Simple Linear Regression

October 25, 2023

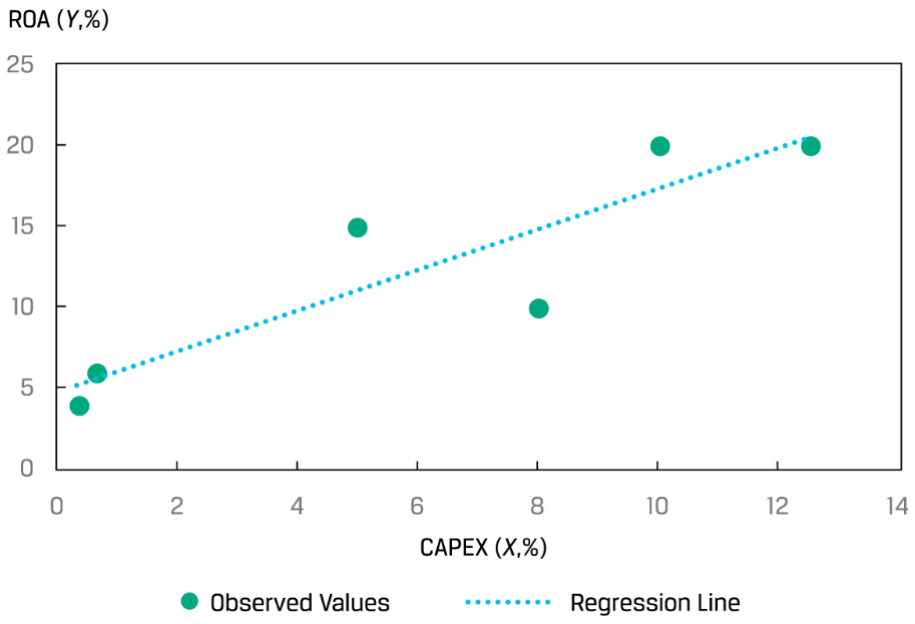
[ ]:
```
from IPython.display import Image
```

1 Estimation of The Simple Linear Regression Model

Hồi quy tuyến tính giả định mối quan hệ tuyến tính giữa biến phụ thuộc và biến độc lập. Mục đích ở đây là khớp một đường thẳng với các quan sát (X, Y) để tối thiểu hóa độ lệch bình phương. Đây chính là tiêu chí bình phương tối thiểu của hồi quy OLS

[ ]:
```
Image(filename = "Pictures/01.png")
```

[ ]:



Phương trình hồi quy tổng thể của hồi quy đơn OLS được trình bày như sau:

$$Y = b_0 + b_1 X + \varepsilon$$

Chúng ta không thể quan sát được các giá trị thực b_0 và b_1 trong phương trình tổng thể. Thay vào đó, chúng ta quan sát các giá trị ước lượng $\hat{b_0}$ và $\hat{b_1}$, các giá trị được ước lượng từ mẫu quan sát.

Điều kiện cần được thỏa mãn khi ước lượng các hệ số này chính là tối thiểu hóa tổng sai số bình phương (SSE)

$$SSE = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} \left(Y_i - \left(\hat{b_0} + \hat{b_1} X_i\right)\right)^2$$

Đối với hồi quy đơn, ước lượng OLS của các hệ số b_0 và b_1 được tính bởi công thức:

$$\hat{b_1} = \frac{Cov(X,Y)}{Var(X)} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

$$\hat{b_0} = \overline{Y} - \hat{b_1}\overline{X}$$

2 Assumptions of the Simple Linear Regression Model

Giả sử rằng chúng ta có n quan sát của biến phụ thuộc Y và biến độc lập X và chúng ta muốn ước tính hồi quy tuyến tính đơn giản của Y hồi quy trên X. Bốn giả định chính dưới đây là cần thiết để có thể rút ra kết luận hợp lệ từ mô hình hồi quy tuyến tính đơn giản:

1. Tuyến tính (`Linearity`): Mối quan hệ giữa Y và X là tuyến tính
2. Phương sai không đổi (`Homoskedasticity`): Phương sai của phần dư hồi quy không đổi với mọi quan sát $Var(e|X) = \sigma^2$
3. Độc lập trung bình (`Independence`): Các cặp Ys và cặp Xs độc lập với bất kỳ cặp nào khác (tức là không tồn tại tương quan chéo, tương quan chuỗi…). Điều này hàm ý phần dư hồi quy không có tương quan với các quan sát: $E(e|X) = 0$
4. Phân phối chuẩn (`Normality`): Phần dư hồi quy có phân phối chuẩn

3 Hypothesis Tests in the Simple Linear Regression Model

3.1 Analysis Variance

Mô hình hồi quy đơn thi thoảng mô tả được mối quan hệ giữa hai biến khá tốt, tuy nhiên trong nhiều trường hợp thì không. Chúng ta cần phân biệt rõ ràng giữa hai kịch bản này để sử dụng hồi quy một cách hiệu quả. Nên nhớ rằng mục tiêu cuối cùng là giải thích sự biến động của biến phụ thuộc

3.2 Breaking Down the Sum of Squares Total into Its Components

Trong hồi quy đơn, chúng ta thường quan tâm đến 3 chỉ tiêu tổng bình phương:

1. Tổng bình phương toàn phần (TSS)

$$TSS = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2$$

2. Tổng bình phương ước lượng (RSS)

$$RSS = \sum_{i=1}^{n} \left(\hat{Y}_i - \overline{Y}\right)^2$$

3. Tổng bình phương sai số (ESS)

$$ESS = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

Trong thực tế, $TSS = RSS + ESS$

3.3 Goodness-of-fit

Phân tích viên có thể sử dụng một số cách để đánh giá mức độ phù hợp của mô hình, nghĩa là mô hình hồi quy phù hợp với dữ liệu. Các thước đo thường gặp bao gồm hệ số xác định (R^2 - `coefficient of determination`), kiểm định F sự phù hợp của mô hình và sai số chuẩn hồi quy

3.3.1 Coefficient of Determination

Hệ số xác định R^2 đo lường tỷ lệ phần trăm biến động của biến phụ thuộc được giải thích từ các biến độc lập

$$R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n} \left(\hat{Y}_i - \overline{Y}\right)^2}{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2}$$

Trong một hồi quy đơn, hệ số xác định chính là bình phương của tương quan Pearson. Phần chứng minh được trình bày bên dưới:

$$R^2 = \frac{\sum_{i=1}^{n} \left(\hat{Y}_i - \overline{Y}\right)^2}{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2} = \frac{\sum_{i=1}^{n} \left(\hat{b_0} + \hat{b_1} X_i - \overline{Y}\right)^2}{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2} = \frac{\sum_{i=1}^{n} \left(\overline{Y} - \hat{b_1}\overline{X} + \hat{b_1} X_i - \overline{Y}\right)^2}{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2} = \frac{\hat{b_1}^2 \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2}$$

$$= \frac{\left(\frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}\right)^2 \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2} = \frac{\left(\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)\right)^2}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2} = r^2$$

3.3.2 F-Test

Ở đây, chúng ta sử dụng kiểm định F để kiểm tra xem liệu tất cả các hệ số góc của hồi quy có đồng thời bằng 0 hay không. Giả thuyết không của kiểm định này là: $b_i = 0 \ \forall i \neq 0$

Đối với hồi quy tuyến tính tổng quát với k biến độc lập trên một mẫu có n quan sát, giá trị kiểm định F được tính bởi công thức:

$$F(k, n-k-1) = \frac{R^2/k}{(1-R^2)/(n-k-1)} = \frac{MSR}{MSE}$$

Trong đó:

- MSR là trung bình bình phương hồi quy, bằng RSS/k trong trường hợp tổng quát
- MSE là trung bình bình phương sai số, bằng $TSS/(n-k-1)$ trong trường hợp tổng quát

Trong hồi quy đơn, giá trị kiểm định F sẽ là:

$$F(1, n-2) = \frac{R^2}{(1-R^2)/(n-2)}$$

3.3.3 Hypothesis Testing of Individual Regression Coefficients

Đôi khi, chúng ta cần thực hiện các kiểm định liên quan đến hệ số tổng thể, chẳng hạn như dấu của hệ số hồi quy, hệ số tổng thể có bằng một giá trị cụ thể nào đó hay không... Khi đó, chúng ta có thể sử dụng kiểm định t để kiểm tra các giả thuyết này. Tại đây, chúng ta thảo luận về 3 tình huống thường gặp:

1. Kiểm định liên quan đến hệ số góc: Giá trị kiểm định t cũng như sai số chuẩn mẫu của hệ số góc được xác định bởi các công thức sau:

$$t_{n-2} = \frac{\hat{b_1} - B_1}{s_{\hat{b_1}}}, \; s_{\hat{b_1}} = \frac{s_e}{TSS_X}$$

2. Kiểm định liên quan đến hệ số chặn: Giá trị kiểm định t cũng như sai số chuẩn mẫu của hệ số góc được xác định bởi các công thức sau:

$$t_{n-2} = \frac{\hat{b_0} - B_0}{s_{\hat{b_0}}}, \; s_{\hat{b_0}} = s_e \sqrt{\frac{n^{-1}\overline{X}^2}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}}$$

3. Kiểm định liên quan đến hệ số tương quan: Với giả thuyết không là hệ số tương quan bằng 0, giá trị kiểm định t với bậc tự do $n-2$ được tính bằng công thức:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

4 Prediction in the Simple Linear Regression Model

4.1 ANOVA and Standard Error of Estimate in Simple Linear Regression

Bảng ANOVA (`analyst of variance`) là một dạng mô tả thường gặp khi chúng ta tóm tắt các giá trị tổng bình phương trong mô hình hồi quy. Hình 2 mô tả cách xác định các giá trị trong một bảng ANOVA của hồi quy đơn

[ ]:
```
Image(filename = "Pictures/02.png")
```

[ ]:

Source	Sum of Squares	Degrees of Freedom	Mean Square	*F*-Statistic
Regression	$\text{SSR} = \sum_{i=1}^{n} \left(\widehat{Y}_i - \overline{Y}\right)^2$	1	$\text{MSR} = \frac{\sum_{i=1}^{n} \left(\widehat{Y}_i - \overline{Y}\right)^2}{1}$	$F = \frac{\text{MSR}}{\text{MSE}} = \frac{\frac{\sum_{i=1}^{n} \left(\widehat{Y}_i - \overline{Y}\right)^2}{1}}{\frac{\sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2}{n-2}}$
Error	$\text{SSE} = \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2$	$n - 2$	$\text{MSE} = \frac{\sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2}{n-2}$	
Total	$\text{SST} = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2$	$n - 1$		

Từ bảng ANOVA, chúng ta có thể xác định sai số chuẩn của ước lượng bằng cách lấy căn bậc hai của MSE

$$s_e = \sqrt{MSE} = \sqrt{\frac{ESS}{n-k-1}} = \sqrt{\frac{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2}{n-k-1}}$$

4.2 Prediction Using Simple Linear Regression and Prediction Intervals

Giả định rằng chúng ta đã xác định được mô hình hồi quy mẫu

$$\hat{Y} = \hat{b_0} + \hat{b_1} X$$

Một trong những lý do chúng ta xác định mô hình này là để thực hiện dự báo cho biến phụ thuộc. Cụ thể hơn, với mô hình hồi quy mẫu này, giá trị ước lượng của Y tại X_f là Y_f và được xác định từ hàm hồi quy ở trên:

$$\hat{Y_f} = \hat{b_0} + \hat{b_1} X_f$$

Tuy nhiên, chúng ta cần nhận thấy rằng giá trị ước tính này không phải là dự báo hoàn hảo, nó chỉ là giá trị được xác định dựa vào mối-quan-hệ-trung-bình-giữa-hai-biến. Do đó, chúng ta cần xác định ước lượng khoảng để mô tả sự không chắc chắn này. Độ lệch chuẩn của dự báo s_f được sử dụng trong nhiệm vụ này, và được xác định bằng công thức:

$$s_f = s_e \sqrt{1 + \frac{1}{n} + \frac{\left(X_f - \overline{X}\right)^2}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}} = s_e \sqrt{1 + \frac{1}{n} + \frac{\left(X_f - \overline{X}\right)^2}{\sum_{i=1}^{n} (n-}$$

5 Functional Forms for Simple Linear Regression

Một số dạng hàm khác nhau có thể được sử dụng để chuyển đổi dữ liệu nhằm cho phép sử dụng chúng trong hồi quy tuyến tính. Những phép biến đổi này bao gồm lấy logarit biến phụ thuộc, logarit biến độc lập, nghịch đảo biến độc lập, bình phương biến độc lập hoặc vi phân biến độc lập. Tại đây, chúng ta đề cập đến ba dạng hàm thường gặp:

1. Log-Lin model:

$$\log Y = b_0 + b_1 X + \varepsilon$$

2. Lin-Log model:

$$Y = b_0 + b_1 \log X + \varepsilon$$

3. Log-Log model:

$$\log Y = b_0 + b_1 \log X + \varepsilon$$