

Mô hình hồi qui tuyến tính bội

Có nhiều ví dụ cho không thể giải thích bằng mô hình hồi qui đơn:

Lượng cầu phụ thuộc vào giá, thu nhập, giá các hàng hoá khác v.v...

Sản lượng phụ thuộc vào giá, các nhập lượng ban đầu, các nhập lượng trung gian, công nghệ v.v...

Đầu tư nước ngoài (FDI) phụ thuộc vào suất sinh lợi của đầu tư, tiền lương, tham nhũng, tính minh bạch v.v...

Khi chúng ta có một tập hợp dữ liệu có chứa các nhiều biến giải thích, trường hợp phổ biến là tất cả các biến này cùng thay đổi, điều đó làm chúng ta không thể cố định ảnh hưởng của một biến giải thích nào đó đến biến phụ thuộc.

Việc xét đến các tác động riêng biệt của nhiều nhân tố có thể được giải thích bằng mô hình hồi qui đa biến.

Hàm hồi qui tổng thể

Mô hình :

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_K X_{Ki} + \varepsilon_i \quad \text{PRF}$$

$$E[Y_i | X_i] = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_K X_{Ki} + E[\varepsilon_i | X_i] \quad \text{PRF}$$

Các hệ số β được gọi là các hệ số hồi qui riêng; mỗi hệ số được giải thích như sau:

$$\frac{\partial E[Y_i | X_i]}{\partial X_k} = \beta_k$$

Để nhận dạng mô hình, chúng ta bổ sung thêm một giả định vào các giả định cổ điển trước đây trong hồi qui đơn. Giả định bổ sung này là :

Những biến hồi qui này không có mối quan hệ tuyến tính hoàn hảo. Tức là, không tồn tại tập hợp các hệ số thỏa mãn biểu thức sau với mọi i :

$$1 + \lambda_2 X_{2i} + \lambda_3 X_{3i} + \dots + \lambda_K X_{Ki} = 0$$

Giả định này được gọi là *không có tính đa cộng tuyến hoàn hảo*.

Hãy ghi nhận rằng chúng ta một lần nữa lại có thể xác định hoàn toàn phân phối xác suất của biến phụ thuộc này:

$$Y_i \sim N(\beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \dots + \beta_K X_{Ki}, \sigma^2)$$

Hàm hồi qui mẫu và các ước lượng bình phương tối thiểu

Chúng ta đề cập tới vấn đề ước lượng bằng cách nhận dạng hàm hồi qui mẫu (SRF):

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \dots + \hat{\beta}_K X_{Ki}$$

Các phần dư được định nghĩa giống như chúng được xác định trong phương pháp hồi qui đơn:

$$e_i = Y_i - \hat{Y}_i$$

Bằng định nghĩa này, chúng ta có thể nhớ lại nguyên tắc bình phương tối thiểu để tìm các ước lượng của các hệ số hồi qui riêng.

Chọn $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_K$ sao cho $\sum e_i^2$ là tối thiểu.

$$\text{Lưu ý là } \sum e_i^2 = \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i} - \dots - \hat{\beta}_K X_{Ki})^2$$

Chúng ta có thiết lập các điều kiện bậc nhất cho phép tính tối thiểu này như sau :

$$\frac{\partial \sum e_i^2}{\partial \hat{\beta}_1} = -2 \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i} - \dots - \hat{\beta}_K X_{Ki}) = 0$$

$$\frac{\partial \sum e_i^2}{\partial \hat{\beta}_2} = -2 \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i} - \dots - \hat{\beta}_K X_{Ki}) X_{2i} = 0$$

⋮

$$\frac{\partial \sum e_i^2}{\partial \hat{\beta}_K} = -2 \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i} - \dots - \hat{\beta}_K X_{Ki}) X_{Ki} = 0$$

Chúng ta có thể giải K phương trình chuẩn này để tìm K hệ số beta chưa biết. Sự trình bày đơn giản nhất của lời giải này ở dưới dạng đại số ma trận.

Trường hợp hàm hồi qui có hai biến giải thích

Chúng ta có thể tìm lời giải cho mô hình có chứa hai biến hồi qui:

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

Các ước lượng tìm được từ phương pháp bình phương tối thiểu là :

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}_2 - \hat{\beta}_3 \bar{X}_3$$
$$\hat{\beta}_2 = \frac{(\sum y_i x_{2i})(\sum x_{3i}^2) - (\sum y_i x_{3i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2}$$
$$\hat{\beta}_3 = \frac{(\sum y_i x_{3i})(\sum x_{2i}^2) - (\sum y_i x_{2i})(\sum x_{2i} x_{3i})}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2}$$

Anh/Chị không cần nhớ các biểu thức này, nhưng chúng sẽ được sử dụng chúng để minh họa các kết quả nào đó.

Ví dụ , nếu trong mẫu của chúng ta, hai biến hồi qui này không có tương quan hoàn hảo với nhau thì chúng ta có

$$\sum x_{2i} x_{3i} = 0$$

Nếu điều này đúng thì các hệ số hồi qui bội đơn giản là các hệ số hồi qui đơn.

Các phương sai

Các phương sai của hồi qui bội cũng trở nên phức tạp hơn. Chúng ta chỉ trình bày phương sai của $\hat{\beta}_2$:

$$\text{VAR}[\hat{\beta}_2] = \frac{\sum x_{3i}^2}{(\sum x_{2i}^2)(\sum x_{3i}^2) - (\sum x_{2i} x_{3i})^2} \sigma^2$$

Hãy nhớ lại định nghĩa của bình phương hệ số tương quan giữa X_2 và X_3 :

$$r_{23}^2 = \frac{(\sum x_{2i}x_{3i})^2}{(\sum x_{2i}^2)(\sum x_{3i}^2)}$$

Biến đổi một chút chúng ta có phương sai của $\hat{\beta}_2$ như sau:

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

Một lần nữa, nếu hai biến hồi qui này không tương quan với nhau, thì phương sai này đơn giản chỉ là phương sai của hồi qui đơn.

Nên lưu ý rằng tác động lên phương sai của mối tương quan mạnh giữa hai biến hồi qui: nó làm tăng phương sai của mỗi ước lượng hệ số. Khi hai biến hồi qui có tương quan với nhau, sẽ khó khăn hơn cho thuật toán hồi qui bội để tách được các tác động riêng biệt của chúng, và khó khăn này được phản ánh trong các ước lượng mà các ước lượng này có các phương sai lớn hơn. Hãy ghi nhận rằng điều này không giống với tính "không hiệu quả" của ước lượng so với ước lượng có phương sai tối thiểu. Mà trái lại nó có nghĩa là chính phương sai tối thiểu này đã lớn hơn. Chúng ta sẽ triển khai điều này ngay sau đây khi mà chúng ta thảo luận chi tiết hơn về tính đa cộng tuyến.

Hãy tìm hiểu tiếp mối quan hệ giữa hồi qui đơn và hồi qui bội. Hãy viết hàm hồi qui có hai biến giải thích dưới dạng độ lệch giữa giá trị quan sát và giá trị trung bình:

$$y_i = \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$$

Bây giờ hãy chỉ ra một cặp hồi qui đơn:

$$\hat{y}_i = \hat{\alpha} x_{3i} \quad v_i = y_i - \hat{y}_i$$

$$\hat{x}_{2i} = \hat{\gamma} x_{3i} \quad w_i = x_{2i} - \hat{x}_{2i}$$

Chúng ta xem mỗi phần dư v_i và w_i giống như dạng phần dư của biến ban đầu y_i và x_{2i} nhưng đã "loại trừ" tác động tuyến tính của x_{3i} . Nói cách khác, các tác động tuyến tính của x_{3i} đã được ước lượng và loại trừ y_i và x_{2i} .

Bây giờ, nếu chúng ta chỉ ra một phép hồi qui đơn thứ ba :

$$v_i = \beta_2 w_i + \xi_i$$

Ước lượng bình phương tối thiểu của β_2 trong phép hồi qui này tách riêng tác động của x_{2i} lên y_i , và cho phép kiểm soát các tác động của x_{3i} .

Ước lượng này được trình bày bằng biểu thức: $\hat{\beta}_2 = \frac{\sum v_i w_i}{\sum w_i^2}$. Liệu Anh/Chị có thể chứng minh rằng biểu thức này cho chúng ta ước lượng hệ số hồi qui bội (hãy nhớ mang mặt nạ bảo vệ khi cố gắng làm điều này).

Phân phối xác suất chọn mẫu của ước lượng bình phương tối thiểu

Để có khả năng xây dựng các khoảng tin cậy đối với các tham số chưa biết và kiểm định giả thuyết về chúng, chúng ta cần biết các phân phối xác suất chọn mẫu của những ước lượng này.

Khi đề cập về phân phối chọn mẫu của các ước lượng, chúng ta cần biết ba nội dung sau đây:

- Kỳ vọng toán học
- Phương sai
- Dạng hàm của phân phối chọn mẫu

Đầu tiên, hãy xét kỳ vọng của ước lượng $\hat{\beta}_2$:

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

Lưu ý rằng các chữ Y hoa đã thay thế cho các dạng độ lệch mà chúng ta đã sử dụng, phù hợp với nguyên tắc của Anh Cao. Bây giờ hãy thế

$$Y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$$

vào trong biểu thức ước lượng và thực hiện vài biến đổi đại số chúng ta sẽ có:

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

Khi chúng ta lấy giá trị kỳ vọng của biểu thức thứ hai, chúng ta thấy ước lượng của tham số là không chệch bởi vì chúng ta có kết quả sau đây:

$$E[\hat{\beta}_2] = \beta_2$$

Chúng ta hầu như đã quen với cách trình bày phương sai của ước lượng.

Cuối cùng, rõ ràng từ biểu thức trên là mỗi ước lượng đều là một tổ hợp tuyến tính của các biến ngẫu nhiên có phân phối chuẩn, vì thế mỗi ước lượng cũng đều có phân phối chuẩn.

Những kết quả suy luận nêu trên cũng đúng đối với hàm hồi qui bội có K biến. Những ước lượng tham số là không chệch, và các phương sai của ước lượng tham số đã biết, và chúng tuân theo phân phối chuẩn. Tuy vậy, trên thực tế không thể biểu diễn được những kết quả này nếu không sử dụng công cụ đại số ma trận.

Chúng ta tóm tắt kết quả điển hình dưới dạng sau: $\hat{\beta}_k \sim N(\beta_k, \sigma_{\hat{\beta}_k}^2)$

Ước lượng phương sai của sai số

Như trong trường hợp hồi qui đơn, ước lượng phương sai sai số dựa vào các phần dư bình phương tối thiểu

$$s_\varepsilon^2 = \frac{\sum e_i^2}{n-K}$$

Thực tế là: $E[s_\varepsilon^2] = \sigma_\varepsilon^2$

Nếu các sai số ngẫu nhiên tuân theo phân phối chuẩn thì chúng ta cũng có

$$\frac{(n-K)s_\varepsilon^2}{\sigma_\varepsilon^2} \sim \chi_{(n-K)}^2$$

Nếu chúng ta viết các sai số chuẩn của các ước lượng hệ số là $s.e.(\hat{\beta}_k) = s_{\hat{\beta}_k} = \hat{\sigma}_{\hat{\beta}_k}$ (Anh/Chị sẽ thấy ký hiệu khác nhau trong các tài liệu khác nhau), thì chúng ta có thể viết

$$\frac{\hat{\beta}_k - \beta_k}{s_{\hat{\beta}_k}} = t\text{-stat} \sim t_{(n-K)}$$

Với hiểu biết về phân phối chọn mẫu của trị thống kê t , chúng ta có khả năng xây dựng các khoảng tin cậy và kiểm định giả thiết như trước đây.

Kiểm định mức độ ý nghĩa chung của mô hình

Trong mô hình hồi qui đơn, giả thiết rằng mô hình của chúng ta không có sức mạnh giải thích, điều này tương đương với giả thiết là tham số độ dốc bằng không.

Trong mô hình hồi qui bội, giả thiết “không” cho rằng mô hình không có sức mạnh giải thích được hiểu là tất cả các hệ số hồi qui riêng (các tham số độ dốc) đều bằng không:

$$H_0 : \beta_2 = \beta_3 = \dots = \beta_K = 0$$

H_1 : Không phải tất cả các tham số đồng thời bằng không

Trị thống kê kiểm định đối với giả thiết này là :

$$F = \frac{ESS/(K-1)}{RSS/(n-K)} \sim F_{(K-1, n-K)}$$

EViews sẽ tính trị thống kê này và giá trị p của nó ở góc phải-phía dưới của bảng kết quả hồi qui. Nếu mô hình có sức giải thích kém, thì chúng ta kỳ vọng rằng ESS sẽ nhỏ và RSS sẽ lớn. Do đó, nguyên tắc ra quyết định của chúng ta là phải bác bỏ giả thuyết “không” nếu F^* lớn (hoặc là nếu giá trị p nhỏ hơn mức ý nghĩa).

Hệ số xác định và các tiêu chuẩn lựa chọn mô hình

Hệ số xác định R^2 được xác định giống như trước đây :

$$R^2 = 1 - \frac{RSS}{TSS}$$

Khi giá trị R^2 lớn hơn cho chúng ta biết mô hình hồi qui “tốt hơn”, nhưng chúng ta cần cảnh giác về việc ý nghĩa “tốt hơn” này đạt được ra sao và nên nhớ rằng mỗi mô hình hồi qui có nhiều thuộc tính cần được xem xét đồng thời khi đánh giá chất lượng của nó. Sẽ là sai lầm khi đánh giá một mô hình chỉ trên cơ sở giá trị hệ số xác định R^2 .

Việc bổ sung thêm các biến hồi qui vào một mô hình hồi qui bội không thể làm giảm giá trị R^2 , cho dù là những biến hồi qui này không phù hợp, vì thế thường có vài nỗ lực gia tăng các biến hồi qui vào mô hình. Tuy nhiên, chúng ta sẽ học được cách tiếp cận sau nữa là sự gia tăng của R^2 sẽ chịu đánh đổi bằng sự giảm độ chính xác của những ước lượng.

Các nhà nghiên cứu nên nhớ rằng việc bổ sung thêm một biến hồi qui cũng làm tăng thêm một hệ số ước lượng, điều này tăng thêm "công việc" mà dữ liệu phải làm. Nói cách khác, với một lượng thông tin đã cho chúng ta phải phân phối chúng cho số lượng hệ số lớn hơn.

Một cách nhằm kết hợp sự đánh đổi giữa cái được tiềm năng của thông tin từ một biến hồi qui tăng thêm và chi phí của việc ước lượng hệ số cho biến đó là việc sử dụng một loạt "tiêu chuẩn lựa chọn mô hình" khác nhau. Nhiều tiêu chuẩn trong số này được mô tả dưới đây. Một nét đặc trưng chuẩn mực của những tiêu chuẩn lựa chọn mô hình này là mỗi tiêu chuẩn đều cân đối giữa sự gia tăng sức mạnh giải thích được đóng góp bởi một biến hồi qui bổ sung với sự giảm mức chính xác khi sử dụng thông tin để ước lượng hệ số bổ sung này.

R^2 điều chỉnh

$$\bar{R}^2 = 1 - \frac{RSS/(n-K)}{TSS/(n-1)} = 1 - (1-R^2) \frac{n-1}{n-K}$$

Nghiên cứu biểu thức này để thấy điều gì xảy ra với \bar{R}^2 khi chúng ta bổ sung thêm một biến hồi qui và ESS không cải thiện.

Sách Ramanathan, in lần thứ năm, trang 152 liệt kê 8 tiêu chuẩn khác để lựa chọn mô hình. Các tiêu chuẩn này có thể hiện khác nhau và các nhà nghiên cứu khác nhau có thể lựa chọn các tiêu chuẩn khác nhau phù hợp với các ứng dụng cụ thể.

Hai tiêu chuẩn phổ biến mà EViews cho chúng ta biết là Tiêu chuẩn Thông tin Akaike (AIC) và Tiêu chuẩn Schwarz:

$$AIC = \left(\frac{RSS}{n} \right) e^{(2K/n)}$$

$$Schwarz = \left(\frac{RSS}{n} \right) n^{(K/n)}$$

Khi sử dụng những tiêu chuẩn này để so sánh các mô hình khác nhau, mô hình nào có giá trị những tiêu chuẩn này thấp hơn sẽ được ưu tiên hơn khi lựa chọn.

Trong các bài tập ứng dụng mà chúng ta làm, Anh/Chị cần lưu ý là R^2 , \bar{R}^2 , và các tiêu chuẩn AIC và Schwarz khác nhau như thế nào.