

ĐA CỘNG TUYẾN

1. Giới Thiệu Đa Cộng Tuyến Trong Kinh Tế Lượng

Nhớ lại giả định ban đầu

- ◆ Giả định CLRM (mô hình hồi qui tuyến tính cổ điển): Các biến độc lập không có mối quan hệ tuyến tính chính xác (exact linear relationship)
 - Nếu điều này xảy ra thì sẽ có hiện tượng đa cộng tuyến, đó là hiện tượng các biến độc lập trong mô hình phụ thuộc lẫn nhau và thể hiện được dưới dạng hàm số.

Ví dụ

- ◆ Đa cộng tuyến hoàn hảo:
 - X_2 X_3 X_4
 - 10 50 52
 - 15 75 75
 - 18 90 97
 - 24 120 129
 - X_2 và X_3 có mối quan hệ tuyến tính chính xác $X_3 = 5X_2$

Ví dụ (tt)

- ◆ Giả sử chúng ta ước lượng hàm tiêu dùng. Y = tiêu dùng, X_2 = thu nhập và X_3 = của cải
- ◆ $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3$
- ◆ $X_3 = 5X_2$
- ◆ $Y = \beta_1 + \beta_2 X_2 + \beta_3 5X_2$
- ◆ $Y = \beta_1 + (\beta_2 + 5\beta_3)X_2$

Ví dụ (tt)

- ◆ Chúng ta có thể ước lượng $(\beta_2 + 5\beta_3)$ nhưng không ước lượng riêng từng hệ số hồi qui
 - Không thể có nghiệm duy nhất cho từng hệ số hồi qui (xem lại cách tính các hệ số hồi qui).
 - Như vậy các hệ số hồi qui sẽ không xác định được.
 - ◆ Sai số chuẩn của các hệ số hồi qui là một vô cùng lớn.

Đa Cộng Tuyến

- ◆ Đa cộng tuyến hoàn hảo thường rất ít khi xảy ra trong thực tế
 - Trừ trường hợp chúng ta rơi vào bẫy biến giả (dummy trap – chúng ta sẽ giới thiệu sau)
- ◆ Đa cộng tuyến không hoàn hảo thường hay xảy ra trong thực tế (Near collinearity) (khi các biến độc lập tương quan khá cao):
 - Trường hợp thứ hai chúng ta có thể ước lượng các hệ số hồi qui
 - ◆ Tuy nhiên sai số chuẩn rất lớn và vì vậy hệ số hồi qui ước lượng không chính xác, kiểm định ít có ý nghĩa thống kê và dễ dàng bác bỏ giả thuyết “không”

Đa Cộng Tuyến

- ◆ Nghiên cứu tình huống

2. Nguồn gốc của Đa Cộng Tuyến

Nguồn gốc Đa cộng tuyến

- ◆ Do phương pháp thu thập dữ liệu
 - các giá trị của các biến độc lập phụ thuộc lẫn nhau trong mẫu, nhưng không phụ thuộc lẫn nhau trong tổng thể
 - Ví dụ: người có thu nhập cao hơn khuynh hướng sẽ có nhiều của cải hơn. Điều này có thể đúng với mẫu mà không đúng với tổng thể
 - ◆ Trong tổng thể sẽ có các quan sát về các cá nhân có thu nhập cao nhưng không có nhiều của cải và ngược lại.

Nguồn gốc Đa cộng tuyến

◆ Dạng hàm mô hình:

- Ví dụ: hồi qui dạng các biến độc lập được bình phương (dạng hàm) sẽ xảy ra đa cộng tuyến và đặc biệt khi phạm vi giá trị ban đầu của biến độc lập là nhỏ.

◆ Các biến độc lập vĩ mô được quan sát theo dữ liệu chuỗi thời gian

- ◆ Ví dụ: Nhập khẩu quốc gia phụ thuộc vào GDP và CPI (các chỉ số này được thu thập từ dữ liệu chuỗi thời gian). Giải thích đa cộng tuyến theo ý nghĩa vĩ mô?

3. Hệ quả của Đa Cộng Tuyến

Hệ quả lý thuyết

◆ Đa cộng tuyến hoàn hảo

- Chúng ta không thể ước lượng được mô hình
- Các phần mềm máy tính sẽ báo các tín hiệu sau
 - ◆ "Matrix singular": ma trận khác thường mà máy tính không thể thực hiện được khi ước lượng các hệ số hồi qui
 - ◆ "Exact collinearity encountered": trường hợp đa cộng tuyến hoàn hảo (chính xác)

Hệ quả lý thuyết

◆ Hệ quả khi có đa cộng tuyến không hoàn hảo

- Ước lượng OLS vẫn BLUE
 - ◆ Ước lượng không chệch: trung bình các ước lượng từ mẫu lập lại sẽ hội tụ đến giá trị ước lượng của tổng thể.

Hệ quả thực tiễn

- ◆ Sai số chuẩn của các hệ số sẽ lớn.
 - Khoảng tin cậy lớn và thống kê t ít ý nghĩa.
 - Các ước lượng không thật chính xác.
 - Do đó chúng ta dễ đi đến không có cơ sở bác bỏ giả thiết "không" và điều này có thể không đúng.

Hệ quả thực tiễn

- ◆ R^2 rất cao cho dù thống kê t ít ý nghĩa
 - Tại sao hệ số xác định lại cao?
 - ◆ Không có nhiều những biến đổi khác biệt giữa các biến số độc lập vì chúng thực sự có mối quan hệ với nhau
 - Dễ dàng bác bỏ giả thuyết "không" của thống kê F và cho rằng mô hình ước lượng có giá trị

Hệ quả thực tiễn

- ◆ Các ước lượng sẽ không chính xác
 - Chỉ cần một sự thay đổi nhỏ trong mẫu dữ liệu sẽ kéo theo sự thay đổi lớn các hệ số ước lượng.
 - Bởi vì các hệ số ước lượng chứa đựng những mối quan hệ mạnh giữa các biến độc lập
- ◆ Dấu vài hệ số sẽ khác với kỳ vọng
 - Do các hệ số này không còn đủ sức giải thích tác động biên lên biến phụ thuộc vì mối quan hệ pha trộn giữa các biến độc lập

Ví dụ

- ◆ Xem kết quả ước lượng hàm tiêu dùng:
 - $Y = 24.77 + 0.94X_2 - 0.04X_3$
 - t (3.67) (1.14) (-0.53)
 - $R^2=0.96$, $F = 92.40$
 - X_2 : thu nhập
 - X_3 : của cải
 - R^2 rất cao giải thích 96% biến đổi của hàm tiêu dùng

Ví dụ

- Không có biến độc lập nào có ý nghĩa (thống kê t quá thấp).
- Có một biến sai dấu.
- Giá trị thống kê F rất cao dẫn đến bác bỏ giả thuyết "không" và cho rằng mô hình ước lượng có ý nghĩa.
 - ◆ Biến thu nhập và của cải tương quan rất mạnh với nhau do đó không thể nào ước lượng được tác động biên chính xác cho thu nhập hoặc của cải lên tiêu dùng.

Ví dụ

- ◆ Thực hiện hồi qui X_3 theo X_2
 - $X_3 = 7.54 + 10.19X_2$
 - (0.26) (62.04) $R^2 = .99$
 - Hầu như chúng ta có đa cộng tuyến hoàn hảo giữa X_2 và X_3
- ◆ Hồi qui tiêu dùng theo thu nhập:
 - $Y = 24.45 + 0.51X_2$
 - (3.81) (14.24) $R^2 = 0.96$

Ví dụ

- Biến thu nhập trở nên có ý nghĩa thống kê, nhưng trước lúc đó trong mô hình đầu thì không có ý nghĩa.
- ◆ Tương tự hồi qui thu nhập Y theo của cải:
 - $Y = 24.41 + 0.05X_3$
 - $t \quad (3.55) \quad (13.29) \quad R^2 = 0.96$
 - Biến của cải trở nên có ý nghĩa thống kê, nhưng trước lúc đó trong mô hình đầu thì không có ý nghĩa.

4. Nhận biết Đa Cộng Tuyến

Các phương pháp nhận biết

- ◆ R^2 cao và thống kê t thấp.
- ◆ Tương quan tuyến tính mạnh giữa các biến độc lập
 - Xây dựng ma trận hệ số tương quan cặp và quan sát để nhận diện độ mạnh của các tương quan từng cặp biến số độc lập
 - Xét về ý nghĩa kinh tế các biến có khả năng tương quan cao

Các phương pháp nhận biết

- Thực hiện hồi qui phụ
- Hồi qui giữa một biến độc lập với tất cả các biến độc lập với nhau và quan sát hệ số R^2 của các hồi qui phụ
- Thực hiện tính thống kê F
 - ◆ $F = [R^2/(k-1)] / [(1-R^2)/(n-k)]$
 - ◆ k số biến độc lập trong hồi qui phụ
 - ◆ Nếu $F > F^*$ thì chúng ta có thể kết luận rằng R^2 khác không theo ý nghĩa thống kê và điều này có nghĩa là có đa cộng tuyến trong mô hình.

Các phương pháp nhận biết

- ◆ Thừa số tăng phương sai (Variance inflation factor-VIF)
 - $VIF = 1/(1-r_{ij}^2)$
 - r_{ij}^2 là hệ số tương quan giữa hai biến độc lập trong mô hình.
 - Khi r_{ij} tăng làm VIF tăng và làm tăng mức độ đa cộng tuyến
 - ◆ Rule of thumb ≥ 10 có hiện tượng đa cộng tuyến giữa hai biến độc lập trong mô hình

5. Các giải pháp khắc phục Đa cộng tuyến

Rules of Thumb khi bỏ qua nhẹ nhàng Đa cộng tuyến

- ◆ Bỏ qua đa cộng tuyến nếu $t > 2$
- ◆ Bỏ qua đa cộng tuyến nếu R^2 của mô hình cao hơn R^2 của mô hình hồi qui phụ.
- ◆ Bỏ qua đa cộng tuyến nếu mục tiêu xây dựng mô hình sử dụng để dự báo chứ không phải kiểm định.

Các giải pháp nếu xét nghiêm ngặt Đa cộng tuyến

- ◆ Bỏ bớt biến độc lập.
 - Ví dụ: bỏ biến của cái ra khỏi mô hình hàm tiêu dùng.
 - Điều này xảy ra với giả định rằng không có mối quan hệ giữa biến phụ thuộc và biến độc lập loại bỏ mô hình.
 - ◆ Nếu lý thuyết khẳng định có mối quan hệ với biến dự định loại bỏ thì việc loại bỏ này sẽ dẫn đến loại bỏ biến quan trọng và chúng ta mắc sai lầm về nhận dạng mô hình (specification error)

Các giải pháp

- ◆ Bổ sung dữ liệu hoặc tìm dữ liệu mới
 - Tìm mẫu dữ liệu khác hoặc gia tăng cỡ mẫu
 - ◆ Nếu mẫu lớn hơn mà vẫn còn multicollinearity thì vẫn có giá trị vì mẫu lớn hơn sẽ làm cho phương sai nhỏ hơn và hệ số ước lượng chính xác hơn so với mẫu nhỏ.

Các giải pháp

- ◆ Thay đổi dạng mô hình
 - Mô hình kinh tế lượng có nhiều dạng hàm khác nhau.
 - Thay đổi dạng mô hình cũng có nghĩa là tái cấu trúc mô hình
- ◆ Sử dụng thông tin hậu nghiệm "priori information"
 - Sử dụng kết quả của các mô hình kinh tế lượng trước ít có đa cộng tuyến
 - Ví dụ: chúng ta có thể biết tác động biên của của cải lên tiêu dùng chỉ bằng 1/10 so với tác động biên của thu nhập lên tiêu dùng.

Giải pháp

- Ví dụ: $\beta_3 = 0.10 \beta_2$
 - ◆ Chạy mô hình với điều kiện tiên nghiệm.
 - ◆ $Y = \beta_1 + \beta_2 X_2 + 0.10 \beta_2 X_3 + e$
 - $Y = \beta_1 + \beta_2 X$ trong đó $X = X_2 + 0.1X_3$
 - ◆ Khi ước lượng được β_2 thì suy ra β_3 từ mối quan hệ tiên nghiệm trên.

Giải pháp

- ◆ Sử dụng sai phân cho các biến của mô hình
 - Sai phân làm cho vấn đề đa cộng tuyến có thể nhẹ đi
 - Quay trở lại ví dụ hàm tiêu dùng.
 - ◆ Thu nhập và của cải có mối quan hệ khá chặt chẽ và do đó không tránh khỏi đa cộng tuyến

Giải pháp

- Chúng ta muốn ước lượng
 - ◆ $Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + e_t$
- Ứng với t-1
 - ◆ $Y_{t-1} = \beta_1 + \beta_2 X_{2t-1} + \beta_3 X_{3t-1} + e_{t-1}$
- Lấy sai phân các biến theo thời gian
 - ◆ $Y_t - Y_{t-1} = \beta_2 (X_{2t} - X_{2t-1}) + \beta_3 (X_{3t} - X_{3t-1}) + v_t$

Giải pháp

- Điều này có thể giải quyết vấn đề đa cộng tuyến vì đa cộng tuyến xảy ra từ bản thân các biến độc lập chứ không xảy ra từ sai phân các biến này.
- Tuy nhiên có thể vi phạm giả định chuẩn về sai số ngẫu nhiên.

Giải pháp

◆ Kết hợp dữ liệu chéo và dữ liệu chuỗi thời gian

- Ví: Nghiên cứu cầu xe hơi và chỉ có dữ liệu chuỗi thời gian.
 - ◆ $\ln Y = \beta_1 + \beta_2 \ln \text{Price} + \beta_3 \ln \text{Income} + e$
 - Y số xe hơi bán ra.
 - ◆ Thông thường giá và thu nhập tương quan mạnh với nhau theo thời gian nên chắc chắn mô hình có đa cộng tuyến khi sử dụng chuỗi thời gian

Giải pháp

- ◆ Giả sử chúng ta có dữ liệu chéo
 - Chúng ta có thể ước lượng độ co giãn theo thu nhập khi sử dụng dữ liệu chéo. Còn độ co giãn theo giá chúng ta phải tìm từ chuỗi dữ liệu theo thời gian
- ◆ Ước lượng hàm hồi quy theo thời gian
 - $Y = \beta_1 + \beta_2 \ln P + e$
 - Khi đó $Y = \ln Y - \beta_3 \ln \text{Income}$
 - Y đại diện cho số xe hơi bán ra sau khi loại trừ tác động của thu nhập
 - Căn cứ vào β_3 cho trước chúng ta ước lượng được độ co giãn cầu xe hơi theo giá nhưng không có hiện tượng Đa cộng tuyến
 - Tuy nhiên chúng ta phải giả định rằng, độ co giãn từ chuỗi thời gian và từ dữ liệu chéo là đồng nhất.