Xây dựng mô hình hồi quy tuyến tính bội

Bởi:

Phạm Trí Cao

Xây dựng mô hình

Giới thiệu

Mô hình hồi quy hai biến mà chúng ta đã nghiên cứu ở chương 3 thường không đủ khả năng giải thích hành vi của biến phụ thuộc. Ở chương 3 chúng ta nói tiêu dùng phụ thuộc vào thu nhập khả dụng, tuy nhiên có nhiều yếu tố khác cũng tác động lên tiêu dùng, ví dụ độ tuổi, mức độ lạc quan vào nền kinh tế, nghề nghiệp… Vì thế chúng ta cần bổ sung thêm biến giải thích(biến độc lập) vào mô hình hồi quy. Mô hình với một biến phụ thuộc với hai hoặc nhiều biến độc lập được gọi là hồi quy bội.

Chúng ta chỉ xem xét hồi quy tuyến tính bội với mô hình tuyến tính với trong tham số, không nhất thiết tuyến tính trong biến số.

Mô hình hồi quy bội cho tổng thể

$$Y_i = \beta_1 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \_ + \beta_k X_{k,i} + \epsilon_i$$

(4.1)

Với $X_{2,i}$, $X_{3,i}$,…,$X_{k,i}$ là giá trị các biến độc lập ứng với quan sát i

β_1? β_2? β_3?…? $\beta_k$là các tham số của hồi quy

$\beta_i$là sai số của hồi quy

Với một quan sát i, chúng ta xác định giá trị kỳ vọng của Yi

$$E[Y|X's] = \beta_1 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \_ + \beta_k X_{k,i}$$

(4.2)

Ý nghĩa của tham số

Các hệ số β được gọi là các hệ số hồi quy riêng

$$\frac{\partial[Y|X's]}{\partial X_m} = \beta_m$$

(4.3)

β_k đo lường tác động riêng phần của biến X_m lên Y với điều kiện các biến số khác trong mô hình không đổi. Cụ thể hơn nếu các biến khác trong mô hình không đổi, giá trị kỳ vọng của Y sẽ tăng β_m đơn vị nếu X_m tăng 1 đơn vị.

Giả định của mô hình

Sử dụng các giả định của mô hình hồi quy hai biến, chúng ta bổ sung thêm giả định sau:

Các biến độc lập của mô hình không có sự phụ thuộc tuyến tính hoàn hảo, nghĩa là không thể tìm được bộ số thực (β_1? β_2?...? β_k) sao cho

$$\lambda_1 + \lambda_2 X_{2i} + \lambda_3 X_{3i} + \_ + \lambda_k X_{ki} = 0$$

với mọi i.

Giả định này còn được được phát biểu là " không có sự đa cộng tuyến hoàn hảo trong mô hình".

Số quan sát n phải lớn hơn số tham số cần ước lượng k.

Biến độc lập X_i phải có sự biến thiên từ quan sát này qua quan sát khác hay $Var(X_i)>0$.

Ước lượng tham số của mô hình hồi quy bội

Hàm hồi quy mẫu và ước lượng tham số theo phương pháp bình phương tối thiểu

Trong thực tế chúng ta thường chỉ có dữ liệu từ mẫu. Từ số liệu mẫu chúng ta ước lượng hồi quy tổng thể.

Hàm hồi quy mẫu

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + ... + \hat{\beta}_k X_{ki} + e_i \quad (4.4)$$

$$e_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i} - ... - \hat{\beta}_k X_{ki}$$

Với các $\hat{\beta}_m$ là ước lượng của tham số β_m. Chúng ta trông đợi $\hat{\beta}_m$ là ước lượng không chệch của β_m, hơn nữa phải là một ước lượng hiệu quả. Với một số giả định chặt chẽ như ở mục 3.3.1 chương 3 và phần bổ sung ở 4.1, thì phương pháp tối thiểu tổng bình phương phần dư cho kết quả ước lượng hiệu quả β_m.

Phương pháp bình phương tối thiểu

Chọn β_1? β_2? …? β_k sao cho

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2,i} - \hat{\beta}_3 X_{3,i} - \_\_ - \hat{\beta}_k X_{k,i}\right)^2$$

(4.5)

đạt cực tiểu.

Điều kiện cực trị của (4.5)

$$\left.\begin{aligned}
&\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \beta_1} = -2\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2,i} - \hat{\beta}_3 X_{3,i} - \ldots - \hat{\beta}_K X_{K,i}\right) = 0 \\
&\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \beta_2} = -2\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2,i} - \hat{\beta}_3 X_{3,i} - \ldots - \hat{\beta}_K X_{K,i}\right) X_{2,i} = 0 \quad (4.6) \\
&\ldots \\
&\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \beta_k} = -2\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2,i} - \hat{\beta}_3 X_{3,i} - \ldots - \hat{\beta}_K X_{K,i}\right) X_{k,i} = 0
\end{aligned}\right\}$$

Hệ phương trình (4.6) được gọi là hệ phương trình chuẩn của hồi quy mẫu (4.4).

Cách giải hệ phương trình (4.4) gọn gàng nhất là dùng ma trận. Do giới hạn của chương trình, bài giảng này không trình bày thuật toán ma trận mà chỉ trình bày kết quả tính toán cho hồi quy bội đơn giản nhất là hồi quy ba biến với hai biến độc lập. Một số tính chất của hồi quy ta thấy được ở hồi quy hai biến độc lập có thể áp dụng cho hồi quy bội tổng quát.

Ước lượng tham số cho mô hình hồi quy ba biến

Hàm hồi quy tổng thể

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

(4.7)

Hàm hồi quy mẫu

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

(4.8)

Nhắc lại các giả định

Kỳ vọng của sai số hồi quy bằng 0:

$$E(e_i | X_{2i}, X_{3i}) = 0$$

Không tự tương quan:

$$cov(e_i, e_j) = 0$$

, i≠j

Phương sai đồng nhất:

$$var(e_i) = \sigma^2$$

Không có tương quan giữa sai số và từng X_m:

$$cov(e_i, X_{2i}) = cov(e_i, X_{3i}) = 0$$

Không có sự đa cộng tuyến hoàn hảo giữa X_2 và X_3.

Dạng hàm của mô hình được xác định một cách đúng đắn.

Với các giả định này, dùng phương pháp bình phương tối thiểu ta nhận được ước lượng các hệ số như sau.

$$\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2\overline{X}_2 - \hat{\beta}_3\overline{X}_3 \quad (4.10)$$

$$\hat{\beta}_2 = \frac{\left(\sum_{i=1}^{n} y_i x_{2,i}\right)\left(\sum_{i=1}^{n} x_{3,i}^2\right) - \left(\sum_{i=1}^{n} y_i x_{3,i}\right)\left(\sum_{i=1}^{n} x_{2,i} x_{3,i}\right)}{\left(\sum_{i=1}^{n} x_{2,i}^2\right)\left(\sum_{i=1}^{n} x_{3,i}^2\right) - \left(\sum_{i=1}^{n} x_{2,i} x_{3,i}\right)^2} \quad (4.11)$$

$$\hat{\beta}_3 = \frac{\left(\sum_{i=1}^{n} y_i x_{3,i}\right)\left(\sum_{i=1}^{n} x_{2,i}^2\right) - \left(\sum_{i=1}^{n} y_i x_{2,i}\right)\left(\sum_{i=1}^{n} x_{2,i} x_{3,i}\right)}{\left(\sum_{i=1}^{n} x_{2,i}^2\right)\left(\sum_{i=1}^{n} x_{3,i}^2\right) - \left(\sum_{i=1}^{n} x_{2,i} x_{3,i}\right)^2} \quad (4.12)$$

Phân phối của ước lượng tham số

Trong phần này chúng ta chỉ quan tâm đến phân phối của các hệ số ước lượng

$$\hat{\beta}_2$$

và

$$\hat{\beta}_3$$

. Hơn nữa vì sự tương tự trong công thức xác định các hệ số ước lượng nên chúng ta chỉ khảo sát

$$\hat{\beta}_2$$

. Ở đây chỉ trình bày kết quả
Các thao tác chứng minh khá phức tạp, để tự chứng minh độc giả hãy nhớ lại các định nghĩa và tính chất của giá trị kỳ vọng, phương sai và hiệp phương sai của biến ngẫu nhiên.
.

$$\hat{\beta}_2$$

là một ước lượng không chệch :

$$E\left(\hat{\beta}_2\right) = \beta_2$$

(4.13)

$$var\left(\hat{\beta}_2\right) = \frac{\sum_{i=1}^{n} x_{3,i}^2}{\left(\sum_{i=1}^{n} x_{2,i}^2\right)\left(\sum_{i=1}^{n} x_{3,i}^2\right) - \left(\sum_{i=1}^{n} x_{2,i} x_{3,i}\right)^2}\sigma^2$$

(4.14)

Nhắc lại hệ số tương quan giữa X_2 và X_3 :

$$r_{X_2X_3} = \frac{\sum_{i=1}^{n} x_{2i}x_{3i}}{\sqrt{\left(\sum_{i=1}^{n} x_{2i}^2\right)}\sqrt{\left(\sum_{i=1}^{n} x_{3i}^2\right)}}$$

Đặt $r_{X_2X_3}$= r23 biến đổi đại số (4.14) ta được

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

(4.15)

Từ các biểu thức (4.13) và (4.15) chúng ta có thể rút ra một số kết luận như sau:

Nếu X2 và X3 có tương quan tuyến tính hoàn hảo thì

$$r_{23}^2$$

=1. Hệ quả là

$$\text{var}(\hat{\beta}_2)$$

vô cùng lớn hay ta không thể xác định được hệ số của mô hình hồi quy.

Nếu X2 và X3 không tương quan tuyến tính hoàn hảo nhưng có tương quan tuyến tính cao thì ước lượng

vẫn không chệch nhưng không hiệu quả.

Những nhận định trên đúng cho cả hồi quy nhiều hơn ba biến.

R^2 và

hiệu chỉnh

Nhắc lại khái niệm về

:

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

Một mô hình có

lớn thì tổng bình phương sai số dự báo nhỏ hay nói cách khác độ phù hợp của mô hình đối với dữ liệu càng lớn. Tuy nhiên một tính chất đặc trưng quan trọng của là nó có xu hướng tăng khi số biến giải thích trong mô hình tăng lên. Nếu chỉ đơn thuần chọn tiêu chí là chọn mô hình có

cao, người ta có xu hướng đưa rất nhiều biến độc lập vào mô hình trong khi tác động riêng phần của các biến đưa vào đối với biến phụ thuộc không có ý nghĩa thống kê.

Để hiệu chỉnh phạt việc đưa thêm biến vào mô hình, người ra đưa ra trị thống kê

hiệu chỉnh(Adjusted R^2)
Công thức của Theil, được sử dụng ở đa số các phần mềm kinh tế lượng. Một công thức khác do Goldberger đề xuất là Modified

$$R^2 = \left(1 - \frac{k}{n}\right) R^2$$

. (Theo Gujarati, Basic Econometrics-3rd, trang 208).

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k}$$

(4.16)

Với n là số quan sát và k là số hệ số cần ước lượng trong mô hình.

Qua thao tác hiệu chỉnh này thì chỉ những biến thực sự làm tăng khả năng giải thích của mô hình mới xứng đáng được đưa vào mô hình.

Kiểm định mức ý nghĩa chung của mô hình

Trong hồi quy bội, mô hình được cho là không có sức mạnh giải thích khi toàn bộ các hệ số hồi quy riêng phần đều bằng không.

Giả thiết

H_0: $\beta_2 = \beta_3 = \cdots = \beta_k = 0$

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

Trị thống kê kiểm định H_0:

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

Quy tắc quyết định

Nếu $F_{tt} > F_{(k-1,n-k,\beta)}$ thì bác bỏ H_0.

Nếu $F_{tt} \leq F_{(k-1,n-k,\beta)}$ thì không thể bác bỏ H_0.

Quan hệ giữa R2 và F

$$F = \frac{ESS/(k-1)}{RSS/(n-k)} = \frac{(n-k)ESS}{(k-1)RSS} = \frac{(n-k)ESS}{(k-1)(TSS-ESS)}$$

$$= \frac{(n-k)ESS/TSS}{(k-1)(1-ESS/TSS)} = \frac{(n-k)R^2}{(k-1)(1-R^2)} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}$$

Ước lượng khoảng và kiểm định giả thiết thống kê cho hệ số hồi quy

Ước lượng phương sai của sai số

$$s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-k}$$

(4.17)

Người ta chứng minh được

là ước lượng không chệch của σ^2, hay

.

Nếu các sai số tuân theo phân phối chuẩn thì

$$\frac{(n-k)s_e^2}{\sigma^2} \sim \chi^2_{(n-k)}$$

.

Ký hiệu

$$s.e(\hat{\beta}_m) = s_{\hat{\beta}_m} = \hat{\sigma}_{\hat{\beta}_m}$$

. Ta có trị thống kê

$$\frac{\hat{\beta}_m - \beta_m}{s.e(\hat{\beta}_m)} \sim t_{(n-k)}$$

Ước lượng khoảng cho β_m với mức ý nghĩa β là

$$\hat{\beta}_m - t_{(n-k,1-\alpha/2)} s.e(\hat{\beta}_m) \le \beta_m \le \hat{\beta}_m + t_{(n-k,1-\alpha/2)} s.e(\hat{\beta}_m)$$

(4.18)

Thông thường chúng ta muốn kiểm định giả thiết H_0 là biến X_m không có tác động riêng phần lên Y.

$H_0 : \beta_m = 0$

$H_1 : \beta_m \neq 0$

Quy tắc quyết định

Nếu /t-stat/ > $t_{(n-k,\beta/2)}$ thì ta bác bỏ H_0.

Nếu /t-stat/$\le t_{(n-k,\beta/2)}$ thì ta không thể bác bỏ H_0.