

Phân tích nhận dạng mô hình

Bài giảng này sẽ thảo luận hai nội dung: những biến hồi qui nào sẽ đưa vào/bỏ ra ngoài một mô hình cụ thể, và những dạng hàm nào có thể là phù hợp đối với biến phụ thuộc và các biến hồi qui này.

Sự loại trừ các biến liên quan

Giả sử chúng ta quan tâm mô hình sau đây:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \dots + \beta_K X_{Ki} + \beta_{(K+1)} X_{(K+1)i} + \dots + \beta_{(K+L)} X_{(K+L)i} + \varepsilon_i$$

Vấn đề đặt ra là liệu tập hợp L biến hồi qui dưới đây

$$X_{(K+1)} + \dots + X_{(K+L)}$$

cần đưa vào mô hình. Một mặt là khi bổ sung các biến hồi qui có thể tăng thêm sức mạnh giải thích, và việc bỏ sót chúng có thể tạo ra "các sai lệch do bỏ sót biến". Nhưng mặt khác, các biến hồi qui bổ sung lại đòi hỏi tăng thêm các hệ số ước lượng, điều này làm giảm bậc tự do các ước lượng của chúng ta (tức là làm cho các phân phối chọn mẫu của chúng phân tán hơn). Hơn nữa, mức độ mà các biến hồi qui có tương quan tuyến tính lẫn nhau cũng làm tăng các sai số chuẩn của các ước lượng này.

Việc đánh đổi này có thể được minh họa khi chúng ta trở lại với mô hình có hai biến giải thích.

Mô hình hồi qui hai biến giải thích có dạng:

$$\text{Mô hình thực (mô hình tổng thể)} \quad Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

Hệ số hồi qui bội cho biến X_2 được tính bằng công thức :

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

Phương sai của $\hat{\beta}_2$ là :

$$\text{VAR}[\hat{\beta}_2] = \frac{1}{\left(\sum x_{2i}^2\right)(1 - r_{23}^2)} \sigma^2$$

Bậc tự do tương ứng cho hồi qui có hai biến giải thích là : (n - 3).

Dường như chi phí phải trả cho việc thực hiện hồi qui bội thay vì chỉ thực hiện hồi qui đơn của Y theo X_2 là phương sai của ước lượng $\hat{\beta}_2$ đã tăng lên nếu X_2 và X_3 tương quan với nhau (mà điều này thì luôn xảy ra cho các dữ liệu kinh tế).

Nhưng chi phí đánh đổi này sẽ thấp so với chi phí của việc bỏ X_3 ra ngoài mô hình khi thực sự là X_3 có liên quan với Y. Xét hồi qui đơn như sau:

Dạng mô hình ước lượng $Y_i = \beta_1 + \beta_2 X_{2i} + \varepsilon_i$

Ước lượng bình phương tối thiểu của $\hat{\beta}_2$ là

$$\hat{\beta}_2 = \frac{\sum x_{2i} Y_i}{\sum x_{2i}^2}$$

Để tính kỳ vọng toán học của ước lượng này, chúng ta thay Y_i bằng biểu thức đại số của dạng mô hình thực:

$$E[\hat{\beta}_2] = E\left[\frac{\sum x_{2i} (Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i)}{\sum x_{2i}^2}\right]$$

$$E[\hat{\beta}_2] = \beta_2 + \beta_3 \frac{\sum x_{2i} X_{3i}}{\sum x_{2i}^2}$$

Như vậy, chúng ta thấy rằng nếu X_3 có liên quan thực sự với Y (có nghĩa là $\beta_3 \neq 0$) và X_2 và X_3 tương quan với nhau, thì hệ số hồi qui đơn $\hat{\beta}_2$ là chệch. Biểu thức cho dưới đây

$$\beta_3 \frac{\sum x_{2i} X_{3i}}{\sum x_{2i}^2}$$

được gọi là chệch do bỏ sót biến và chúng ta có thể lưu ý biểu thức dưới đây

$$\hat{V}_2 = \frac{\sum x_{2i}x_{3i}}{\sum x_{2i}^2}$$

là hệ số độ dốc của hồi qui đơn của X_3 theo X_2 .

Kết quả này khái quát cho các mô hình tổng quát hơn. Tránh hiện tượng chệch do bỏ sót biến là điều mong muốn cho dù các biến hồi qui bổ sung sẽ tạo ra các sai số chuẩn lớn hơn. Mong muốn tránh bỏ sót biến có thể khuyến khích chúng ta để dành khi bổ sung các biến hồi qui vào mô hình của mình.

Tuy nhiên, chúng ta cần hỏi xem điều gì xảy ra nếu chúng ta bổ sung các biến hồi qui mà thực tế chúng không phù hợp.

Đưa vào các biến không liên quan

Để phân tích trường hợp này, chúng ta trở lại một lần nữa với mô hình có hai biến hồi qui, chỉ có lần này chúng ta giả định rằng X_3 là không có quan hệ đến Y (có nghĩa là $\beta_3 = 0$).

Mô hình thực $Y_i = \beta_1 + \beta_2 X_{2i} + \varepsilon_i$

Mô hình ước lượng $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$

Một lần nữa, chúng ta lấy các hệ số ước lượng và tính các kỳ vọng của chúng:

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

Thay Y_i bằng mô hình thực và biến đổi một chút:

$$\hat{\beta}_2 = \frac{\beta_2 (\sum x_{2i}^2)(\sum x_{3i}^2) - \beta_2 (\sum x_{2i} x_{3i})^2}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2} + \frac{(\sum \varepsilon_i x_{2i})(\sum x_{3i}^2) - (\sum \varepsilon_i x_{3i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2}$$

Rõ ràng, số hạng thứ nhất là β_2 và số hạng thứ hai có kỳ vọng bằng không, vì thế ước lượng này là không chệch.

Phương sai của ước lượng này là:

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

và chúng ta thấy rằng việc đưa X_3 không phù hợp vào làm tăng phương sai nếu X_2 và X_3 tương quan với nhau.

Bây giờ hãy xét ước lượng hệ số cho biến không phù hợp nhưng lại được đưa vào mô hình:

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

Một lần nữa, thay Y_i bằng mô hình thực và biến đổi một chút

$$\hat{\beta}_3 = 0 + \frac{(\sum \varepsilon_i x_{3i})(\sum x_{2i}^2) - (\sum \varepsilon_i x_{2i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2}$$

Kỳ vọng của ước lượng này bằng không.

Do đó chúng ta thấy rằng khi chúng ta đưa các biến không phù hợp vào, chúng ta nhận được các ước lượng không chệch của tất cả các hệ số, nhưng chi phí phải trả là những phương sai tối thiểu lớn hơn so với trường hợp mà chúng ta đã không đưa biến không phù hợp vào mô hình.

Xây dựng mô hình từ tổng quát tới đơn giản

Những kết quả mà chúng ta vừa hình thành gợi ý rằng chiến lược xây dựng mô hình nên đi từ tổng quát tới đơn giản tốt hơn là đi từ đơn giản tới tổng quát. Các bước là:

- [Sử dụng lý thuyết kinh tế, nghiên cứu trước đây, và kinh nghiệm để xác định một mô hình tổng quát (trong trường hợp này, "tổng quát" có nghĩa là một mô hình bao gồm tất cả mọi biến có thể có liên quan).
- [Ước lượng mô hình.
- [Nếu bất cứ hệ số nào trong những hệ số ước lượng không có ý nghĩa thống kê, thì chúng ta nên bỏ đi biến ít ý nghĩa nhất và ước lượng lại mô hình với các biến số còn lại. Nên loại bỏ từng biến một bởi vì ảnh hưởng của việc loại bỏ lên các phương sai của những biến còn lại. Nếu hồi qui lần thứ nhất cho chúng ta thấy

có hai biến không có ý nghĩa thống kê, thì biến ít ý nghĩa nhất sẽ bị bỏ ra trước, điều này có thể làm tăng mức ý nghĩa của biến kia.

[Xét chi phí phải trả cho các sai lầm loại 1 và loại 2 khi lựa chọn một mức ý nghĩa.

Sai lầm loại 1 trong trường hợp này có nghĩa là Anh/Chị giữ lại một biến không phù hợp, và vì thế làm tăng các phương sai của các ước lượng hệ số khi áp dụng phương pháp bình phương tối thiểu.

Sai lầm loại 2 trong trường hợp này có nghĩa là Anh/Chị bỏ sót một biến phù hợp, và vì thế tạo ra ước lượng chệch do bỏ sót biến trong những hệ số còn lại.

Những cân nhắc này đề xuất rằng nhà nghiên cứu có thể muốn kiểm soát khả năng xảy ra ước lượng chệch do bỏ sót biến bằng cách lựa chọn một mức ý nghĩa khá lớn cho việc kiểm định các biến trong qui trình từ tổng quát tới đơn giản.

[Nghiên cứu các hệ số của các biến còn lại đối với những thay đổi lớn khi chúng ta bỏ sót một biến; các thay đổi lớn có thể tạo ra ước lượng chệch do bỏ sót biến ngay cả khi hệ số của biến mà Anh/Chị định bỏ không có ý nghĩa thống kê.

Các kiểm định Wald

Đôi khi, quan tâm của chúng ta tập trung vào kiểm định giả thiết kết hợp ví dụ như giả thiết dưới đây:

$$H_0 : \beta_{(K+1)} = \beta_{(K+2)} = \dots = \beta_{(K+L)} = 0$$

hay vào giả thiết thể hiện một sự kết hợp tuyến tính như thế này:

$$H_0 : \beta_2 + \beta_3 = 1$$

Chúng ta có thể xem các mô hình dưới giả thiết “không” và giả thiết thay thế là các mô hình *giới hạn* và *không có giới hạn*. Chúng ta coi “giới hạn” có nghĩa là các hệ số của mô hình phải thỏa mãn các giới hạn tuyến tính được xác định bởi các giả thiết nêu ra.

Trong trường hợp thứ nhất, các mô hình có giới hạn và không có giới hạn trông có thể giống như thế này :

$$Y_i = \beta_1 + \beta_2 X_{2i} + \dots + \beta_K X_{Ki} + \xi_i$$

$$Y_i = \beta_1 + \beta_2 X_{2i} + \dots + \beta_K X_{Ki} + \beta_{(K+1)} X_{(K+1)i} + \dots + \beta_{(K+L)} X_{(K+L)i} + \epsilon_i$$

Lưu ý là các thành phần sai số sẽ khác nhau vì sai số thứ nhất bao hàm tác động của tất cả mọi biến bị bỏ sót.

Trong trường hợp thứ hai, các mô hình có giới hạn và không có giới hạn có thể thể hiện dưới dạng sau:

$$Y_i = \beta_1 + \beta_2 X_{2i} + (1 - \beta_2) X_{3i} + \xi_i$$

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

Mỗi mô hình mà chúng ta ước lượng bằng bình phương bình phương tối thiểu thì tổng bình phương phần dư được mô tả là

$$RSS = \sum e_i^2$$

Nếu chúng ta áp đặt một giới hạn ràng buộc (tức là, nếu dữ liệu không thích hợp với giới hạn do chúng ta áp đặt), thì việc nguyên tắc tối thiểu hoá cũng không còn hiệu lực và RSS sẽ lớn hơn. Thực tế chúng ta có thể chứng minh rằng:

$$RSS_U \leq RSS_R$$

Điều này đề xuất một kiểm định thống kê: nếu RSS đã tăng lên có ý nghĩa thống kê, thì chúng ta cần bác bỏ những giới hạn đặt ra trong giả thuyết “không” này.

Kiểm định Wald chính thức kiểm tra ý tưởng này:

$$\frac{(RSS_R - RSS_U) / (df_R - df_U)}{RSS_U / df_U} = F^* \sim F_{((df_R - df_U), df_U)}$$

Giải thích của kiểm định này phản ánh sự thực là nếu các giới hạn không phải là bắt buộc, thì chúng ta kỳ vọng biến ngẫu nhiên này sẽ tiến về không; các giá trị lớn của trị thống kê này đề xuất việc bác bỏ giả thuyết “không”. Vì vậy, chúng ta có nguyên tắc ra quyết định sau đây cho kiểm định đuôi bên phải:

Nếu $F_{((1-\alpha), (df_R - df_U), df_U)} < F^*$ thì chúng ta bác bỏ giả thiết “không”.

Việc thực hiện các nội dung trên bằng EViews được thảo luận trong Bài tập 12.