

Định lý Gauss-Markov

Lần trước chúng ta đã tính các ước lượng bằng phương pháp bình phương tối thiểu cho tung độ gốc và độ dốc của mô hình hồi qui đơn. Chúng ta đã thấy rằng các ước lượng có phân phối chuẩn và không chệch, và chúng ta đã tìm ra phương sai cho ước lượng của hệ số độ dốc.

Ước lượng bình phương tối thiểu là tốt nhất? Theo Định lý Gauss-Markov :

Trong các ước lượng không chệch tuyến tính, các ước lượng bình phương tối thiểu là tốt nhất. Điều này có nghĩa là không tồn tại hàm ước lượng không chệch tuyến tính nào khác có phương sai nhỏ hơn ước lượng bình phương tối thiểu này.

Chúng ta có thể chứng minh điều này cho hệ số độ dốc như sau :

$$\hat{\beta}_2 = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}$$

Để tiện lợi về ký hiệu, hãy định nghĩa: $w_i = \frac{x_i}{\sum_{i=1}^n x_i^2}$.

Bây giờ chúng ta có thể viết $\hat{\beta}_2 = \sum_{i=1}^n w_i Y_i$

Chúng ta cũng có thể định nghĩa một ước lượng tuyến tính khác như sau :

$$\tilde{\beta}_2 = \sum_{i=1}^n (w_i + d_i) Y_i$$

Chúng ta có thể chọn tập hợp các số d_i bằng mọi cách mà chúng ta muốn để có được một ước lượng tuyến tính khác, nhưng chúng phải thỏa mãn một ràng buộc là ước lượng không chệch. Mục tiêu của chúng ta là nếu chúng ta có khả năng tìm ra một tập hợp những số như vậy mà chúng sẽ cho ra ước lượng không chệch với một phương sai nhỏ hơn so với phương sai của ước lượng theo phương pháp bình phương tối thiểu.

$$\tilde{\beta}_2 = \sum_{i=1}^n (w_i + d_i) Y_i = \hat{\beta}_2 + \sum_{i=1}^n d_i Y_i$$

Để giữ tính chất không chệch chúng ta cần

$$\begin{aligned} E[\tilde{\beta}_2] &= E[\hat{\beta}_2] + E\left[\sum d_i (\beta_1 + \beta_2 X_i \varepsilon_i)\right] \\ &= \beta_2 + \beta_1 \sum d_i + \beta_2 \sum d_i X_i + \sum d_i E[\varepsilon_i] \end{aligned}$$

Nói chung biểu thức kỳ vọng này bằng β_2 nếu có hai điều kiện xảy ra :

$$\sum d_i = 0 \quad X \quad \text{và} \quad \sum d_i X_i = 0$$

Tiếp theo chúng ta tính phương sai của ước lượng mới này :

$$\text{VAR}[\tilde{\beta}_2] = \text{VAR}\left[\sum_{i=1}^n (w_i + d_i) Y_i\right]$$

Vì các thành phần của tổng là độc lập thống kê, nên chúng ta có thể viết :

$$\begin{aligned} \text{VAR}[\tilde{\beta}_2] &= \sum_{i=1}^n \text{VAR}[(w_i + d_i) Y_i] = \sum (w_i + d_i)^2 \text{VAR}[Y_i] \\ &= \sigma^2 \sum w_i^2 + \sigma^2 \sum d_i^2 + 2 \times \sigma^2 \sum w_i d_i \end{aligned}$$

Bây giờ xét thành phần cuối cùng :

$$\sum \left(\frac{X_i - \bar{X}}{\sum x_i^2} \right) d_i = \frac{1}{\sum x_i^2} [\sum X_i d_i + \bar{X} \sum d_i] = 0$$

chúng ta có được điều này vì điều kiện ước lượng không chệch .

Bây giờ hãy xét thành phần thứ nhất:

$$\sum \left(\frac{x_i}{\sum x_i^2} \right)^2 = \frac{1}{(\sum x_i^2)^2} \sum x_i^2 = \frac{1}{\sum x_i^2}$$

Và chúng ta nhận ra rằng thành phần thứ nhất là phương sai của ước lượng theo phương pháp bình phương tối thiểu:

$$\text{VAR}[\hat{\beta}_2]$$

Với mục đích làm tối thiểu phương sai của ước lượng không chệch tuyến tính mới này, chúng ta phải chọn các phần tử d_i để cho thành phần thứ hai nhỏ tới mức tối đa. Do

thành phần thứ hai là tổng bình phương, nên cách tốt nhất làm cho nó nhỏ đi là cho $d_i = 0$ với mọi i .

Hàm ước lượng không chệch tuyến tính có phương sai tối thiểu trở thành ước lượng bình phương tối thiểu. Chúng ta thường thấy ước lượng bình phương tối thiểu này được gọi là **BLUE**: có nghĩa là ước lượng không chệch tuyến tính tốt nhất.

Ước lượng phương sai của sai số

Ước lượng của chúng ta về phương sai của sai số dựa trên tổng bình phương các phần dư (là đại lượng mà chúng ta tối thiểu hoá trong thủ tục bình phương tối thiểu):

$$s_{\varepsilon}^2 = \frac{\sum_{i=1}^n e_i^2}{n-2}$$

Ước lượng này là không chệch: $E[s_{\varepsilon}^2] = \sigma_{\varepsilon}^2$

Định lý :

$$\frac{(n-1)s_{\varepsilon}^2}{\sigma^2} \sim \chi^2_{(n-2)}$$

Chứng minh những biểu thức này thì dễ nếu chúng ta có thể sử dụng đại số ma trận, nhưng nếu bằng cách khác đi thì việc chứng minh sẽ trở nên nhàm chán, vì thế chúng ta sẽ không thực hiện các chứng minh về các biểu thức trên.

Thay thế ước lượng σ vào các biểu thức $\sigma_{\hat{\beta}_1}^2$ và $\sigma_{\hat{\beta}_2}^2$, ta thu được các ước lượng cho các phương sai của các hệ số tung độ gốc và độ dốc của chúng ta. Các căn bậc hai dương của những đại lượng này là các *sai số chuẩn* của các ước lượng này. Các sai số chuẩn này thường được gọi là

$$\text{s.e.}(\hat{\beta}_1) = s_{\hat{\beta}_1} = \hat{\sigma}_{\hat{\beta}_1}$$

$$\text{s.e.}(\hat{\beta}_2) = s_{\hat{\beta}_2} = \hat{\sigma}_{\hat{\beta}_2}$$

Anh/Chị có thể sử dụng bất cứ ký hiệu nào mà mình thích.

Các trị thống kê t

Làm việc với các định nghĩa của các ước lượng và các phương sai của chúng, chúng ta có thể dễ dàng có được:

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim t_{(n-2)}$$

$$\frac{\hat{\beta}_2 - \beta_2}{\hat{\sigma}_{\hat{\beta}_2}} \sim t_{(n-2)}$$

Thông tin này tạo điều kiện cho chúng ta đưa ra những khẳng định xác suất có thể làm cơ sở cho việc xây dựng các khoảng tin cậy và kiểm định giả thuyết.

Thảo luận về khoảng tin cậy :

Thảo luận về kiểm định giả thuyết :

Hệ số xác định: trị thống kê Goodness-of-fit

SRF là

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$

Thay định nghĩa của $\hat{\beta}_1$ vào biểu thức trên và biến đổi, ta có:

$$\hat{Y}_i - \bar{Y} = \hat{\beta}_2 (X_i - \bar{X}) \Rightarrow \hat{y}_i = \hat{\beta}_2 x_i$$

Chúng ta cũng có thể dùng định nghĩa phần dư sau đây:

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

Vì thế

$$Y_i - e_i - \bar{Y} = \hat{\beta}_2 (X_i - \bar{X}) \Rightarrow y_i = \hat{\beta}_2 x_i + e_i$$

Do đó

$$y_i = \hat{y}_i + e_i$$

Bây giờ xét biến thiên của Y_i :

$$\sum (Y_i - \bar{Y})^2 = \sum y_i^2 = \sum (\hat{y}_i + e_i)^2$$

$$\sum y_i^2 = \sum \hat{y}_i^2 = \sum e_i^2$$

Ba thành phần này thường được gọi là :

TSS = Tổng bình phương phần cần được giải thích

ESS = Tổng bình phương phần được giải thích

RSS = Tổng bình phương phần không được giải thích (phần dư)

Tỉ số giữa tổng biến thiên được giải thích bởi mô hình cho tổng bình phương cần được giải thích được gọi là hệ số xác định, hay là trị thống kê “good of fit” :

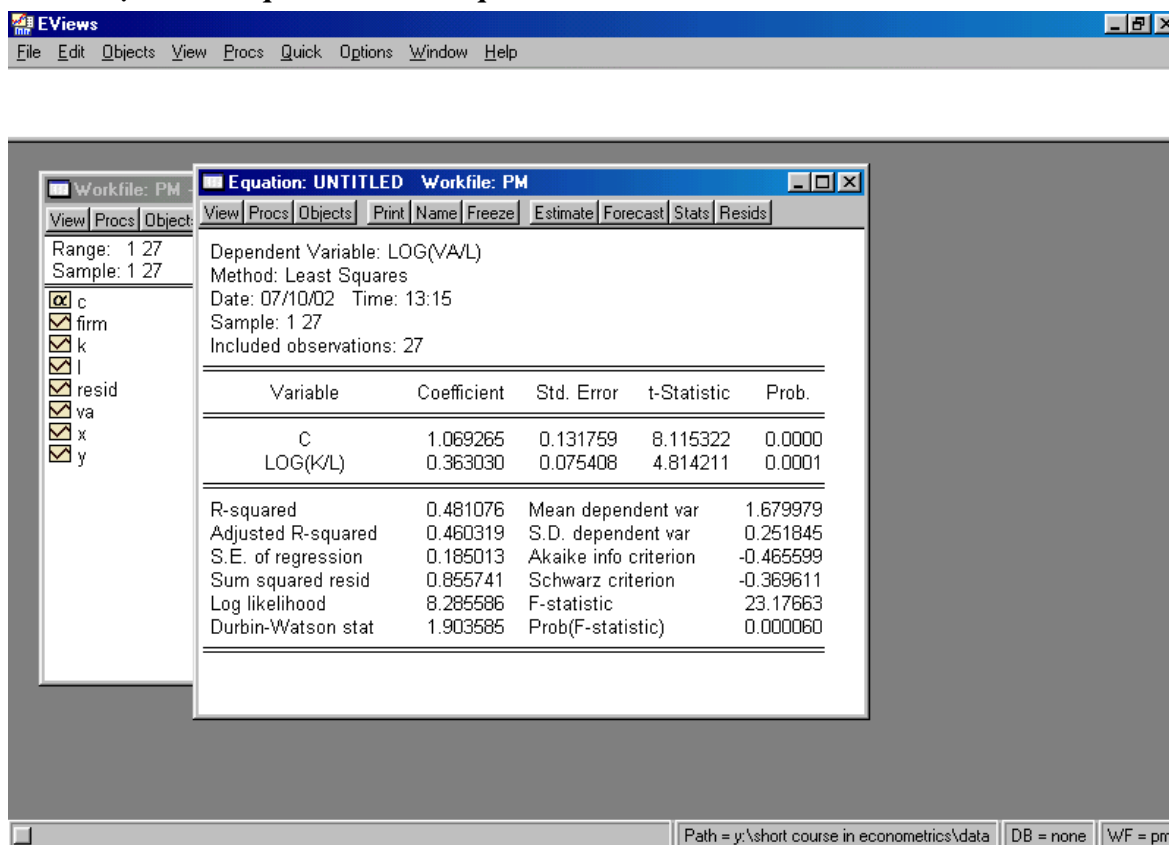
$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = \frac{\hat{\beta}_2^2 \sum x_i^2}{\sum y_i^2}$$

Chúng ta có thể thấy hệ số này đóng vai trò tương tự như hệ số tương quan.

Thực ra, đối với hồi qui đơn, hệ số xác định giống như hệ số tương quan bình phương. Nếu dữ liệu của chúng ta nằm chính xác trên SRF, chúng ta có một “good of fit” hoàn hảo. Với một “good-of fit” hoàn hảo, tất cả các phần dư đều bằng không và $R^2 = 1$.

Mặt khác, nếu mô hình của chúng ta không có sức mạnh giải thích ($\hat{\beta}_2 = 0$) thì $R^2 = 0$.

Thảo luận về kết quả LS cho hồi qui đơn



Đối với phương pháp hồi qui đơn này, biến phụ thuộc là logarit của sản lượng / lao động và biến giải thích là logarit của tỉ số vốn/lao động. Chúng ta có thể xem nó như là một mô hình giải thích năng suất lao động theo lượng vốn hiện trạng bị cho mỗi lao động.

Vì mô hình có dạng logarit kép nên hệ số độ dốc là hệ số co giãn của năng suất lao động theo tỉ lệ vốn/lao động.

Nếu (K / L) tăng lên 10% thì (VA / L) tăng lên 3,6%.

Một khoảng tin cậy 95% có thể được tính cho độ co giãn này, dựa vào thống kê t chúng ta có khẳng định xác suất sau đây:

$$P\left(t_{(0,025,n-2)} \leq \frac{\hat{\beta}_2 - \beta_2}{\hat{\sigma}_{\hat{\beta}_2}} \leq t_{(0,975,n-2)} \right) = 0,95$$

Nếu chúng ta biến đổi lại biểu thức này nhằm đưa tham số chưa biết vào giữa, thì chúng ta thu được khoảng tin cậy sau đây:

$$\hat{\beta}_2 \pm t_{(0,975,n-2)} \hat{\sigma}_{\hat{\beta}_2}$$

Sử dụng EViews chúng ta tìm được $t_{(0,975, n-2)} = 2,0595$

Khoảng tin cậy của chúng ta là

$$0,3630 \pm 2,0595 \times 0,0754 = [0,2077, 0,5183]$$

Chúng ta có thể kiểm định giả thiết “không” rằng tham số độ dốc bằng không với mức ý nghĩa 5% bằng cách ghi nhận rằng khoảng tin cậy này không chứa giá trị không.

Mặt khác, cách tiếp cận thông lệ sẽ là so sánh trị thống kê t nhận được với một giá trị tới hạn thích hợp hay so sánh giá trị của p với một mức ý nghĩa thích hợp.

Phát biểu giả thiết :

$$H_0 : \beta_2 = 0$$

$$H_1 : \beta_2 \neq 0$$

Chúng ta đọc được từ bảng kết quả là $t - Statistic_{\beta_2} = 4,8142$

Giá trị này lớn hơn giá trị tới hạn $t_{(0,975, n-2)} = 2,0595$ mà chúng ta tìm thấy trước đây, vì thế chúng ta bác bỏ giả thuyết “không” này.

Cuối cùng, chúng ta ghi nhận rằng giá trị của p bằng 0,0001. Giá trị này nhỏ hơn nhiều so với các mức ý nghĩa thông lệ, vì thế trên cơ sở này chúng ta sẽ dứt khoát bác bỏ giả thiết “không”.

Mặc dù hệ số góc có ý nghĩa thống kê mạnh, giá trị R^2 chỉ bằng 0,48. Có nghĩa là mô hình của chúng ta chỉ giải thích 48% biến thiên của biến logarit năng suất lao động.

Điều này cảnh báo cho chúng ta khả năng là các biến giải thích quan trọng đã bị bỏ sót và mô hình của chúng ta là chệch (vì vậy thực ra là vô ích). Chúng ta sẽ nghiên cứu chi tiết về bản chất của hiện tượng chệch do bỏ sót biến ngay sau bài giảng này.

Các trị thống kê khác mà Anh/Chị có thể nhận thấy là:

Sai số chuẩn hồi qui :
$$\hat{\sigma} = \sqrt{\frac{\sum e_i^2}{n-2}}$$

Tổng bình phương phần dư :
$$\sum e_i^2$$

Trung bình của biến phụ thuộc
$$\bar{Y}$$

$$\text{Độ lệch chuẩn của biến phụ thuộc } s_Y = \sqrt{\frac{\sum (Y_i - \bar{Y})^2}{n-1}}$$

Các trị thống kê mà Anh/Chị chưa biết sẽ được giới thiệu sau.

Chẩn đoán phần dư

Những tính chất của các ước lượng và các trị thống kê của chúng ta phụ thuộc vào các giả định cổ điển. Rất hữu ích khi ta có thể kiểm định những giả định này. Bây giờ chúng ta chỉ ra rằng những phần dư này phản ánh hành vi của các thành phần nhiễu ngẫu nhiên:

$$e_i = Y_i - \hat{Y}_i = \beta_1 + \beta_2 X_i + \varepsilon_i - \hat{\beta}_1 - \hat{\beta}_2 X_i$$

$$e_i = (\beta_1 - \hat{\beta}_1) + (\beta_2 - \hat{\beta}_2) X_i + \varepsilon_i$$

Thay $\hat{\beta}_1 = \beta_1 + \beta_2 \bar{X} - \hat{\beta}_2 \bar{X}$ và biến đổi đại số một chút để có

$$e_i = \varepsilon_i + \left(\frac{x_i \sum x_i \varepsilon_i}{\sum x_i^2} \right)$$

Trong tương lai chúng ta sẽ làm một số kiểm định chính thức về các phần dư. Hiện thời, Anh/Chị có thể chỉ cần nhấp vào nút Resids trong bảng kết quả LS để xem các phần dư biến đổi như thế nào. Anh/Chị cũng có thể khảo sát các trị thống kê mô tả và biểu đồ tần suất của vectơ phần dư trong workfile để thấy có phải chúng dường như tuân theo phân phối chuẩn.