

Dự đoán bằng mô hình hồi qui đơn

Xét mô hình năng suất lao động mà chúng ta đã nghiên cứu:

$$\log\left(\frac{VA_i}{L_i}\right) = \beta_1 + \beta_2 \log\left(\frac{K_i}{L_i}\right) + \varepsilon_i$$

Trong tập dữ liệu của chúng ta về công nghiệp kim loại sơ chế, biến giải thích nằm trong khoảng 0,55 đến 2,46. Hãy tưởng tượng là chúng ta quan tâm đến việc dự đoán giá trị của $\log\left(\frac{VA_i}{L_i}\right)$ với $\log\left(\frac{K_i}{L_i}\right) = 5.0$.

Để việc trình bày được thuận tiện, chúng ta tạo các định nghĩa sau:

$$Y_i = \log\left(\frac{VA_i}{L_i}\right)$$

$$X_i = \log\left(\frac{K_i}{L_i}\right)$$

Thường thì $i = 1, \dots, n$ và trong trường hợp này $n = 27$.

Chúng ta đặt một giá trị nằm “ngoài mẫu” là $X_0 = 5.0$.

Cách tính giá trị dự báo của giá trị nằm ngoài mẫu của biến hồi qui hoàn toàn giống với cách tính các giá trị ước lượng của biến phụ thuộc:

$$\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_0$$

Các tính chất của giá trị dự báo (và của các giá trị ước lượng)

Dự báo bằng phương pháp bình phương tối thiểu (LS) là một dự báo không chệch cho trung bình tổng thể:

$$E[\hat{Y}_0] = E[\hat{\beta}_1 + \hat{\beta}_2 X_0] = \beta_1 + \beta_2 X_0 = E[\hat{Y}_0 | X_0]$$

bởi vì các ước lượng LS là các ước lượng tham số không chệch.

Hiển nhiên là dưới giả định sai số ngẫu nhiên tuân theo phân phối chuẩn, thì các giá trị ước lượng và giá trị dự báo cũng tuân theo phân phối chuẩn vì các tham số ước lượng cũng có phân phối chuẩn.

Sau cùng, chúng ta khảo sát các phương sai của các giá trị ước lượng và giá trị dự báo. Các phương sai này phụ thuộc vào việc chúng ta có dự định ước lượng trung bình tổng thể hay giá trị thực Y_i riêng biệt

Ta có

$$\text{VAR}[\hat{Y}_0] = \text{VAR}[\hat{\beta}_1 + \hat{\beta}_2 X_0] = \text{VAR}[\hat{\beta}_1] + X_0^2 \times \text{VAR}[\hat{\beta}_2] + 2X_0 \text{COV}[\hat{\beta}_1, \hat{\beta}_2]$$

Sử dụng công thức thích hợp và thực hiện một vài động tác rút gọn chúng ta có:

$$\text{VAR}[\hat{Y}_0] = \sigma^2 \left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2} \right]$$

Lưu ý là chúng ta có thể nâng cao mức chính xác của dự báo bằng cách chọn cỡ mẫu lớn. Cũng cần lưu ý là độ chính xác sẽ giảm khi giá trị nằm ngoài mẫu của biến độc lập nằm xa giá trị trung bình mẫu.

Nếu chúng ta quan tâm việc dự báo kết quả Y_i , thì sai số dự báo của chúng ta là

$$Y_0 - \hat{Y}_0 = \beta_1 + \beta_2 X_0 + \varepsilon_0 - \hat{\beta}_1 - \hat{\beta}_2 X_0 = (\beta_1 - \hat{\beta}_1) + X_0 (\beta_2 - \hat{\beta}_2) + \varepsilon_0$$

Bây giờ chúng ta thấy là phần sai số ngẫu nhiên có thêm một nguồn sai số nữa. Nếu chúng ta tính phương sai bằng cách bình phương hai vế và tính kỳ vọng, chúng ta được biểu thức sau:

$$\text{VAR}[\hat{Y}_0] = \sigma^2 \left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2} \right]$$

Khoảng tin cậy

Chúng ta có thể muốn tính các khoảng tin cậy cho các giá trị ước lượng hoặc các giá trị dự báo. Cách thực hiện chúng khá rõ ràng dựa vào các thông tin chúng ta đã có.

Định nghĩa Sai Số Chuẩn của Dự báo là:

$$s_{\hat{Y}_0} = \left(s^2 \left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2} \right] \right)^{1/2}$$

Kế đến lưu ý là

$$t_{\hat{Y}_0} = \frac{Y_0 - \hat{Y}_0}{s_{\hat{Y}_0}} \text{ tuân theo phân phối } t \text{ với bậc tự do là } (n-2)$$

Chúng ta tìm được khoảng tin cậy là

$$\hat{Y}_0 \pm t_{(1-\alpha/2, n-2)} S_{\hat{Y}_0}$$

Thông thường trong thực tế để tính khoảng tin cậy cho mỗi giá trị dự báo và giá trị ước lượng, rồi sau đó vẽ các đường cong nối các điểm này lại. Chúng ta sẽ được *dải tin cậy* cho SRF.

Bài tập 10 hướng dẫn bạn qui trình cần thiết để tạo đồ thị sau:

