[10] Bởi vì gần như đa cộng tuyến tự thân nó không vi phạm các giả định khác đã được liệt kê trong chương 7, các ước lượng OLS là BLUE như đã xác định.

[11] Christopher H. Achen, *Interpreting and Using Regression, (Diễn dịch và Sử dụng Hồi qui),* Sage Publications, Beverly Hills, Calif., 1982, trang 82-83.

[12] Peter Kennedy, *Hướng dẫn môn Kinh tế lượng, (A guide to economics),* 3d ed., The MIT Press, Cambride, Mass., 1992, trang 177.

Thứ hai, cũng đúng là cộng tuyến không xóa bỏ tính chất phương sai nhỏ nhất: Trong loại các hàm ước lượng không thiên lệch tuyến tính, các hàm ước lượng OLS có phương sai nhỏ nhất; nghĩa là, các hàm ước lượng này có hiệu quả. Nhưng không có nghĩa là phương sai của một hàm ước lượng OLS sẽ phải nhất thiết nhỏ (tương đối so với giá trị của hàm ước lượng này) trong bất kỳ mẫu cho trước nào, như chúng ta sẽ chứng minh một cách ngắn gọn.

Thứ ba, *đa cộng tuyến đặc biệt là một hiện tượng mẫu (hồi qui)* theo nghĩa là cho dù các biến X không tương quan tuyến tính trong tổng thể, chúng cũng có thể tương quan trong một mẫu cụ thể nào đó: Khi chúng ta đặt ra lý thuyết hoặc là hàm hồi qui tổng thể (population regression function - PRF), chúng ta tin rằng mọi biến X trong mô hình này có ảnh hưởng riêng biệt hoặc độc lập đến biến phụ thuộc Y. Nhưng có thể là trong một mẫu cho trước bất kỳ được sử dụng để kiểm tra PRF một số hoặc toàn bộ các biến X đều cộng tuyến cao đến độ chúng ta không thể tách ảnh hưởng của riêng từng biến lên Y. Vì vậy có thể nói mẫu của chúng ta khiến công việc của chúng ta xấu đi mặc dù lý thuyết cho rằng mọi biến X đều quan trọng. Tóm lại, mẫu có thể không đủ "giàu" để chứa được mọi biến X trong phân tích.

Để minh họa, xem lại ví dụ về tiêu dùng - thu nhập trong chương 3. Các nhà kinh tế lượng lý luận rằng, ngoài thu nhập, sự giàu có của người tiêu dùng cũng là một yếu tố quyết định quan trọng của chi tiêu cho tiêu dùng. Vì vậy, chúng ta có thể viết

$$\text{Tiêu dùng}_i = \beta_1 + \beta_2 \text{ Thu nhập}_i + \beta_3 \text{ Sự giàu có}_i + u_i$$

Bây giờ có vẻ như khi chúng ta có dữ liệu về thu nhập và sự giàu có, hai biến này có lẽ tương quan chặt, nếu không muốn nói là hoàn hảo: Những người giàu có hơn thường có thu nhập cao hơn. Vì vậy, mặc dù trong lý thuyết về thu nhập và sự giàu có là những nhân tố logic để giải thích hành vi chi tiêu cho tiêu dùng, trong thực tế (đó là trong mẫu) khó có thể phân biệt được các tác động riêng biệt của thu nhập và sự giàu có đến chi tiêu cho tiêu dùng.

Một cách lý tưởng, để đánh giá các tác động riêng biệt của sự giàu có và thu nhập lên chi tiêu cho tiêu dùng chúng ta cần có đủ số quan sát mẫu về những cá nhân giàu có với thu nhập thấp, và những người có thu nhập cao nhưng ít giàu (nhớ lại giả định 8). Mặc dù điều này có vẻ như có thể thực hiện trong những nghiên cứu chéo liên khu vực (cross-sectional studies) (bằng cách tăng cỡ mẫu), nhưng rất khó đạt được trong chuỗi thời gian tổng hợp (aggregate time series work).

Vì tất cả các lý do trên, sự thật là các hàm ước lượng OLS là BLUE mặc dù đa cộng tuyến có rất ít cách giải quyết trong thực tế. Chúng ta phải xem những gì xảy ra hoặc có vẻ như sẽ xảy ra trong một mẫu cho trước bất kỳ, đề tài này được thảo luận trong phần sau.

10.5 HỆ QUẢ THỰC TẾ CỦA ĐA CỘNG TUYẾN

Trong các trường hợp gần như đa cộng tuyến hoặc đa cộng tuyến cao, chúng ta thường phải đối đầu với các hệ quả sau:

1. Mặc dù BLUE, nhưng các hàm ước lượng OLS có phương sai và đồng phương sai lớn, gây khó khăn cho việc ước lượng chính xác.
2. Vì hệ quả 1, khoảng tin cậy có khuynh hướng rộng hơn nhiều, dẫn đến việc dễ dàng chấp nhận "giả thiết H_0 zero" (zero null-hypothesis) (đó là hệ số thực của tập hợp chính bằng 0) hơn.
3. Cũng vì hệ quả 1, tỷ số t của một hoặc nhiều hệ số có khuynh hướng không có ý nghĩa thống kê.
4. Mặc dù tỷ số t của một hoặc nhiều hệ số không có ý nghĩa thống kê, R^2, dùng để đánh giá độ thích hợp, có thể rất cao.
5. Các hàm ước lượng OLS và các sai số chuẩn của chúng có thể rất nhạy đối với các thay đổi nhỏ trong dữ liệu.

Các hệ quả trên có thể được xác định như sau.

Phương sai và đồng phương sai của các ước lượng OLS lớn

Để thấy được phương sai và đồng phương sai lớn, hãy nhớ lại đối với mô hình (10.2.1) phương sai và đồng phương sai của $\hat{\beta}_2$ và $\hat{\beta}_3$ được tính như sau

$$\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x^2_{2i}(1 - r^2_{23})} \qquad (7.4.12)$$

$$\text{var}(\hat{\beta}_3) = \frac{\sigma^2}{\sum x^2_{3i}(1 - r^2_{23})} \qquad (7.4.15)$$

$$\text{cov}(\hat{\beta}_2, \hat{\beta}_3) = \frac{-r^2_{23}\sigma^2}{(1 - r^2_{23})\sqrt{\sum x^2_{2i}\sum x^2_{3i}}} \qquad (7.4.17)$$

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

Từ (7.4.12) và (7.4.15) ta thấy rõ ràng khi r_{23} tiến đến 1, đó là khi sự cộng tuyến gia tăng, phương sai của hai hàm ước lượng tăng và trong giới hạn khi $r_{23} = 1$, các hàm ước lượng này là vô hạn. Từ (7.4.17) cũng rõ ràng là khi r_{23} tiến đến 1, đồng phương sai của hai ước lượng cũng tăng về giá trị tuyệt đối.[Chú ý:$\text{cov}(\hat{\beta}_2, \hat{\beta}_3) = \text{cov}(\hat{\beta}_3, \hat{\beta}_2)$]

Tốc độ gia tăng của phương sai và đồng phương sai có thể thấy được qua **yếu tố lạm phát phương sai** (variance-inflation factor _ VIF), được định nghĩa như sau

$$\text{VIF} = \frac{1}{(1 - r^2_{23})} \qquad (10.5.1)$$

VIF cho thấy phương sai của một hàm ước lượng *tăng nhanh* như thế nào bởi sự hiện diện của đa cộng tuyến. Khi r^2_{23} bằng 1, VIF tiến đến vô hạn. Đó là khi độ cộng tuyến gia tăng, phương sai của hàm ước lượng gia tăng, và trong giới hạn của độ cộng tuyến, phương sai có thể trở thành vô hạn. Như đã thấy, nếu không có cộng tuyến giữa X_2 và X_3, VIF sẽ bằng 1.

Sử dụng định nghĩa này, chúng ta có thể diễn tả (7.4.12) và (7.4.15) như sau

$$\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x^2_{2i}}\text{VIF} \qquad (10.5.2)$$

$$\text{var}(\hat{\beta}_3) = \frac{\sigma^2}{\sum x^2_{3i}}\text{VIF} \qquad (10.5.3)$$

các biểu thức cho thấy phương sai của $\hat{\beta}_2$ và $\hat{\beta}_3$ tỷ lệ với VIF.

Để có khái niệm về phương sai và đồng phương sai tăng như thế nào khi r_{23} tăng, hãy xem bảng 10.1, trong đó trình bày các giá trị phương sai và đồng phương sai ứng với các giá trị của r_{23}. Như trong bảng này, gia tăng r_{23} có ảnh hưởng nghiêm trọng đến phương sai và đồng phương sai ước lượng của các hàm ước lượng OLS. Khi r_{23} = 0.50, $\text{var}(\hat{\beta}_2)$ bằng 1.33 lần phương sai khi r_{23} = 0, nhưng khi r_{23} bằng 0.95 thì $\text{var}(\hat{\beta}_2)$ lớn gấp 10 lần khi không có đa cộng tuyến. Và kỳ lạ thay, khi r_{23} tăng từ 0,95 đến 0.995 đã làm phương sai ước lượng tăng gấp 100 lần so với khi không có cộng tuyến. Ảnh hưởng nghiêm trọng này cũng tương tự đối với đồng phương sai. Tất cả điều này có thể thấy qua hình 10.2

Nhân tiện, các kết quả vừa được thảo luận trên đây cũng có thể dễ dàng mở rộng cho mô hình k biến (xem bài tập 10.15 và 10.16).

Bảng 10. 1 Ảnh hưởng của sự gia tăng r_{23} đến $\text{var}(\hat{\beta}_2)$ và $\text{cov}(\hat{\beta}_2, \hat{\beta}_3)$

Giá trị của 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	$\frac{\sigma^2}{\sum x^2_{2i}} = A$	—	0
0.50	1.33	1.33xA	1.33	0.67xB
0.70	1.96	1.96xA	1.96	1.37xB
0.80	2.78	2.78xA	2.78	2.22xB
0.90	5.76	5.76xA	5.76	4.73xB
0.95	10.26	10.26xA	10.26	9.74xB
0.97	16.92	16.92xA	16.92	16.41xB
0.99	50.25	50.25xA	50.25	49.75xB
0.995	100.00	100.00xA	100.00	99.50xB
0.999	500.00	500.00xA	500.00	499.50xB

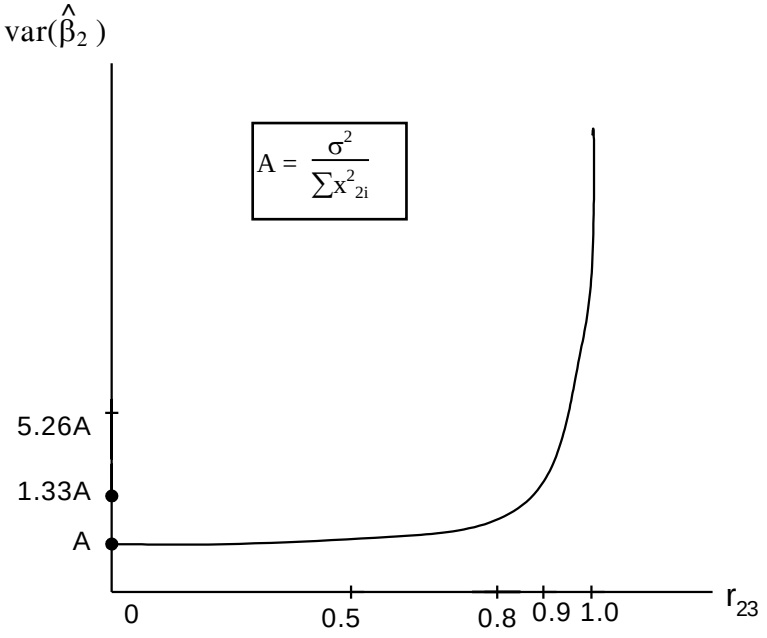
Ghi chú: $A = \frac{\sigma^2}{\sum x^2_{2i}}$

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

* Để tìm ảnh hưởng của sự gia tăng r_{23} lên var $(\hat{\beta}_3)$, chú ý là $A = \frac{\sigma^2}{\sum x^2_{3i}}$ khi r_{23} = 0, nhưng các yếu tố phóng đại phương sai và đồng phương sai vẫn giữ nguyên

Khoảng tin cậy rộng hơn

Vì các sai số chuẩn lớn nên khoảng tin cậy đối với các thông số tổng thể liên quan cũng có khuynh hướng lớn hơn, có thể thấy từ bảng 10.2. Ví dụ, khi $r_{23} = 0.95$, khoảng tin cậy cho β_2 lớn hơn $\sqrt{10.26}$ so với khi $r_{23} = 0$, khoảng bằng 3.



Hình 10. 2 var$(\hat{\beta}_2)$ như là một hàm của r_{23}.

Bảng 10. 2 Tác động của sự gia tăng cộng tuyến lên khoảng tin cậy 95% đối với $\hat{\beta}_2$: $\hat{\beta}_2 \pm 1.96$ se$(\hat{\beta}_2)$

Giá trị của r_{23}	Độ tin cậy 95% cho $\hat{\beta}_2$

0.00	$\hat{\beta}_2 \pm 1.96\sqrt{\frac{\sigma^2}{\sum x^2_{2i}}}$
0.50	$\hat{\beta}_2 \pm 1.96\sqrt{(1.33)}\sqrt{\frac{\sigma^2}{\sum x^2_{2i}}}$
0.95	$\hat{\beta}_2 \pm 1.96\sqrt{(10.26)}$
0.99	$\sqrt{\frac{\sigma^2}{\sum x^2_{2i}}}$
0.999	$\hat{\beta}_2 \pm 1.96\sqrt{(100)}\sqrt{\frac{\sigma^2}{\sum x^2_{2i}}}$
	$\hat{\beta}_2 \pm 1.96\sqrt{(500)}\sqrt{\frac{\sigma^2}{\sum x^2_{2i}}}$

Chú ý: Chúng ta đang sử dụng phân phối chuẩn vì để thuận tiện ta giả định là đã biết σ^2. Vì vậy sử dụng 1.96 và khoảng tin cậy 95% cho phân phối chuẩn.
Sai số chuẩn tùy thuộc vào các giá trị khác nhau của r_{23} được lấy từ bảng 10.1.

Do đó, trong trường hợp đa cộng tuyến cao, dữ liệu mẫu có thể thích hợp với một tập hợp nhiều loại giả thiết. Chính vì vậy, xác suất để chấp nhận giả thiết sai (đó chính là sai lầm loại II) gia tăng.

Tỉ số t "không có ý nghĩa"

Nhớ lại là để kiểm tra giả thiết H_o:$\beta_2 = 0$, chúng ta sử dụng tỉ số t, đó là $\hat{\beta}_2$ /se($\hat{\beta}_2$), và so sánh giá trị ước lượng của t với giá trị *t* tới hạn từ bảng t. Nhưng như chúng ta đã thấy, trong trường hợp cộng tuyến cao sai số chuẩn ước lượng tăng nghiêm trọng, do đó làm cho giá trị t nhỏ hơn. Chính vì vậy, trong những trường hợp như thế, chúng ta sẽ dễ dàng chấp nhận giả thiết H_0 là giá trị tương ứng thực của tổng thể là bằng 0.[13]

R^2 cao nhưng tỷ số t ít có ý nghĩa.

Xem mô hình hồi qui tuyến tính k biến sau:

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

Trong trường hợp đa cộng tuyến cao, thì có thể tìm thấy, như chúng ta đã lưu ý là một hoặc nhiều hệ số độ dốc riêng phần sẽ không có ý nghĩa thống kê quan trọng dựa trên cở sở kiểm định t. Tuy nhiên, R^2 trong những trường hợp này lại rất cao, trên 0.9, vậy dựa trên kiểm định F thì có thể bác bỏ giả thiết cho rằng $\beta_2 = \beta_3 = ... = \beta_k = 0$. Thật sự thì đây là một trong những dấu hiệu của đa cộng tuyến – giá trị t không có ý nghĩa nhưng R^2 lại cao (và giá trị F có ý nghĩa)!

[13] Nói theo ngôn ngữ của khoảng tin cậy, giá trị $\beta_2 = 0$ sẽ càng gia tăng khả năng nằm trong vùng chấp nhận khi mức độ cộng tuyến gia tăng.

Chúng ta sẽ xác định dấu hiệu này trong phần sau, nhưng kết luận này không có gì đáng ngạc nhiên trong thảo luận của chúng ta về kiểm định riêng biệt so với kiểm định liên kết trong chương 8. Như bạn có thể nhớ lại, vấn đề thực sự ở đây là đồng phương sai giữa các hàm ước lượng, mà như công thức (7.4.17) cho thấy, thì liên quan đến mối tương quan giữa các biến hồi qui độc lập.

Độ nhạy của hàm ước lượng OLS và của sai số chuẩn của các hàm này đối với những thay đổi nhỏ trong dữ liệu

Chỉ cần đa cộng tuyến không hoàn hảo thì việc ước lượng các hệ số hồi qui có thể thực hiện được nhưng các giá trị ước lượng và sai số chuẩn của chúng trở nên vô cùng nhạy ngay cả đối với thay đổi nhỏ nhất trong số liệu.

Để thấy được điều này, xem Bảng 10.3. Dựa trên những số liệu này, chúng ta có hàm hồi qui bội sau:

$$\begin{aligned} \hat{Y}_i &= 1.1939 + 0.4463X_{2i} + 0.0030X_{3i} \\ &\quad (0.7737) \quad (0.1848) \quad (0.0851) \\ t &= (1.5431) \quad (2.4151) \quad (0.0358) \qquad (10.5.4) \\ &\quad R^2 = 0.8101 \qquad r_{23} = 0.5523 \\ \text{cov}(\hat{\beta}_2, \hat{\beta}_3) &= -0.00868 \qquad df = 2 \end{aligned}$$

Hàm hồi qui (10.5.4) cho thấy không có hệ số hồi qui nào tự thân có ý nghĩa ở mức ý nghĩa qui ước là 1 hoặc 5%, mặc dù $\hat{\beta}_2$ có ý nghĩa ở mức ý nghĩa 10% dựa trên kiểm định t một phía.

Bây giờ xem xét Bảng 10.4. Khác biệt duy nhất giữa Bảng 10.3 và Bảng 10.4 là giá trị thứ ba và thứ tư của X_3 đổi chỗ cho nhau. Sử dụng số liệu trong Bảng 10.4, bây giờ ta có:

$$\begin{aligned} \hat{Y}_i &= 1.2108 + 0.4014X_{2i} + 0.0270X_{3i} \\ &\quad (0.7480) \quad (0.2721) \quad (0.1252) \\ t &= (1.6187) \quad (1.4752) \quad (0.2158) \qquad (10.5.5) \\ &\quad R^2 = 0.8143 \qquad r_{23} = 0.8258 \\ \text{cov}(\hat{\beta}_2, \hat{\beta}_3) &= -0.0282 \qquad df = 2 \end{aligned}$$

Bảng 10. 3
Số liệu lý thuyết của Y, X_2, và X_3

Y	X_2	X_3
1	2	4
2	0	2
3	4	12
4	6	0
5	8	16

Bảng 10. 4
Số liệu lý thuyết của Y, X_2, và X_3

Y	X_2	X_3
1	2	4
2	0	2
3	4	0
4	6	12
5	8	16

Do kết quả của một thay đổi nhỏ trong số liệu, chúng ta có thể thấy rằng $\hat{\beta}_2$, giá trị mà đã có ý nghĩa thống kê trước đây ở mức ý nghĩa 10%, hiện giờ không còn có ý nghĩa ở mức ý nghĩa này nữa. Cũng lưu ý rằng trong (10.5.4) cov($\hat{\beta}_2$,$\hat{\beta}_3$) = -0.00868 trong khi trong (10.5.5) giá trị này là -0.0282 tăng gấp 3 lần. Tất cả những thay đổi này có lẽ đã góp phần làm gia tăng đa cộng tuyến. Trong (10.5.4) r_{23} = 0.5523, trong khi trong (10.5.5) giá trị này lại là 0.8285. Tương tự, các sai số chuẩn của $\hat{\beta}_2$ và $\hat{\beta}_3$ tăng giữa hai hàm hồi qui, đó là hiện tượng thường gặp của cộng tuyến.

Trước đây chúng ta lưu ý là với đa cộng tuyến cao, ta không thể ước lượng được các hệ số hồi qui riêng phần một cách chính xác nhưng tổ hợp tuyến tính của các hệ số này lại có thể được ước lượng chính xác. Sự việc này có thể được chứng minh bằng các hàm hồi qui (10.5.4) và (10.5.5). Trong hàm hồi qui đầu, tổng của hai hệ số độ dốc riêng phần là 0.4493 và trong hàm thứ hai thì giá trị này là 0.4284, gần như là một. Không chỉ như thế, các sai số chuẩn cũng gần như giống nhau, 0.1550 và 0.1823.[14] Tuy nhiên, lưu ý rằng hệ số của X_3 đã thay đổi nghiêm trọng, từ 0.003 đến 0.027.

Hệ quả của cỡ mẫu nhỏ

Rập khuôn theo các hệ quả của đa cộng tuyến, và một cách hài hước, Goldberger trích dẫn chính xác các hệ quả tương tự của cỡ mẫu nhỏ, đó là, phân tích dựa trên cỡ mẫu nhỏ.[15] Người đọc nên xem phân tích của Goldberger để hiểu tại sao ông ta coi cỡ mẫu nhỏ quan trọng (hoặc không quan trọng) tương tự như đa cộng tuyến.

10.6 VÍ DỤ MINH HỌA: CHI TIÊU CHO TIÊU DÙNG TRONG QUAN HỆ VỚI THU NHẬP VÀ SỰ GIÀU CÓ

Để minh họa những điểm đã thảo luận trên đây, chúng ta hãy xem lại ví dụ tiêu thụ-thu nhập trong chương 3. Trong bảng 10.5 chúng ta lấy lại số liệu của bảng 3.2 và thêm vào đó số liệu về sự giàu có của người tiêu dùng, sau đó, dựa vào bảng 10.5 chúng ta có các hàm hồi qui sau:

$$\begin{aligned} \hat{Y}_i &= 24.7747 + 0.9415X_{2i} - 0.0424X_{3i} \\ &\quad (6.7525) \quad (0.8229) \quad (0.0.807) \\ t &= (3.6690) \quad (1.1442) \quad (-0.5261) \qquad (10.6.1) \end{aligned}$$

$$R^2 = 0.9635 \quad \bar{R}^2 = 0.9531 \quad df = 7$$

Hàm hồi qui (10.6.1) cho thấy thu nhập và sự giàu có cùng giải thích về việc 96% của sự biến đổi về chi tiêu cho tiêu dùng, và tuy nhiên không có hệ số độ dốc nào có ý nghĩa thống kê riêng

[14] Các sai số chuẩn này được tính theo công thức

$$se(\hat{\beta}_2 + \hat{\beta}_3) = \sqrt{var(\hat{\beta}_2) + var(\hat{\beta}_3) + 2cov(\hat{\beta}_2, \hat{\beta}_3)}$$

[15] Goldberger, op. cit., trang 248-250.

biệt. Hơn thế nữa, biến giàu có không những chỉ có ý nghĩa thống kê mà còn có dấu sai. Một tiên nghiệm, thường thì chúng ta kỳ vọng một tương quan dương giữa tiêu dùng và sự giàu có. Mặc dù $\hat{\beta}_2$ và $\hat{\beta}_3$ không có ý nghĩa thống kê riêng biệt, nếu chúng ta kiểm định giả thiết cho rằng $\hat{\beta}_2 = \hat{\beta}_3$ và đồng thời bằng 0, giả thiết này có thể bị bác bỏ, như bảng 10.6 cho thấy. Với giả định thường gặp chúng ta có

$$F = \frac{4282.7770}{46.3494}$$
$$= 92.4019 \qquad (10.6.2)$$

Giá trị F này rõ ràng rất có ý nghĩa.

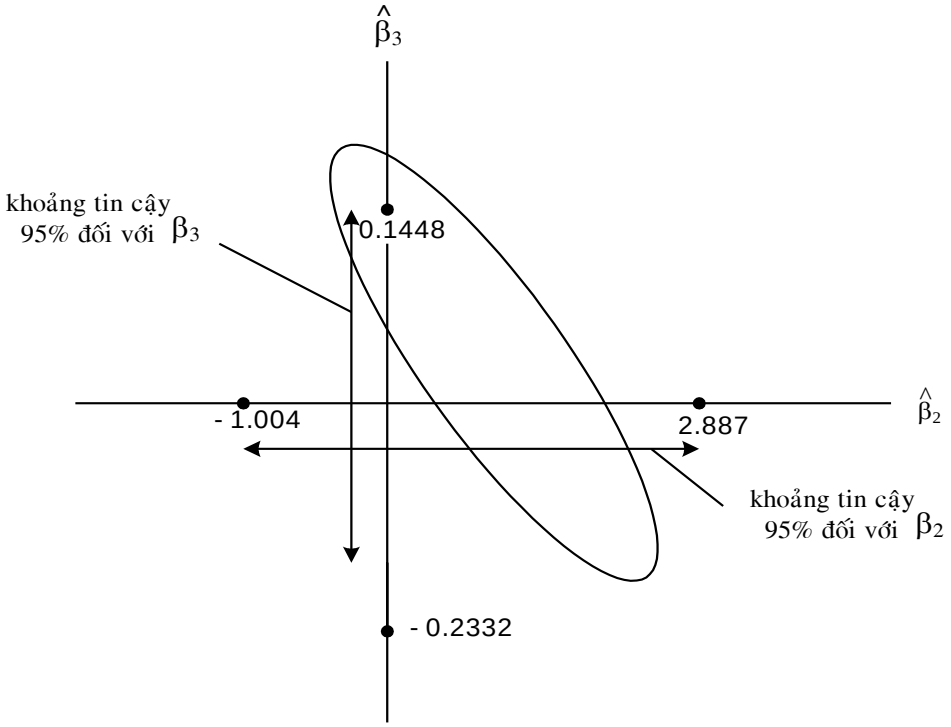
Rất thú vị nếu nhìn kết quả này dưới dạng hình học. (Hình 10.3). Dựa vào hàm hồi qui (10.6.1), chúng ta đã thiết lập khoảng tin cậy 95% cho β_2 và β_3 theo thủ tục thông thường đã thảo luận ở chương 8. Như những khoảng này cho thấy, riêng mỗi khoảng đều có chứa giá trị 0. ì vậy, một cách riêng biệt, chúng ta có thể chấp nhận giả thiết cho rằng: hai hệ số độ dốc riêng phần đồng thời bằng 0. Nhưng khi chúng ta thiết lập một khoảng tin cậy kết hợp để kiểm định giả thiết là $\hat{\beta}_2 = \hat{\beta}_3 = 0$, giả thiết này không thể chấp nhận được vì khoảng tin cậy liên kết, thật sự là hình elip, không chứa điểm 0.[16]. Như đã trình bày, khi cộng tuyến cao, thì kiểm định các biến hồi qui độc lập riêng biệt không đáng tin cậy; trong những trường hợp như vậy, kiểm định F tổng thể sẽ cho thấy có mối quan hệ giữa Y và các biến hồi qui độc lập khác hay không.

Ví dụ của chúng ta trình bày một cách nghiêm trọng những gì mà vấn đề cộng tuyến gây ra. Sự thực là, kiểm định F là có ý nghĩa nhưng các giá trị t của X_2 và X_3 riêng biệt thì không có ý nghĩa; tức là hai biến này tương quan chặt đến độ không thể tách riêng các ảnh hưởng cá nhân của thu nhập hoặc sự giàu có đến tiêu dùng. Từ sự kiện này, nếu chúng ta lập hàm hồi qui của X_3 theo X_2, ta có

$$\hat{X}_{3i} = 7.5454 + 10.1909X_{2i}$$
$$(29.4758) \quad (0.1643) \qquad (10.6.3)$$
$$t = (0.2560) \quad (62.0405) \qquad R^2 = 0.9979$$

cho thấy là có sự đa cộng tuyến gần như hoàn hảo giữa X_3 và X_2.

[16] Như đã lưu ý ở phần 5.3, đề tài về khoảng tin cậy liên kết phức tạp hơn. Độc giả quan tâm có thể xem phần tham khảo được trích ở đó.



Hình 10. 3: Khoảng tin cậy riêng cho β_2 và β_3 và khoảng tin cậy kết hợp (elip) cho β_2 và β_3

Bây giờ chúng ta xem điều gì xảy ra nếu chúng ta lập hàm hồi qui của Y chỉ theo X_2.

$$\hat{Y}_i = 24.4545 + 0.5091X_{2i}$$
$$(6.4138) \quad (0.0357) \qquad (10.6.4)$$
$$t = (3.8128) \quad (14.2432) \qquad R^2 = 0.9621$$

Trong (10.6.1) biến thu nhập đã không có ý nghĩa thống kê trong khi bây giờ biến này lại có ý nghĩa cao. Nếu thay vì lập hồi qui Y theo X_2 ta lập hàm hồi qui theo X_3, ta có

$$\hat{Y}_i = 24.411 + 0.0498X_{2i}$$
$$(6.874) \quad (0.0037) \qquad (10.6.5)$$
$$t = (3.551) \quad (13.29) \qquad R^2 = 0.9567$$

Chúng ta thấy là sự giàu có bây giờ có ảnh hưởng quan trọng đến chi tiêu cho tiêu dùng, trong khi ở (10.6.1) biến này không có ảnh hưởng đến chi tiêu cho tiêu dùng.

Các hàm hồi qui (10.6.4) và (10.6.5) trình bày khá rõ ràng là trong những trường hợp cực đoan của đa cộng tuyến bỏ qua biến cộng tuyến cao thường sẽ khiến cho biến X khác có ý nghĩa thống kê. Kết quả này đưa ra cách để tránh khỏi vấn đề cộng tuyến cực đoan là bỏ qua biến cộng tuyến, nhưng chúng ta sẽ thảo luận vấn đề này nhiều hơn ở phần 10.8.

10.7 PHÁT HIỆN VẤN ĐỀ ĐA CỘNG TUYẾN

Sau khi tìm hiểu bản chất và các hệ quả của đa cộng tuyến, câu hỏi thường đặt ra là: bằng cách nào chúng ta biết được cộng tuyến tồn tại trong một tình huống cho trước, đặc biệt là trong những mô hình liên quan đến nhiều hơn hai biến giải thích? Lúc này, thật là hữu ích nếu chúng ta nằm lòng những khuyến cáo của Kmenta:

1. Đa cộng tuyến là một câu hỏi về mức độ, không phải về sự phân biệt có ý giữa sự hiện diện hay không hiện diện của đa cộng tuyến mà là giữa các mức độ khác nhau của đa cộng tuyến.
2. Vì đa cộng tuyến đề cập đến điều kiện của các biến giải thích đã được giả định là không ngẫu nhiên, đây là đặc điểm của mẫu chứ không phải của tổng thể.

Vì vậy, chúng ta không "kiểm định đa cộng tuyến" nhưng có thể, nếu chúng ta muốn, đo lường mức độ đa cộng tuyến trong bất kỳ một mẫu cụ thể nào.[17]

Bởi vì đa cộng tuyến là một hiện tượng mẫu rất quan trọng xuất hiện ngoài tập số liệu phi thực nghiệm lớn được thu thập trong hầu hết các ngành khoa học xã hội, chúng ta không có một phương pháp duy nhất nào để phát hiện nó hoặc đo lường độ mạnh của nó. Những gì chúng ta có là một vài qui tắc kinh nghiệm, một số thông thường và một số ngoại lệ, nhưng các qui tắc kinh nghiệm thì đều giống nhau. Bây giờ chúng ta xem xét một vài trường hợp của các qui tắc kinh nghiệm này.

1. **R^2 cao nhưng tỷ số *t* ít có ý nghĩa.** Như đã lưu ý, đây là hiện tượng " cổ điển" của đa cộng tuyến. Nếu R^2 cao hơn 0.8, kiểm định F trong hầu hết các trường hợp sẽ bác bỏ giả thiết: các hệ số độ dốc riêng phần đồng thời bằng 0, nhưng các kiểm định *t* riêng biệt sẽ cho thấy là không có hoặc rất ít các hệ số độ dốc này khác không, theo ý nghĩa thống kê. Sự thật này đã được minh họa rõ ràng bằng ví dụ của chúng ta về tiêu dùng - thu nhập - sự giàu có.

   Mặc dù chuẩn đoán này là hợp lý, nhưng khuyết điểm của nó là "quá nhấn mạnh theo hướng là đa cộng tuyến được xem như có hại chỉ khi mọi ảnh hưởng của các biến giải thích lên biến Y không thể tách riêng được."[18]

2. **Các hệ số tương quan từng đôi** (pair-wise correlations) **giữa các biến hồi qui độc lập**. Một qui tắc kinh nghiệm khác được nêu ra là nếu hệ số tương quan từng đôi hoặc bậc 0 giữa hai biến hồi qui độc lập cao, trên 0.8, thì đa cộng tuyến trở thành một vấn đề nghiêm trọng. Vấn đề đối với tiêu chuẩn này là, mặc dù hệ số tương quan bậc 0 cao có thể cho là có cộng tuyến, nhưng không nhất thiết là các hệ số này phải cao thì mới có sự cộng tuyến trong mọi trường hợp cụ thể. Nói theo kỹ thuật, *tương quan bậc 0 cao là điều kiện đủ nhưng không phải là điều kiện cần cho sự hiện diện của đa cộng tuyến vì đa cộng tuyến có thể tồn tại ngay*

[17] Jan Kmenta, *Elements of Econometrics,* (Các thành tố của Kinh tế lượng), 2d., ed., Macmillan, New York, 1986, p. 431.

cả khi hệ số tương quan đơn hoặc hệ số tương quan bậc 0 tương đối thấp (nhỏ hơn 0.50). Để thấy mối liên hệ này, giả sử chúng ta có mô hình bốn biến:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i$$

và giả sử là

$$X_{4i} = \lambda_2 X_{2i} + \lambda_3 X_{3i}$$

với λ_2 và λ_3 là các hằng số không đồng thời bằng 0. Rõ ràng là, X_4 là một tổ hợp tuyến tính chính xác của X_2 và X_3, với $R^2_{4.23} = 1$, hệ số xác định trong hàm hồi qui của X_4 theo X_2 và X_3.

Bây giờ nhớ lại cộng thức (7.9.6) ở chương 7, chúng ta có thể viết

$$R^2_{4.23} = \frac{r^2_{42} + r^2_{43} - 2r_{42}r_{43}}{1 - r^2_{23}}$$

Nhưng vì $R^2_{4.23} = 1$ do cộng tuyến hoàn hảo, chúng ta có

$$1 = \frac{r^2_{42} + r^2_{43} - 2r_{42}r_{43}}{1 - r^2_{23}} \qquad (10.7.2)$$

Thật không khó để nhận ra là (10.7.2) thỏa khi $r_{42} = 0.5$, $r_{43} = 0.5$ và $r_{23} = -0.5$, đây là những giá trị không quá cao.

Vì vậy, trong mô hình liên quan đến nhiều hơn hai biến giải thích, hệ số tương quan bậc 0 hay hệ số tương quan đơn sẽ không cung cấp một chỉ dẫn đáng tin cậy về sự hiện diện của đa cộng tuyến. Dĩ nhiên, nếu chỉ có hai biến giải thích, các hệ số tương quan bậc 0 là đủ rồi.

3. **Kiểm tra các hệ số tương quan riêng phần.** Vì vấn đề vừa nêu chỉ dựa vào các hệ số tương quan bậc 0, Farrar và Glauber đề nghị là chúng ta nên quan tâm đến các hệ số tương quan riêng phần.[19] Vì vậy, trong hàm hồi qui của Y theo X_2, X_3 và X_4, một phát hiện là $R^2_{1.234}$ thì rất cao nhưng $r^2_{12.34}$, $r^2_{13.24}$ và $r^2_{14.23}$ thì tương đối thấp có thể ngụ ý là các biến X_2, X_3 và X_4 có tương quan lẫn nhau cao và ít nhất một trong những biến này là không cần thiết.
   Mặc dù một nghiên cứu về các hệ số tương quan có lẽ sẽ có ích nhưng không có gì bảo đảm là những hệ số này sẽ đem lại một chỉ dẫn đáng tin cậy về đa cộng tuyến, vì có thể ngẫu nhiên cả R^2 và mọi hệ số tương quan riêng phần đều đủ cao. Nhưng quan trọng hơn là, C. Robert Wichers đã chỉ ra [20] là kiểm định Farrar - Glauber về hệ số tương quan riêng phần

[18] Ibid., trang 439.
[19] D. E. Farrar và R. R. Glauber, " Multicollinearity in Regression Analysis: The Problem Revisited," (Đa cộng tuyến trong phân tích hồi qui: Vấn đề được xem xét lại), *Review of Econometrics and Statistics,* số 49, 1967, trang 92-107.
[20] "The Detection of Multicollinearity: A Comment", (Sự phát hiện đa cộng tuyến: Một lời bình luận), *Review of econometrics and Statistics,* số 57, 1975, trang 365-366.

không đủ hiệu quả trong việc so sánh một hệ số tương quan riêng phần cho trước với các kiểu đa cộng tuyến khác.

Kiểm định Farrar - Glauber cũng đã bị T.Krishma,[21] John O'Hagan và Brendan McCabe.[22] chỉ trích kịch liệt.

4. **Các hàm hồi qui phụ trợ**. Từ khi vấn đề đa cộng tuyến phát sinh vì một hay nhiều biến hồi qui độc lập là tổ hợp tuyến tính hoàn hảo hoặc gần như hoàn hảo của các biến hồi qui độc lập khác nào, một cách để tìm xem biến X nào có quan hệ với các biến X khác, là lập hàm hồi qui cho mỗi biến X_i theo các biến X còn lại và tính R^2 tương ứng, mà ta đặt là R^2_i; mỗi một hàm hồi qui trong những hàm hồi qui này gọi là **hàm hồi qui phụ trợ**, phụ cho hàm hồi qui chính của Y theo các biến X. Kế đó, mối liên hệ sau giữa F và R^2 đã được thiết lập trong (8.5.11), biến

$$R_i = \frac{R^2_{x1.\,x2x3...xk} / (k-2)}{(1 - R^2_{x1.\,x2x3...xk}) / (n - k + 1)} \qquad (10.7.3)$$

tuân theo phân phối F với độ tự do k - 2 và n - k + 1. Trong biểu thức (10.7.3) *n* đại diện cho cỡ mẫu, k đại diện cho số biến giải thích gồm cả số hạng tung độ gốc, và $R^2_{x1.\,x2x3...\,xk}$ là hệ số xác định trong hàm hồi qui của biến X_i theo các biến X còn lại. [23]

Nếu giá trị F tính được cao hơn giá trị F_i, điều đó có nghĩa là biến X_i cụ thể này cộng tuyến với các biến X khác; nếu giá trị F tính được không vượt quá giá trị tới hạn F_i, chúng ta nói rằng X_i không cộng tuyến với các biến X khác, trong trường hợp này chúng ta có thể vẫn duy trì biến đó trong mô hình. Nếu F_i có ý nghĩa thống kê, chúng ta sẽ vẫn phải giải quyết xem biến X_i cụ thể này nên bị bỏ khỏi mô hình hay không. Câu hỏi này sẽ đượcđề cập đến trong phần 10.8.

> Nhưng phương pháp này không phải là không có trở ngại, bởi vì...nếu vấn đề đa cộng tuyến chỉ liên quan đến một vài biến đến nỗi các hàm hồi qui phụ trợ không bị ảnh hưởng từ đa cộng tuyến mở rộng, các hệ số độ dốc ước lượng có thể cho thấy bản chất của sự phụ thuộc tuyến tính giữa các biến hồi qui độc lập. Không may thay, nếu có nhiều liên kết tuyến tính phức tạp, đường cong thực nghiệm này có lẽ không có nhiều giá trị vì sẽ khó xác định các quan hệ giữa các biến một cách tách biệt.[24]

Thay vì kiểm định thông thường mọi giá trị R^2 phụ, ta có thể sử dụng qui tắc kinh nghiệm của Klien, kinh nghiệm này cho là vấn đề đa cộng tuyến có lẽ là một vấn đề phức tạp chỉ khi R^2

---

[21] Multicollinearity in Regression Analysis", (Đa cộng tuyến trong phân tích hồi qui), Review of Econometrics anhd Statistics, số 57, 1975, trang 366-368.

[22] "Test for the Severity of Multicollinearity in Regression Analysis: A comment" (Kiểm định tính nghiêm trọng của đa cộng tuyến trong phân tích hồi qui), *Review of Econometrics and Statistics*, số 57, 1975, trang 368 - 370.

[23] Ví dụ, R^2_{x2} có thể có được bằng cách lập hàm hồi qui X_2 như sau: $X_{2i} = a_1 + a_3X_{3i} + \tilde{a}X_{4i} + ... + a_kX_{ki} + \hat{u}_i$.

[24] George G. Judge, R. Carter Hill, William E. Griffiths, Helmut Lutkepohl, và Tsoung-Chao Lee, *Introduction to the Theory and Practice of Econometrics,* (Nhập môn Lý thuyết và Thực hành môn Kinh tế lượng), John Wiley & Sons, New York, 1982, trang 621.

có được từ một hàm hồi qui phụ trợ có giá trị lớn hơn R^2 toàn diện, đó là, R^2 có từ hàm hồi qui của Y theo mọi biến hồi qui độc lập.[25] Dĩ nhiên, như mọi qui tắc kinh nghiệm khác, cần phải cân nhắc khi sử dụng kinh nghiệm này.

5. **Giá đặc trưng và chỉ số điều kiện.** Nếu bạn kiểm tra sản lượng SAS của hàm sản xuất của Cobb-Douglas cho trong phụ lục 7A.7, bạn sẽ thấy là SAS sử dụng giá trị đặc trưng và chỉ số điều kiện để chẩn đoán đa cộng tuyến. Chúng ta sẽ không thảo luận về giá trị đặc trưng ở đây, vì điều đó sẽ dẫn chúng ta vào đề tài về ma trận đại số, vượt ngoài phạm vi cuốn sách này. Tuy nhiên, từ những giá trị đặc trưng, chúng ta có thể có được cái gọi là **số điều kiện k** (condition number k), được định nghĩa là

$$k = \frac{\text{giá trị đặc trưng lớn nhất}}{\text{giá trị đặc trưng nhỏ nhất}}$$

và chỉ số điều kiện (condition index) (CI) được định nghĩa là

$$CI = \sqrt{\frac{\text{giá trị đặc trưng lớn nhất}}{\text{giá trị đặc trưng nhỏ nhất}}} = \sqrt{k}$$

kế đó chúng ta có qui tắc kinh nghiệm này. Nếu k nằm giữa 100 và 1000 thì có sự đa cộng tuyến từ trung bình đến cao và nếu giá trị này cao hơn 1000 thì có sự đa cộng tuyến rất cao. Hay nếu CI ($= \sqrt{k}$) giữa 10 và 30, có sự đa cộng tuyến từ trung bình đến cao và nếu giá trị này cao hơn 30 thì có sự đa cộng tuyến rất cao.

Đối với ví dụ minh họa, k = 3.0/0.00002422 hoặc bằng khoảng 123,864 và $CI = \sqrt{123864} \approx 352$; cả giá trị k và CI vì vậy dự đoán là có sự đa cộng tuyến rất cao. Dĩ nhiên, k và CI có thể tính được giữa đặc trưng lớn nhất và bất kỳ giá trị đặc trưng khác như được làm trong tài liệu. (Lưu ý: tài liệu này không tính toán một cách rõ ràng giá trị k, nhưng chỉ đơn giản tính giá trị bình phương của CI.) Nhân đây, lưu ý rằng một giá trị đặc trưng thấp (so sánh tương đối với giá trị đặc trưng lớn nhất) thường là một dấu hiệu xác định của các phụ thuộc gần như tuyến tính trong số liệu.

6. Một vài tác giả tin rằng chỉ số điều kiện là cách chẩn đoán đa cộng tuyến sẵn có tốt nhất. Những ý kiến này không được tiếp nhận rộng rãi. Đối với chúng ta, CI chỉ là một qui tắc kinh nghiệm, có lẽ phức tạp hơn một chút. Nhưng để cụ thể hơn, độc giả có thể xem thêm các tài liệu tham khảo.[26]

[25] Lawrence R. Klien, *An Introduction to Econometrics,* (Nhập môn kinh tế lượng), Prentice- Hall, Englewood Cliffs, N. J., 1962, trang 101.

7. **Dung sai** (Tolerance) **và nhân tố lạm phát - phương sai.** Đối với mô hình hồi qui đa biến [Y, tung độ gốc và (k - 1) biến hồi qui độc lập], như chúng ta đã thấy trong (7.5.6) phương sai của hệ số hồi qui riêng phần có thể được diễn tả

$$\text{var}(\hat{\beta}_j) = \frac{\sigma^2}{\sum x^2_j} \cdot \left(\frac{1}{1 - R^2_j}\right) \qquad (7.5.6)$$

$$= \frac{\sigma^2}{\sum x^2_j} \cdot \text{VIF}_j \qquad (10.7.4)$$

với β_j là hệ số hồi qui (riêng phần) của biến hồi qui độc lập X_j, R^2_j là giá trị R^2 trong hàm hồi qui (phụ trợ) của X_j theo (k - 2) biến hồi qui độc lập còn lại vàVIF_j là nhân tố lạm phát phương sai được giới thiệu lần đầu tiên trong phần 10.5. Khi R^2_j tăng dần đến 1, đó là, vì sự cộng tuyến của X_j với các biến hồi qui độc lập khác tăng, VIF cũng tăng và trong giới hạn VIF có thể trở thành vô hạn.

Vì vậy một số tác giả dùng VIF như là một dấu hiệu xác định của đa cộng tuyến: Giá trị VIF càng lớn thì biến X_j càng "phức tạp" hoặc càng cộng tuyến cao. Nhưng VIF cao đến như thế nào trước khi một biến hồi qui độc lập trở nên rắc rối? **Như một qui tắc kinh nghiệm**, nếu VIF của một biến vượt quá 10 (điều này xảy ra nếu R^2_j vượt quá 0.9), biến này được nói là cộng tuyến cao.[27]

Các tác giả khác sử dụng phép đo dung sai để phát hiện đa cộng tuyến. Được định nghĩa như sau

$$\text{TOL}_j = (1 - R^2_j) = (1/\text{VIF}_j) \qquad (10.7.5)$$

Rõ ràng là, $\text{TOL}_j = 1$ nếu X_j không tương quan với các biến hồi qui độc lập khác, trong khi đó $\text{TOL}_j = 0$ nếu X_j liên kết hoàn toàn với cá biến hồi qui độc lập khác.

VIF (hoặc dung sai) như một phép đo độ cộng tuyến không tránh khỏi được các nhà phê bình. Như (10.7.4) trình bày, $\text{var}(\hat{\beta}_j)$ phụ thuộc ba yếu tố: σ^2, $\sum x^2_j$, và VIF_j. Một giá trị VIF cao có thể được cân bằng bởi σ^2 thấp hoặc $\sum x^2_j$ cao. Nói cách khác, một giá trị VIF cao thì không phải là điều kiện cần và đủ để có phương sai và sai số chuẩn cao. Vì vậy, đa cộng tuyến cao, như được đo lường bằng giá trị VIF cao, có lẽ không phải là điều kiện cần để gây ra sai số chuẩn cao. Trong thảo luận này, thuật ngữ cao và thấp được sử dụng với nghĩa tương đối.

---

[26] Đặc biệt xem D. A. Belsley, E. Kuh, và R. E. Welsch, *Regression Diagonistics: Identifying Influential Data and Sources of Collinearity*, (Chẩn đoán hồi qui: Xác định ảnh hưởng của cộng tuyến đến số liệu và các nguồn số liệu), John Wiley & Sons, New York, 1980, chương 3. Tuy nhiên, cuốn sách này không dành cho người mới học.

[27] xem David G. Kleinbaum, Lawrence L. Kupper, và Keith E. Muller, *Applied Regression Analysis and Other Multivariate Methods,* (Phân tích hồi qui ứng dụng và các phương pháp đa biến khác), 2d. ed., PWS-Kent, Boston, Mass., 1988, trang 210.

Để kết luận phần thảo luận của chúng ta và việc phát hiện đa cộng tuyến, chúng ta nhấn mạnh là nhiều phương pháp khác nhau mà chúng ta đã thảo luận đều có bản chất "thả câu" ("fishinng expeditions,") vì vậy chúng ta không thể nói phương pháp nào sẽ tốt trong bất kỳ một trường hợp ứng dụng cụ thể nào. Đáng tiếc là, chúng ta không thể làm được gì nhiều, vì đa cộng tuyến thì rất riêng biệt đối với mỗi mẫu cho trước mà nhà nghiên cứu có lẽ không kiểm soát được hết, đặc biệt là nếu số liệu về bản chất là phi thực nghiệm - trường hợp mà nhà nghiên cứu thường gặp trong các ngành khoa học xã hội.

Một lần nữa, nhại lại của đa cộng tuyến, Goldberger trích ra một số cách phát hiện cỡ mẫu nhỏ chẳng hạn như xây dựng giá trị tới hạn của một cỡ mẫu, n*, như vậy nảy sinh vấn đề cỡ mẫu nhỏ chỉ khi nào cỡ mẫu thật, n, nhỏ hơn n*. Quan điểm việc nhại lại của Goldberger là nhấn mạnh rằng cỡ mẫu nhỏ và việc thiếu các sự biến thiên của các biến giải thích có thể gây ra nhiều vấn đề mà ít nhất cũng nghiêm trọng như các vấn đề liên quan đến đa cộng tuyến.

10.8 CÁC BIỆN PHÁP GIẢI QUYẾT

Có thể làm gì nếu vấn đề đa cộng tuyến trở nên nghiêm trọng? Như trong trường hợp phát hiện đa cộng tuyến, không còn lời hướng dẫn nào đáng tin cậy nữa vì đa cộng tuyến đặc biệt là một vấn đề về mẫu. Tuy nhiên, chúng ta có thể cố gắng tuân theo các qui tắc kinh nghiệm, việc thành công còn phụ thuộc vào mức độ nghiêm trọng của vấn đề cộng tuyến.

1. **Thông tin đầu tiên**. Giả sử chúng ta xem xét mô hình

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

với Y = tiêu dùng, X_2 = thu nhập và X_3 = sự giàu có. Như đã lưu ý trước đây, biến thu nhập và biến sự giàu có có khuynh hướng cộng tuyến cao. Nhưng giả sử đầu tiên chúng ta tin là $\beta_3 = 0.10\beta_2$; đó là, tỷ lệ thay đổi của tiêu dùng theo sự giàu có bằng 1/10 tỷ lệ thay đổi tương ứng theo thu nhập. Chúng ta có thể tạo hàm hồi qui sau

$$Y_i = \beta_1 + \beta_2 X_{2i} + 0.10\beta_2 X_{3i} + u_i = \beta_1 + \beta_2 X_i + u_i$$

với $X_i = X_{2i} + 0.1X_{3i}$. Một khi chúng ta có $\hat{\beta}_2$, chúng ta có thể ước lượng $\hat{\beta}_3$ từ mối quan hệ cơ bản giữa β_2 và β_3.

Bằng cách nào chúng ta có được thông tin đầu tiên? Thông tin này có thể từ các công việc thực tế trước đây trong đó đã xảy ra nhiều vấn đề cộng tuyến nhưng ít nghiêm trọng hơn hoặc từ các lý thuyết tương ứng trong lĩnh vực nghiên cứu. Ví dụ, trong hàm sản xuất của Cobb-Douglas (7.10.1), nếu chúng ta kỳ vọng sinh lợi không đổi theo qui mô, thì $(\beta_2 + \beta_3) = 1$ trong trường hợp mà chúng ta có thể sử dụng hàm hồi qui (8.7.13), lập hàm hồi qui của tỉ số sản lượng / lao động theo tỉ số vốn/lao động. Nếu có cộng tuyến giữa lao động và vốn, như các trường hợp

thông thường trong phần lớn số liệu mẫu, một sự biến đổi như vậy có thể làm giảm hoặc loại bỏ được vấn đề đa cộng tuyến. Nhưng có một khuyến cáo ở đây là về việc ấn định một ràng buộc tiên nghiệm như vậy, "... vì nói chung chúng ta sẽ muốn kiểm định một dự đoán tiên nghiệm của học thuyết kinh tế hơn là chỉ đơn giản đặt chúng trên những số liệu mà theo những số liệu này có thể chúng không đúng."[28] Tuy nhiên, từ phần 8.7, chúng ta biết cách kiểm định một cách rõ ràng sự hiệu lực của những ràng buộc như vậy.

2. Kết hợp số liệu chéo (cross-sectional) **và số liệu chuỗi thời gian.** Một biến thể của kỹ thuật thông tin tương lai hoặc kỹ thuật thông tin tiên nghiệm là tổ hợp của dữ liệu chéo (liên vùng) và dữ liệu chuỗi thời gian, được gọi là *góp chung số liệu* (pooling the data). Giả sử là chúng ta muốn nghiên cứu về nhu cầu của xe máy ở Hoa Kỳ và giả sử là chúng ta có số liệu chuỗi thời gian về số lượng xe được bán ra, giá trung bình của xe hơi và thu nhập của người tiêu dùng. Cũng giả sử là

$$\ln Y_t = \beta_1 + \beta_2 \ln P_t + \beta_3 \ln I_t + u_t$$

với Y = số xe hơi bán ra, P = giá trung bình, I = thu nhập, và t = thời gian. Mục tiêu của chúng ta là ước lượng độ co giãn của giá β_2 và độ co giãn của thu nhập β_3.

Trong số liệu chuỗi thời gian, các biến giá cả và thu nhập nói chung có khuynh hướng cộng tuyến cao. Vì vậy, nếu chúng ta sử dụng hàm hồi qui trước đây, chúng ta sẽ gặp phải vấn đề đa cộng tuyến thường gặp. Tobin đã đề nghị một cách tránh khỏi vấn đề này.[29] Ông ta nói rằng nếu chúng ta có số liệu chéo (ví dụ, số liệu từ danh sách khách hàng, hoặc từ các nghiên cứu về ngân sách được nhiều tổ chức tư nhân hoặc chính phủ thực hiện), chúng ta có thể có được ước lượng khá tin cậy của độ co giãn β_3 bởi vì trong tập số liệu ở cùng một thời điểm như vậy, giá cả không thay đổi quá nhiều. Hãy xem độ co giãn về giá ước lượng theo số liệu chéo là $\hat{\beta}_3$. Sử dụng giá trị ước lượng này, chúng ta có thể viết được hàm hồi qui chuỗi thời gian trước đây như sau

$$Y^*_t = \beta_1 + \beta_2 \ln P_t + u_t$$

với $Y^* = \ln Y - \hat{\beta}_3 \ln I$, đó là, Y^* đại diện cho giá trị của Y sau khi tách bỏ ảnh hưởng của thu nhập lên biến này. Bây giờ chúng ta có thể có một giá trị ước lượng của độ co giãn của giá cả β_2 từ hàm hồi qui trên.

Mặc dù đây là một kỹ thuật hấp dẫn, nhưng góp chung số liệu chuỗi thời gian và số liệu chéo về cách thức vừa đề nghị có thể tạo ra các vấn đề về diễn dịch, bởi vì chúng ta ngầm giả định rằng độ co giãn giá cả ước lượng theo số liệu chéo thì cũng giống như giá trị được ước lượng theo

[28] Mark B. Stewart and Kenneth F. Wallis, *Introduction Econometrics*, (Nhập môn kinh tế lượng), 2d, ed., John Wiley & Sons, A Halstesd Press Book, New York, 1981, trang 154.
[29] J. Tobin, "A Statistical Demand Function for Food in the USA," (Hàm cầu thống kê của thức ăn ở Hoa Kỳ) *journal of the Royal Statistical Society,* Ser. A, 1950, trang 113-141

phân tích chuỗi thời gian thuần túy.[30] Tuy nhiên, kỹ thuật này đã được sử dụng trong nhiều ứng dụng và rất đáng giá trong những trường hợp các ước lượng dữ liệu chéo không biến đổi nhiều giữa một phần dữ liệu này và một phần dữ liệu khác: Một ví dụ về kỹ thuật này được cung cấp trong bài tập 10.25.

3. Bỏ qua một hoặc nhiều biến và các thiên lệch đặc trưng. Khi đối diện với vấn đề đa cộng tuyến nghiêm trọng, một trong những việc "đơn giản" nhất có thể làm là bỏ bớt một trong những biến cộng tuyến. Vì vậy, trong ví dụ minh họa của chúng ta về tiêu dùng-thu nhập-sự giàu có, khi chúng ta bỏ đi biến sự giàu có, chúng ta có hàm hồi qui (10.6.4), cho thấy là, trong khi ở mô hình nguyên thủy, biến thu nhập không có ý nghĩa thống kê, bây giờ biến này có ý nghĩa "cao".

Nhưng khi bỏ một biến khỏi mô hình chúng ta có thể phạm phải một **thiên lệch đặc trưng** hoặc **sai số đặc trưng**. Thiên lệch đặc trưng xuất hiện từ những đặc trưng không đúng của mô hình sử dụng để phân tích, vì vậy, nếu học thuyết kinh tế cho rằng thu nhập và sự giàu có có thể đều có mặt trong mô hình giải thích cho việc chi tiêu cho tiêu dùng, việc bỏ qua biến sự giàu có sẽ tạo thành thiên lệch đặc trưng.

Mặc dù chúng ta sẽ thảo luận đề tài về thiên lệch đặc trưng trong chương 13, chúng ta đã lướt qua vấn đề này trong phần 7.7 ở đó chúng ta đã thấy là nếu mô hình đúng thì

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

nhưng chúng ta đã làm thích hợp mô hình một cách sai lầm

$$Y_i = b_1 + b_{12} X_{2i} + \hat{u}_i \qquad (7.7.1)$$

kế đó

$$E(b_{12}) = \beta_2 + \beta_3 b_{32} \qquad (7.7.4)$$

với b_{32} = hệ số độ dốc trong hàm hồi qui của X_3 theo X_2. Vì vậy, rõ ràng từ (7.7.4) là b_{12} sẽ là một ước lượng thiên lệch của β_2 miễn là b_{23} khác 0 (giả sử là β_3 khác 0; nếu không thì sẽ vô nghĩa nếu đưa X_3 vào mô hình nguyên thủy).[31] Dĩ nhiên, nếu $b_{32} = 0$, chúng ta không gặp phải vấn đề đa cộng tuyến. Cũng thấy rõ ràng từ (7.7.4) là nếu cả b_{32} và β_3 đều dương, $E(b_{12})$ sẽ lớn hơn β_2; vì vậy, về trung bình b_{12} sẽ là ước lượng quá cao của β_2, dẫn đến thiên lệch dương. Tương tự, nếu tích $b_{32}\beta_3$ âm, về trung bình b_{12} sẽ thấp hơn β_2, dẫn đến thiên lệch âm.

[30] Để thông qua phần thảo luận này và ứng dụng kỹ thuật góp chung số liệu, xem Edwin Kuh, C*apital Stock Growth: A Micro-Econometric Approach,* (Sự tăng trưởng của vốn cổ phần: Một phương pháp kinh tế vi lượng), North-Holland Publishing Company, Amsterdam, 1963, chương 5 và 6.

[31] Lưu ý là nếu b_{32} không tiến đến 0 khi cỡ mẫu tăng vô hạn, kế đó b_{12} sẽ không chỉ thiên lệch mà còn không nhất quán.

Từ thảo luận trên, rõ ràng là việc bỏ một biến khỏi mô hình để làm giảm bớt vấn đề đa cộng tuyến có thể sẽ dẫn đến thiên lệch đặc trưng. Vì vậy, phương pháp giải quyết có lẽ lại còn làm cho vấn đề xấu thêm trong một số trường hợp, bởi vì, trong khi đa cộng tuyến có thể cản trở việc ước lượng được chính xác các thông số của mô hình, thì việc bỏ qua một biến có lẽ làm cho chúng ta lạc hướng trầm trọng khi tìm đến giá trị thực của các thông số. Nhớ lại các hàm ước lượng OLS là BLUE mặc dù gần như cộng tuyến.

4. Biến đổi các biến. Giả sử là chúng ta có số liệu chuỗi thời gian về chi tiêu cho tiêu dùng, thu nhập và sự giàu có. Một lý do của sự đa cộng tuyến cao giữa thu nhập và sự giàu có trong số liệu này là do theo thời gian cả hai biến này đều có khuynh hướng dịch chuyển theo cùng một hướng. Một cách để giảm thiểu sự phụ thuộc này là làm như sau.

Nếu quan hệ

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t \qquad (10.8.1)$$

có giá trị ở thời điểm t, nó cũng phải có giá trị ở thời điểm t - 1 bởi vì gốc thời gian là chọn tùy ý theo bất kỳ cách nào. Vì vậy, chúng ta có

$$Y_{t-1} = \beta_1 + \beta_2 X_{2,t-1} + \beta_3 X_{3,t-1} + u_{t-1} \qquad (10.8.2)$$

Nếu lấy (10.8.1) trừ (10.8.2) ta có

$$Y_t - Y_{t-1} = \beta_2 (X_{2t} - X_{2,t-1}) + \beta_3 (X_{3t} - X_{3,t-1}) + \nu_t \qquad (10.8.3)$$

với $\nu_t = u_t - u_{t-1}$. Biểu thức (10.8.3) được gọi là **dạng hiệu số thứ nhất** (the first difference form) vì chúng ta sử dụng hàm hồi qui, không theo biến nguyên thủy mà theo hiệu số giữa các giá trị liên tục của các biến.

Mô hình hồi qui hiệu số thứ nhất thường làm giảm mức độ nghiêm trọng của đa cộng tuyến vì, mặc dù các mức độ của X_2 và X_3 có thể tương quan cao, nhưng không có lý do chính đáng nào để tin là các hiệu số giữa chúng sẽ tương quan cao.

Tuy nhiên, sự biến đổi hiệu số thứ nhất lại tạo thêm một số vấn đề. Số hạng sai số ν_t xuất hiện trong (10.8.3) có thể không thỏa một trong những giả định của mô hình hồi qui tuyến tính cổ điển, đó là, các nhiễu này không quan hệ với nhau theo chuỗi thời gian. Như chúng ta sẽ thấy trong chương 12, nếu số hạng nguyên thủy u_t độc lập hoặc không tương quan theo chuỗi, thì số hạng sai số ν_t có được ở trên sẽ tương quan theo chuỗi thời gian trong hầu hết mọi trường hợp. Một lần nữa phương pháp giải quyết có lẽ lại làm vấn đề xấu thêm! Hơn thế nữa, do thủ tục hiệu số nên sẽ mất bớt một giá trị quan sát và vì vậy độ tự do bị giảm đi một. Trong một mẫu nhỏ điều

này có thể là một vấn đề cần được để ý đến. Hơn nữa, thủ tục hiệu số thứ nhất có lẽ không thích hợp với số liệu chéo vì số liệu này không có một trật tự logic cho các quan sát.

5. Số liệu bổ sung hoặc số liệu mới. Vì vấn đề đa cộng tuyến là một đặc tính của mẫu, có thể là trong một mẫu khác các biến cộng tuyến có lẽ sẽ không nghiêm trọng như trong mẫu đầu tiên. Thỉnh thoảng chỉ đơn giản gia tăng cỡ mẫu (nếu có thể) cũng có thể làm giảm bớt vấn đề cộng tuyến. Ví dụ, trong mô hình ba biến chúng ta đã thấy là

$$\text{var}\,(\hat{\beta}_2) = \frac{\sigma^2}{\sum x^2_{2i}\,(1 - r^2_{23})}$$

Bây giờ khi cỡ mẫu tăng, $\sum x^2_{2i}$ nói chung sẽ tăng. (Tại sao?) Vì vậy, đối với bất kỳ r_{23} nào cho trước, phương sai của $\hat{\beta}_2$ sẽ giảm, do đó kéo theo sai số chuẩn giảm; điều này giúp chúng ta ước lượng β_2 chính xác hơn.

Để minh họa, xem hàm hồi qui sau của chi tiêu cho tiêu dùng Y theo thu nhập X_2 và sự giàu có X_3 dựa trên 10 quan sát:[32]

$$\hat{Y}_i = 24.337 + 0.8716X_{2i} - 0.0349X_{3i} \qquad (10.8.4)$$

$$t = (3.875) \quad (2.7726) \quad (-1.1595) \quad R^2 = 0.9682$$

Hệ số của biến giàu có trong hàm hồi qui này không chỉ có dấu sai mà còn không có ý nghĩa thống kê ở mức ý nghĩa 5%. Nhưng khi cỡ mẫu tăng lên 40 lần quan sát (vấn đề cỡ mẫu nhỏ?), ta có các kết quả sau

$$\hat{Y}_i = 2.0907 + 0.7299X_{2i} + 0.0605X_{3i} \qquad (10.8.5)$$

$$t = (0.8713) \quad (6.0014) \quad (2.0014) \quad R^2 = 0.9672$$

Bây giờ hệ số biến giàu có không chỉ có dấu đúng mà còn có ý nghĩa thống kê ở mức ý nghĩa 5%.

Có thêm số liệu bổ sung hoặc số liệu "tốt hơn" không phải luôn luôn dễ dàng, vì như Judge và những người khác đã lưu ý:

> Không may thay, các nhà kinh tế học ít khi có thể có được số liệu bổ sung mà không phải chịu những khoảng chi phí quá lớn, với ít lựa chọn cho các giá trị của các biến giải thích mà họ mong muốn. Thêm vào đó, khi bổ sung những biến mới trong nhiều trường hợp không thể kiểm soát

[32] Tôi biết ơn Albert Zucker vì đã cung cấp cho tôi các kết quả đưa ra trong những hàm hồi qui sau.

được, chúng ta phải biết là bổ sung thêm các quan sát có được từ một quá trình khác với các quan sát kết hợp với tập số liệu ban đầu; đó là, chúng ta phải chắc chắn tăng cấu trúc kinh tế kết hợp với những quan sát mới phải giống như cấu trúc ban đầu.[33]

6. Giảm cộng tuyến trong các hàm hồi qui đa thức. Trong phần 7.11 chúng ta đã thảo luận về mô hình hồi qui đa thức. Một thuộc tính đặc biệt của các mô hình này là các biến giải thích xuất hiện với nhiều số mũ khác nhau. Vì vậy, hàm tổng chi phí bậc ba là hàm hồi qui của tổng chi phí theo sản lượng, (sản lượng)^2, và (sản lượng)^3, như trong (7.11.4), các số hạng sản lượng khác nhau sẽ tương quan với nhau, làm cho khó ước lượng chính xác các hệ số độ dốc khác nhau.[34] Trong thực tế mặc dù người ta tìm thấy là nếu (các) biến giải thích được diễn tả dưới dạng độ lệch (đó là, độ lệch so với giá trị trung bình), đa cộng tuyến thật sự giảm bớt. Nhưng ngay cả sau đó vấn đề này có thể vẫn còn tồn tại, [35] trong trường hợp đó chúng ta có thể muốn xem xét các kỹ thuật như **các đa thức trực giao.**[36]

7. Các phương pháp khác giải quyết vấn đề đa cộng tuyến. Các kỹ thuật thống kê đa biến như **phân tích nhân tố** (factor analysis) và **các thành tố cơ bản** (principal components) hoặc các kỹ thuật như **hồi qui dạng sóng** (ridge regression) thường được sử dụng để "giải quyết" vấn đề đa cộng tuyến. Nhưng đáng tiếc là những kỹ thuật này ngoài phạm vi của cuốn sách, vì chúng ta không thể thảo luận những kỹ thuật này một cách hoàn chỉnh mà không sử dụng đến ma trận đại số.[37]

10.9 CÓ NHẤT THIẾT ĐA CỘNG TUYẾN LÀ XẤU KHÔNG? CÓ LẼ KHÔNG NẾU NHƯ MỤC TIÊU CHỈ ĐƠN THUẦN LÀ TIÊN ĐOÁN

Người ta đã nói là nếu mục tiêu chính của phân tích hồi qui là tiên đoán hoặc dự báo, thì đa cộng tuyến không phải là một vấn đề nghiêm trọng bởi vì giá trị R^2 càng cao thì tiên đoán càng chính tốt.[38] nhưng điều này có thể là "...miễn là các giá trị của các biến giải thích mà đối với các biến này người ta mong rằng các dự báo phải tuân theo sự phụ thuộc gần như tuyến tính chính xác

[33] Judge et al., op. cit., trang 625. Xem thêm phần 10.9

[34] Như đã lưu ý, tương quan giữa X, X^2 và X^3 là phi tuyến, nghiêm khắc mà nói thì, các hàm hồi qui đa thức không vi phạm các giả định phi đa cộng tuyến của mô hình cổ điển.

[35] Xem R. A. Bradley và S. S. Srivastava, "Correlation and Polynomial Regression," (Tương quan và Các hàm hồi qui đa thức), *American Statistician,* số 33, 1979, trang 11-14.

[36] Xem Norman Draper và Harry Smith, *Applied Regression Analysis,* (Phân tích hồi qui ứng dụng), 2d ed., John Wiley & Sons, New York, 1981, trang 266-274.

[37] Có thể đọc thêm về những kỹ thuật này trong ứng dụng ở Samprit Chaterjee và Bertram Price, R*egression Analysis by Example*, (Phân tích hồi qui bằng ví dụ), John Wiley & Sons, New York, 1977, chương 7, 8. Xem thêm H. D. Vinod, "A Survey of Ridge Regression and Related Techniques for Improvements over Ordinary Least Square", *Review of Economics and Statistics,* số 60, tháng 2, 1963, trang 121-131.

[38] Xem thêm R. C. Geary, "Some Results about Relation between Stochastic Variables: A Discussion Document," (Một số kết quả về mối quan hệ giữa các biến ngẫu nhiên: Một tài liệu thảo luận), *Review of International Statistical Institute,* so61 31, 1963, trang 163-181.

như ma trận [dữ liệu] X thiết kế ban đầu."[39] vì vậy, nếu trong một hàm hồi qui ước lượng có $X_2 \approx 2X_3$, thì trong một mẫu ở tương lai được dùng để dự báo Y, X_2 cũng sẽ gần bằng $2X_3$, một điều kiện thật khó gặp trong thực tế (xem ghi chú 33), trong trường hợp này dự đoán sẽ gia tăng sự không chắc chắn.[40] Hơn nữa, nếu mục tiêu của phân tích này không chỉ là dự báo mà còn là ước lượng tin cậy của các thông số, đa cộng tuyến nghiêm trọng có thể sẽ là một vấn đề bởi vì chúng ta đã thấy là đa cộng tuyến nghiêm trọng dẫn đến sai số của các hàm ước lượng sẽ lớn.

Tuy nhiên có một tình huống, đa cộng tuyến có lẽ không gây ra vấn đề nghiêm trọng. Đó là trường hợp khi R^2 cao và hệ số hồi qui có ý nghĩa một cách riêng biệt như được thấy qua các giá trị t cao hơn. Tuy nhiên, các chẩn đoán đa cộng tuyến, chỉ số điều kiện, chỉ ra là có sự cộng tuyến nghiêm trọng trong số liệu. Khi nào một tình huống như vậy xuất hiện? Như Johnston lưu ý:

> Trường hợp này xảy ra nếu các hệ số riêng phần cao hơn giá trị thực, vì thế không xuất hiện các tác động mặc dù sai số chuẩn gia tăng và/hoặc bởi vì bản thân giá trị thực quá lớn đến nỗi ngay cả một ước lượng theo chiều đi xuống cũng vẫn có vẻ như có ý nghĩa.[41]

10.10 TÓM TẮT VÀ KẾT LUẬN

1. Một trong những giả định của mô hình hồi qui tuyến tính cổ điển là không có vấn đề đa cộng tuyến giữa các biến giải thích X. Nói rộng ra là, vấn đề đa cộng tuyến đề cập đến tình huống trong đó tồn tại một mối quan hệ tuyến tính hoàn hảo hoặc gần như hoàn hảo giữa các biến X.

2. Các hệ quả của đa cộng tuyến là: Nếu tồn tại cộng tuyến hoàn hảo giữa các biến X, thì hệ số hồi qui của chúng là không xác định và các sai số chuẩn của chúng là vô hạn. Nếu cộng tuyến cao nhưng không hoàn hảo thì việc ước lượng của các hệ số hồi qui là có thể thực hiện được nhưng sai số chuẩn của chúng có khuynh hướng rất lớn. Kết quả là, các giá trị tổng thể của các hệ số không thể được ước lượng một cách chính xác. Tuy nhiên, nếu mục tiêu là ước lượng tổ hợp tuyến tính của các hệ số này, *các hàm ước lượng,* thì việc này có thể thực hiện được ngay cả với sự hiện diện của đa cộng tuyến hoàn hảo.

3. Mặc dù không có phương pháp chắc chắn nào để phát hiện cộng tuyến, nhưng có một số chỉ dẫn như sau:

(a) Dấu hiệu rõ nhất của đa cộng tuyến là khi R^2 rất cao nhưng không có hệ số hồi qui nào có ý nghĩa thống kê dựa trên kiểm định qui ước t. Trường hợp này dĩ nhiên là cực đoan.

---

[39] Judge et al, op cit., trang 619. Bạn cũng có thể tìm thấy ở trang này bằng chứng, mặc dù cộng tuyến, là tại sao chúng ta có thể có các giá trị dự báo trung bình tốt hơn nếu cấu trúc cộng tuyến hiện tại vẫn tiếp tục ở các mẫu trong tương lai

[40] Để thảo luận thật tốt, xem thêm E. Malinvaud, *Statistical methods of Econometrics,* 2d ed., North Holland Publishing Company, Amsterdam, 1970, trang 220-221.

[41] J. Johnston, *Econometric Methods,* (Các phương pháp kinh tế lượng), 3d ed., McGraw Hill, New York, 1984, trang 249

(b) Trong các mô hình chỉ liên quan đến hai biến giải thích, một phát hiện tốt về cộng tuyến có thể có được bằng cách kiểm tra hệ số tương quan bậc 0 hay hệ số tương quan đơn giữa hai biến. Nếu hệ số này cao, thì thông thường đó chính là do đa cộng tuyến.

(c) Tuy nhiên, hệ số tương quan bậc 0 có thể dẫn đến sai lầm trong mô hình có nhiều hơn hai biến giải thích bởi vì có thể có hệ số tương quan bậc 0 thấp nhưng vẫn có đa cộng tuyến cao. Trong những trường hợp như thế, có lẽ chúng ta cần phải kiểm tra các hệ số tương quan riêng phần.

(d) Nếu R^2 cao nhưng hệ số tương quan riêng phần thấp, thì có thể có đa cộng tuyến. Ở đây một hoặc nhiều biến có thể là không cần thiết. Nhưng nếu R^2 cao và các hệ số tương quan riêng phần cũng cao, thì có lẻ không thể phát hiện được đa cộng tuyến ngay. Cũng như C. Robert, Krishna Kuma, John O'Hagan và Brendan McCabe đã nêu, có một số vấn đề thống kê với kiểm định các hệ số tương quan riêng phần do Farrar và Glauber đề nghị.

(e) Vì vậy, chúng ta có thể lập hàm hồi qui mỗi biến X_i theo các biến X còn lại trong mô hình và tìm ra các hệ số tương ứng của R^2_i. Một giá trị R^2_i cao có thể cho là X_i tương quan chặt với các biến X còn lại. Do đó, chúng ta có thể bỏ biến nay khỏi mô hình, miễn là nó không gây ra các thiên lệch đặc trưng nghiêm trọng.

4. Phát hiện ra đa cộng tuyến chỉ là một nửa nhiệm vụ. Nửa còn lại liên quan đến việc giải quyết vấn đề này bằng cách nào. Một lần nữa lại không có phương pháp nào chắc chắn, chỉ có một ít qui tắc kinh nghiệm. Một số qui tắc kinh nghiệm được nêu sau: (1) sử dụng thông tin tiên nghiệm hay thông tin ngoại lai, (2) kết hợp số liệu chéo và số liệu chuỗi thời gian, (3) bỏ qua biến cộng tuyến cao, (4) biến đổi số liệu, và (5) thêm số liệu bổ sung hoặc số liệu mới. Dĩ nhiên, qui tắc kinh nghiệm nào trong những qui tắc trên được áp dụng sẽ phụ thuộc vào bản chất của số liệu và mức độ nghiêm trọng của vấn đề cộng tuyến.

5. Chúng ta đã lưu ý đến vai trò của đa cộng tuyến trong dự báo và chỉ ra là trừ phi cấu trúc cộng tuyến vẫn tiếp tục trong mẫu tương lai, thật là nguy hiểm khi sử dụng hàm hồi qui ước lượng, đã bị tác hại của đa cộng tuyến, cho mục đích dự báo.

6. Mặc dù đa cộng tuyến đã nhận được sự quan tâm rộng rãi (có người cho rằng là quá mức) trong các tài liệu, một vấn đề không kém quan trọng mà chúng ta gặp phải trong nghiên cứu lý thuyết là vấn đề cỡ mẫu nhỏ, sự nhỏ của cỡ mẫu. Theo Goldberger, "Khi một bài báo nghiên cứu phàn nàn về đa cộng tuyến, đọc giả phải xem liệu những lời phàn nàn này có còn thuyết phục nếu "vấn đề cỡ mẫu nhỏ" được thay thế cho "vấn đề đa cộng tuyến".[42] Ông ta đề nghị là người đọc phải quyết định số lần quan sát *n* nhỏ đến cỡ nào trước khi quyết định là họ có vấn đề về cỡ mẫu nhỏ, như khi họ quyết định giá trị R^2 cao cỡ nào trong một hàm hồi qui phụ trợ trước khi nói rằng vấn đề cộng tuyến là nghiêm trọng.

[42] Goldberger, op cit., trang 250

BÀI TẬP

Câu hỏi

10.1. Trong mô hình hồi qui tuyến tính k biến có k biểu thức thông thường để ước lượng k giá trị chưa biết. Những biểu thức thông thường này được cho trong (9.8.3). Giả sử là X_k là tổ hợp tuyến tính của các biến X còn lại. Bằng cách nào bạn cho thấy là trong trường hợp này không thể ước lượng k hệ số hồi qui?

10.2. Xét một tập hợp các số liệu lý thuyết ở phần sau. Giả sử bạn muốn áp dụng mô hình sau cho số liệu đã cho

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

Y	X_2	X_3
- 10	1	1
- 8	2	3
- 6	3	5
- 4	4	7
- 2	5	9
0	6	11
2	7	13
4	8	15
6	9	17
8	10	19
10	11	21

(a) Bạn có thể ước lượng ba thông số chưa biết hay không? Tại sao có hoặc tại sao không?
(b) Nếu không, hàm tuyến tính nào của các thông số này, hàm ước lượng, bạn có thể ước lượng được? Trình bày những tính toán cần thiết

10.3. Nhớ lại chương 8, phần 5, ở đó chúng ta đã xét đến đóng góp biên tế hoặc gia tăng của một biến giải thích. Ví dụ thảo luận ở đó liên quan đến hàm hồi qui của chi tiêu cho tiêu dùng cá nhân Y theo thu nhập khả dụng của cá nhân X_2, và xu hướng X_3. Khi chúng ta đưa biến X_2 vào mô hình trước và sau đó đưa biến X_3 vào, ta có bảng 8.7. Nhưng giả sử là chúng ta đưa X_3 vào trước và sau đó đến X_2. Bảng ANOVA tương ứng với thay đổi này như sau:

Bảng ANOVA khi đưa X_3 vào trước

Nguồn thay đổi	SS		df	MSS
ESS do chỉ X_3	Q_1 =	64,536.2529	1	64,536.2529
ESS do thêm X_2	Q_2 =	1,428.8471	1	1,428.8471
ESS do X_2 và X_3	Q_3 =	65,965.1000	2	32,982.5500
Do các biến còn lại	Q_4 =	77.1693	12	6.4310
Tổng	Q_5 =	66,042.2693		

Mặc dù ESS do X_2 và X_3 hợp lại thì giống nhau trong các bảng, nhưng vị trí giữa hai biến thì khác. Trong bảng 8.7, khi đưa X_2 vào trước, đóng góp của biến này vào ESS là 65,898.2353, nhưng khi đưa X_2 vào như bảng trên, đóng góp của biến này chỉ có 1,428.8471. Điều này cũng đúng với X_3. Bạn giải thích hiện tượng này như thế nào?

10.4. Nếu quan hệ $\lambda_1 X_{1i} + \lambda_2 X_{2i} + \lambda_3 X_{3i} = 0$ vẫn đúng với mọi giá trị cũa λ_1, λ_2, và λ_3, hãy ước lượng $r_{12.3}$, $r_{13.2}$ và $r_{23.1}$. Cũng vậy, tìm $R^2_{1.23}$, $R^2_{2.13}$, và $R^2_{3.12}$. Mức độ đa cộng tuyến trong trường hợp này là gì? *Lưu ý*: $R^2_{1.23}$ là hệ số xác định trong hàm hồi qui của biến Y theo X_2 và X_3 . Các giá trị R^2 khác cũng được giải thích tương tự.

10.5 Xét mô hình sau:

$$Y_t = \beta_1 + \beta_2 X_t + \beta_3 X_{t-1} + \beta_4 X_{t-2} + \beta_5 X_{t-3} + \beta_6 X_{t-4} + u_t$$

với Y = tiêu dùng, X = thu nhập, và t = thời gian. Mô hình trên đòi hỏi là chi tiêu cho tiêu dùng ở thời điểm t là một hàm không chỉ của thu nhập và thời gian mà còn của thu nhập của những thời kỳ trước. Vì vậy, chi tiêu cho tiêu dùng trong quí 1 năm 1976 là một hàm của thu nhập trong quí đó và 4 quí của năm 1975. Mô hình như vậy gọi là **mô hình trễ pha phân phối**, (distributed lag models), và chúng ta sẽ thảo luận mô hình này ở một chương sau.

(a) Bạn có nghĩ là có vấn đề đa cộng tuyến trong mô hình như vậy hay không và tại sao?

(b) Nếu bạn nghĩ là có cộng tuyến, bạn sẽ giải quyết như thế nào?

10.6. Xem ví dụ minh họa của phần 10.6. Bạn sẽ điều hòa sự khác biệt trong thiên hướng gia tăng tiêu dùng giữa (10.6.1) và (10.6.4) như thế nào?

10.7. Trong số liệu liên quan đến chuỗi thời gian kinh tế như GNP, nguồn cung tiền tệ, thu nhập, thất nghiệp, vv... người ta thường nghi ngờ có sự hiện diện của đa cộng tuyến. Tại sao?

10.8. Giả sử mô hình

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

với r_{23}, hệ số tương quan giữa X_2 và X_3 , là 0. Vì vậy, một số người đề nghị là bạn sử dụng hàm hồi qui sau:

$$Y_i = \alpha_1 + \alpha_2 X_{2i} + u_{1i}$$
$$Y_i = \gamma_1 + \gamma_3 X_{3i} + u_{2I}$$

(a) Liệu có $\hat{\alpha}_2 = \hat{\beta}_2$ và $\hat{\gamma}_3 = \hat{\beta}_3$ hay không? Tại sao?

(b) Liệu $\hat{\beta}_1$ có bằng $\hat{\alpha}_1$ hoặc $\hat{\gamma}_1$ hoặc bằng một số tổ hợp của chúng hay không?

(c) Liệu có var $(\hat{\alpha}_2)$ = var $(\hat{\beta}_2)$ và var $(\hat{\gamma}_3)$ = var $(\hat{\beta}_3)$ hay không?

10.9. Đề cập đến ví dụ minh họa của chương 7, ở đó chúng ta sử dụng hàm sản xuất của Cobb-Douglas cho khu vực nông nghiệp của Đài Loan. Các kết quả của hàm hồi qui này cho ở (7.10.4) cho thấy là cả hệ số lao động và hệ số vốn đều có ý nghĩa thống kê riêng biệt.

(a) Hãy tìm xem các biến lao động và vốn có tương quan cao hay không?

(b) Nếu câu (a) bạn trả lời là có, bạn có thể bỏ biến lao động khỏi mô hình và lập hàm hồi qui của biến sản lượng chỉ theo nhập lượng vốn hay không?

(c) Nếu làm như vậy, bạn sẽ phạm phải thiên lệch đặc trưng loại gì? Hãy xác định bản chất của thiên lệch này.

10.10. Đề cập đến ví dụ 7.4. Với vấn đề này, ma trận tương quan cho như sau:

	X_i	X_i^2	X^3_i
X_i	1	0.9742	0.9284
X^2_i		1.0	0.9872
X^3_i			1.0

(a) "Vì hệ số tương quan bậc 0 là rất cao, nên có có lẽ có đa cộng tuyến nghiêm trọng." Hãy bình luận câu nhận xét trên.

(b) Bạn có thể bỏ biến $\mathbf{X_i^2}$ và $\mathbf{X^3_i}$ khỏi mô hình được hay không?

(c) Nếu bạn bỏ các biến trên, việc gì sẽ xảy ra với giá trị của hệ số của biến X_i ?

10.11. Hồi qui theo từng bước. Để quyết định tập hợp tốt nhất của các biến giải thích cho một mô hình hồi qui, những nhà nghiên cứu thường dùng phương pháp hồi qui dạng sóng. Trong phương pháp này chúng ta có thể tiến hành bằng cách đưa từng biến X vào (**hồi qui theo từng bước về phía trước)** hoặc bằng cách đưa toàn bộ các biến X vào một hàm hồi qui đa biến rồi đẩy từng biến một ra ngoài (**hồi qui theo từng bước về phía sau)**. Quyết định thêm hoặc bỏ một biến thường dựa trên cơ sở phần đóng góp của biến đó vào ESS, như được đánh giá bằng kiểm định F. Với những gì bạn đã biết về đa cộng tuyến, bạn có đề nghị một thủ tục nào khác hay không? Tại sao hoặc tại sao không?*

10.12. Xác định *và nêu lý do,* các câu sau đây là đúng, sai hoặc không chắc chắn:

(a) Mặc dù đa cộng tuyến hoàn hảo, hàm ước lượng OLS là BLUE

(b) Trong trường hợp đa cộng tuyến cao, không thể đánh giá mức độ ý nghĩa riêng của một hoặc nhiều hệ số hồi qui riêng phần

(c) Nếu một hàm hồi qui phụ trợ cho thấy là một R^2_i cụ thể có giá trị cao, thì có bằng chứng xác đáng về tính cộng tuyến cao hay không.

(d) Các hệ số tương quan từng đôi cao không có nghĩa là có đa cộng tuyến cao

(e) Đa cộng tuyến thì vô hại nếu mục tiêu của phân tích chỉ là dự báo

(f) Nếu giữ các yếu tố khác không đổi, VIF càng cao thì các giá trị phương sai của hàm OLS càng cao

(g) Dung sai (TOL) là một công cụ đo lường đa cộng tuyến tốt hơn VIF

(h) Bạn sẽ không có được giá trị R^2 cao trong hàm hồi qui đa biến nếu mọi hệ số độ dốc riêng phần đều không có ý nghĩa thống kê một cách riêng biệt theo kiểm định t

(i) Trong hàm hồi qui của Y theo X_2 và X_3 , giả sử có sự thay đổi nhỏ trong giá trị của X_3. Điều này sẽ làm tăng var ($\hat{\beta}_3$). Ở trạng thái cực đoan, nếu mọi X_3 đều giống nhau thì var ($\hat{\beta}_3$) là vô hạn

10.13. (a) Chứng tỏ là nếu $r_{1\,i} = 0$ với i = 2,3,...,k thì $R_{1.\ 23...k} = 0$

(b) Phát hiện này có gì quan trọng đối với hàm hồi qui của biến X_1 (= Y) theo X_2, X_3, ...X_k?

10.14. GIả sử mọi hệ số tương quan bậc 0 của X_1 (=Y), X_2,..., X_k đều bằng r.

(a) $R^2_{1.\ 23...k}$ bằng bao nhiêu?

(b) Các giá trị của các hệ số tương quan bậc 1 là gì?

** **10.15.** Trong ma trận ký hiệu chúng ta đã thấy trong chương 9

$$\hat{\boldsymbol{\beta}} = \mathbf{(X'X)^{-1}X'y}$$

(a) Điều gì xảy ra với $\hat{\boldsymbol{\beta}}$ khi có cộng tuyến hoàn hảo giữa các biến X?

(b) Bằng cách nào bạn biết được có tồn tại cộng tuyến hoàn hảo?

** **10.16.** Sử dụng ma trận ký hiệu chúng ta có được ở (9.3.13)

$$\text{var-cov}\,(\hat{\boldsymbol{\beta}}) = \sigma^2\,\mathbf{(X'X)^{-1}}$$

* Xem các lý do của bạn có đúng với các lý do của Arthur S. Goldberger và D> b. Jochems, "Lưu ý về tố thiểu từng bước (Stepwise Least-Square)," *Journal of the American Statistical Association,* số 56, tháng 3, 1961, trang 105-110.

** Lựa chọn

Điều gì xảy ra với giá trị ma trận var-cov (a) khi có đa cộng tuyến hoàn hảo và (b) khi cộng tuyến cao nhưng không hoàn hảo

** **10.17.** Xét **ma trận tương quan** sau:

$$\mathbf{R} = \begin{array}{c|cccc} & X_2 & X_3 & \dots & X_k \\ X_2 & 1 & r_{23} & \dots & r_{2k} \\ X_3 & r_{32} & 1 & \dots & r_{3k} \\ \dots & \dots & \dots & \dots & \dots \\ X_k & r_{k2} & r_{k3} & \dots & 1 \end{array}$$

Bằng cách nào bạn tìm được từ ma trận tương quan này (a) có cộng tuyến hoàn hảo hay không, (b) có cộng tuyến chưa hoàn hảo hay không, và (c) các biến X không tương quan.

Gợi ý: Bạn có thể dùng $|\mathbf{R}|$ để trả lời các câu hỏi này, với $|\mathbf{R}|$ là định thức của ma trận **R**.

** **10.18. Các biến giải thích trực giao**. Giả sử trong mô hình

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

X_2 đến X_k đều không tương quan. Những biến như vậy gọi là **biến trực giao**. Nếu là trường hợp này thì:

(a) Cấu trúc của ma trận ($\mathbf{X}'\mathbf{X}$) sẽ là gì?

(b) Bạn có được biểu thức $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ bằng cách nào?

(c) Bản chất của ma trận var-cov của $\hat{\boldsymbol{\beta}}$ là gì?

(d) Giả sử là bạn đang tiến hành hồi qui và sau đó bạn muốn đưa một biến trực giao khác, biến X_{k+1}, vào mô hình. Bạn có phải tính lại tất cả mọi hệ số $\hat{\beta}_2$ và $\hat{\beta}_k$ trước đây hay không? Tại sao có và tại sao không?

10.19. Xét mô hình sau:

$$GNP_t = \beta_1 + \beta_2 M_t + \beta_3 M_{t-1} + \beta_4 (M_t - M_{t-1}) + u_t$$

với GNP_t = GNP vào thời điểm t, M_t = nguồn cung tiền tệ ở thời điểm t, M_{t-1} = nguồn cung tiền tệ tại thời điểm (t - 1) và $(M_t - M_{t-1})$ = thay đổi về nguồn cung tiền tệ giữa thời điểm t và thời điểm (t - 1). Mô hình này đòi hỏi là mức GNP ở thời điểm t là một hàm của nguồn cung tiền tệ ở thời điểm t và (t - 1) cũng như sự thay đổi nguồn cung tiền tệ giữa các thời kỳ này.

(a) Giả sử bạn có số liệu để ước lượng mô hình trên, bạn có thể ước lượng được mọi hệ số của mô hình này hay không? Tại sao có và tại sao không?

(b) Nếu không, các hệ số nào có thể ước lượng được?

(c) Giả sử là số hạng $\beta_3 M_{t-1}$ không có mặt trong mô hình này. Câu trả lời của bạn có giống câu (a) không?

(d) Lập lại câu (c), với giả định là số hạng $\beta_2 M_t$ không có mặt trong mô hình.

10.20. Chứng tỏ là (7.4.7) và (7.4.8) cũng có thể được diễn tả như sau

$$\hat{\beta}_2 = \frac{\left(\sum y_i x_{2i}\right)\left(\sum x^2_{3i}\right) - \left(\sum y_i x_{3i}\right)\left(\sum x_{2i} x_{3i}\right)}{\left(\sum x^2_{2i}\right)\left(\sum x^2_{3i}\right)(1 - r^2_{23})}$$

$$\hat{\beta}_3 = \frac{\left(\sum y_i x_{3i}\right)\left(\sum x^2_{2i}\right) - \left(\sum y_i x_{2i}\right)\left(\sum x_{2i} x_{3i}\right)}{\left(\sum x^2_{2i}\right)\left(\sum x^2_{3i}\right)(1 - r^2_{23})}$$

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

10.21. Sử dụng (7.4.12) và (7.4.15), chứng tỏ là khi có cộng tuyến hoàn hảo thì các phương sai của $\hat{\beta}_2$ và $\hat{\beta}_3$ là vô hạn.

10.22. Kiểm chứng lại phát biểu: các sai số chuẩn của tổng các hệ số độ dốc ước lượng từ (10.5.4) và (10.5.5) theo thứ tự là 0.1992 và 0.1825. (Xem phần 10.5).

10.23. Với mô hình hồi qui k biến (9.1.1) có thể thấy là phương sai của hệ số hồi qui riêng phần thứ k (k = 2, 3, ..., k) có thể biểu diễn như sau*

$$\text{var}(\hat{\beta}_k) = \frac{1}{n-k}\frac{\sigma^2_y}{\sigma^2_k}\left(\frac{1 - R^2}{1 - R^2_k}\right)$$

với σ^2_y = phương sai của Y, σ^2_k = phương sai của biến giải thích thứ k, R^2_k = R^2 từ hàm hồi qui của X_k theo các biến X còn lại, và R^2 = hệ số xác định từ hàm hồi qui đa biến (9.1.1), đó là, hàm hồi qui của Y theo các biến X còn lại.

(a) Tất cả vẫn giữ nguyên, nếu σ^2_k tăng, chuyện gì sẽ xảy ra với var $(\hat{\beta}_k)$? Có những liên quan gì đến vấn đề đa cộng tuyến?

(b) Chuyện gì xảy ra với công thức trên khi cộng tuyến hoàn hảo?

(c) Phát biểu sau là đúng hay sai: "Phương sai của $\hat{\beta}_k$ giảm khi R^2 tăng, vì vậy ảnh hưởng của R^2_k cao có thể được bù lại bằng R^2 cao."

10.24. Căn cứ vào số liệu hàng năm của khu vực sản xuất của Hoa Kỳ trong thời gian 1899-1922, Dougherty có được kết quả hồi qui sau:*

$$\widehat{\log Y} = 2.81 - 0.53 \log K + 0.91 \log L + 0.047t \quad (1)$$
$$\text{se} = (1.38) \quad (0.34) \quad (0.14) \quad (0.021) \quad R^2 = 0.97, \quad F = 189.8$$

với Y = chỉ số của sản lượng thật, K = chỉ số của nhập lượng vốn thực, L = chỉ số nhập lượng của lao động thực, t = thời gian hoặc xu hướng.

Sử dụng cùng số liệu, ông ta cũng đã có được hàm hồi qui sau:

$$\widehat{\log (Y/L)} = -0.11 + 0.11 \log (K/L) + 0.047t \quad (2)$$
$$\text{se} = (0.04) \quad (0.15) \quad (0.006) \quad R^2 = 0.65, \quad F = 19.5$$

(a) Có đa cộng tuyến trong hàm hồi qui (1) hay không? Làm sao bạn biết?

(b) Trong hàm hồi qui (1), dấu tiên nghiệm của log K là gì? Các kết quả này có phù hợp với kỳ vọng này không? Tại sao có hoặc tại sao không?

(c) Bạn chứng minh dạng hàm hồi qui (1) như thế nào:(*Hướng dẫn:* Hàm sản xuất Cobb - Douglas.)

(d) Giải thích hàm hồi qui (1). Biến xu hướng đóng vai trò gì trong hàm hồi qui này?

(e) Tính logic của hàm hồi qui ước lượng (2) là gì?

---

* R. Stone đưa ra công thức này, "The Analysis of Market Demand," (Phân tích nhu cầu thị trường), *Journal of the Royal Statistical Society,* số B7, 1945, trang 297. Cũng nhớ lại (7.5.6). muốn biết thêm, xem Peter Kennedy, *A Guide to Econometrics,* (Hướng dẫn Kinh tế lượng), 2d ed., The MIT Press, Cambridge, Mass., 1985, trang 156.

* Christopher Dougherty, *Introduction to Econometrics,* (Nhập môn kinh tế lượng), Oxford University Press, New York, 1992, trang 159-160

(f) Nếu có đa cộng tuyến trong hàm hồi qui (1), thì vấn đề đa cộng tuyến này có bị giảm bớt trong hàm hồi qui (2) hay không? Bằng cách nào bạn biết được?

(g) Nếu hàm hồi qui (2) là một dạng giới hạn của hàm hồi qui (1), thì tác giả đã đặt ra sự giới hạn gì? (*Hướng dẫn*: quay lại phần phạm vi.) Bằng cách nào bạn biết được sự giới hạn này có hiệu lực hay không? Bạn sử dụng kiểm định gì? Trình bày mọi tính toán của bạn.

(h) Các giá trị R^2 của hai hàm hồi qui trên có thể so sánh được hay không? Tại sao có hoặc tại sao không? Bạn có thể làm cho chúng trở thành so sánh được bằng cách nào, nếu như hiện tại chúng không thể so sánh được?

Bài toán

10.25. Klein và Goldberger đã cố gắng để sử dụng mô hình hồi qui sau vào kinh tế Hoa Kỳ:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i$$

với Y = tiêu dùng, X_2 = thu nhập tiền lương, X_3 = thu nhập không phải từ tiền lương, không phải từ nông trại, và X_4 = thu nhập từ nông trại. Nhưng vì người ta kỳ vọng là X_2, X_3, và X_4 cộng tuyến cao, nên họ đã có được các giá trị ước lượng của β_3 và β_4 từ phân tích gộp là như sau: $\beta_3 = 0.75\beta_2$ và $\beta_4 = 0.625\beta_2$. Sử dụng các giá trị ước lượng này, họ thiết lập lại hàm tiêu dùng như sau:

$$Y_i = \beta_1 + \beta_2 (X_{2i} + 0.75 X_{3i} + 0.625X_{4i}) + u_i = \beta_1 + \beta_2 Z_i + u_i$$

với $Z_i = X_{2i} + 0.75 X_{3i} + 0.625X_{4i}$.

(a) Hãy làm cho mô hình đã hiệu chỉnh này thích hợp với các số liệu đi kèm và tìm các ước lượng của β_1 đến β_4.

(b) Bạn giải thích biến Z như thế nào?

Năm	Y	X_2	X_3	X_4	Năm	Y	X_2	X_3	X_4
1936	62.8	43.41	17.10	3.96	1946	95.7	76.73	28.26	9.76
1937	65.0	46.44	18.65	5.48	1947	98.3	75.91	27.91	9.31
1938	63.9	44.35	17.09	4.37	1948	100.3	77.62	32.30	9.85
1939	67.5	47.82	19.28	4.51	1949	103.2	78.01	31.39	7.21
1940	71.3	51.02	23.24	4.88	1950	108.9	83.57	35.61	7.39
1941	76.6	58.71	28.11	6.37	1951	108.5	90.59	37.58	7.98
1945*	86.3	87.69	30.29	8.96	1952	111.4	95.47	35.17	7.42

*Số liệu trong những năm chiến tranh 1942-1944 bị thiếu. Số liệu của những năm khác là hàng triệu của 1939 đô - la.

Nguồn: L. R. Klein và A. S. Goldberger, *An Economic Model of the United States,* (Mô hình kinh tế của Mỹ) 1929-1952, North Holland Publishing Company, Amsterdam, 1964, trang 131

10.26. Bảng sau đây cho số liệu về nhập khẩu, GNP, và chỉ số giá tiêu dùng (CPI) của Mỹ trong thời kỳ 1970-1983.

Hàng hóa nhập khẩu, GNP, và CPI, Mỹ, 1970 – 1983

Năm	Nhập khẩu hàng hóa (triệu $)	GNP (tỉ $)	CPI, mọi hạng mục (1967 = 100)
1970	39,866	992.7	116.3

1971	45,579	1,077.6	121.3
1972	55,797	1,185.9	125.3
1973	70,499	1,326.4	133.1
1974	103,811	1,434.2	147.7
1975	98,185	1,549.2	161.2
1976	124,228	1,718.0	170.5
1977	151,907	1,918.3	181.5
1978	176,010	2,163.9	195.4
1979	212,028	2,417.8	217.4
1980	249,781	2,631.7	246.8
1981	265,086	2,957.8	272.4
1982	247,667	3,069.3	289.1
1983	261,312	3,304.8	298.4

Nguồn: Economic Report of the President, 1985. Số liệu về nhập khẩu từ bảng B-98 (trang 344), GNP từ bảng B-1 (trang 232) và CPI từ bảng B-52 (trang 291)

Bạn hãy xem mô hình sau:

$$\ln \text{Nhập khẩu}_t = \beta_1 + \beta_2 \ln \text{GNP}_t + \beta_3 \ln \text{CPI}_t + u_i$$

(a) Ước lượng các thông số của mô hình này, sử dụng số liệu cho trong bảng.

(b) Bạn có nghi ngờ là có đa cộng tuyến trong số liệu hay không?

(c) Kiểm tra bản chất của cộng tuyến, sử dụng chỉ số điều kiện.

(d) Lập hàm hồi qui: (1) $\ln \text{Nhập khẩu}_t = A_1 + A_2 \ln \text{GNP}_t$

(2) $\ln \text{Nhập khẩu}_t = B_1 + B_2 \ln \text{CPI}_t$

(3) $\ln \text{GNP}_t = C_1 + C_2 \ln \text{CPI}_t$

Dựa vào những hàm hồi qui này, bạn có thể nói gì về bản chất của đa cộng tuyến trong số liệu?

(e) Giả sử là có đa cộng tuyến trong số liệu nhưng $\hat{\beta}_2$ và $\hat{\beta}_3$ có ý nghĩa riêng biệt ở mức ý nghĩa 5% và kiểm định F toàn diện cũng có ý nghĩa. Trong trường hợp này chúng ta có nên quan tâm về vấn đề cộng tuyến hay không?

10.27. Liên quan đến bài tập 7.23 về hàm nhu cầu gà ở Mỹ.

(a) Sử dụng mô hình logarit tuyến tính, hoặc logarit kép (double-log), để ước lượng các hàm hồi qui phụ trợ khác nhau. Có bao nhiêu hàm này?

(b) Từ những hàm hồi qui phụ trợ này, bạn quyết định xem hàm hồi qui nào thì cộng tuyến cao bằng cách nào? Bạn sử dụng kiểm định gì? Trình bày chi tiết các tính toán của bạn.

(c) Nếu có cộng tuyến cao trong số liệu, những biến nào bạn sẽ bỏ đi để giảm mức độ trầm trọng của vấn đề cộng tuyến? Nếu bạn làm như vậy, bạn sẽ gặp phải vấn đề kinh tế lượng gì?

(d) Bạn có đề nghị nào khác cách bỏ một số biến để giảm bớt vấn đề cộng tuyến? Giải thích.

10.28. Bảng kèm theo đây trình bày số liệu về loại xe hơi chở khách mới được bán ở Mỹ như một hàm của nhiều biến.

(a) Xây dựng một mô hình tuyến tính hoặc logarit tuyến tính để ước lượng hàm cầu về xe ô tô ở Mỹ.

(b) Nếu bạn quyết định chọn tất cả các biến hồi qui độc lập cho trong bảng làm biến giải thích, bạn có nghĩ là sẽ gặp phải vấn đề đa cộng tuyến không? Tại sao?

(c) Nếu gặp phải vấn đề đó, bạn sẽ giải quyết bằng cách nào? Nêu các giả định của bạn một cách rõ ràng và trình bày mọi tính toán thật chi tiết.

Năm	Y	X_2	X_3	X_4	X_5	X_6
1971	10,227	112.0	121.3	776.8	4.89	79,367
1972	10,872	111.0	125.3	839.6	4.55	82,153
1973	11,350	111.1	133.1	949.8	7.38	85,064
1974	8,775	117.5	147.7	1,038.4	8.61	86,784
1975	7,539	127.6	161.2	1,142.8	6.16	85,846
1976	9,994	135.7	170.5	1,252.6	5.22	88,752
1977	11,046	142.9	181.5	1,379.3	5.50	92,017
1978	11,194	153.8	195.3	1,551.2	7.78	96,048
1979	10,559	166.0	217.7	1,729.3	10.25	98,824
1980	8,979	179.3	247.0	1,918.0	11.28	99,303
1981	8,535	190.2	272.3	2,127.6	13.73	100,397
1982	4,980	197.6	286.6	2,261.4	11.20	99,526
1983	9,179	202.6	297.4	2,428.1	8.69	100,834
1984	10,394	208.5	307.6	2,670.6	9.65	105,005
1985	11,039	215.2	318.5	2,841.1	7.75	107,150
1986	11,450	224.4	323.4	3,002.1	6.31	109,597

Y = Xe hơi chở khách mới được bán (hàng ngàn), không điều chỉnh theo mùa

X_2 = Xe hơi mới, Chỉ số giá tiêu dùng,1967 = 100, không điều chỉnh theo mùa

X_3 = Chỉ số giá tiêu dùng, mọi mục, mọi người tiêu dùng thành thị, 1967 = 100, không điều chỉnh theo mùa

X_4 = thu nhập cá nhân có thể chi tiêu được (PDI), tỉ đô-la, không điều chỉnh theo mùa

X_6 = lực lượng lao động đô thị có nghề nghiệp (hàng ngàn), không điều chỉnh theo mùa

Nguồn: Business Statistics, 1986, A Supplement to the Curent Survey of Business U. S. Deparment of Commerce