Hồi quy tuyến tính đơn

Hoàng Văn Hà
University of Science, VNU - HCM
hvha@hcmus.edu.vn

Mục lục

Phân tích hồi quy

Bài toán: trong các hoạt động về khoa học - kỹ thuật, y học, kinh tế - xã hội, . . . ta có nhu cầu xác định mối liên giữa hai hay nhiều biến ngẫu nhiên với nhau.
Ví dụ:

- Mối liên hệ giữa chiều cao và cỡ giày của một người, từ đó một cửa hàng bán giầy dép có thể xác định chính xác cỡ giầy của một khách hàng khi biết chiều cao,
- Độ giãn nở của một loại vật liệu theo nhiệt độ môi trường,
- Hàm lượng thuốc gây mê và thời gian ngủ của bệnh nhân,
- Doanh thu khi bán 1 loại sản phẩm và số tiền chi cho quảng cáo và khuyến mãi,
- . . .

Để giải quyết các vấn đề trên, ta sử dụng kỹ thuật phân tích hồi quy (regression analysis).

Phân tích hồi quy

- Phân tích hồi quy được sử dụng để xác định mối liên hệ giữa:
  - một biến phụ thuộc Y (biến đáp ứng), và
  - một hay nhiều biến độc lập X_1, $X_2, \ldots, X_p$. Các biến này còn được gọi là biến giải thích.
    - Biến phụ thuộc Y là biến liên tục,
    - Các biến độc lập $X_1, X_2, \ldots, X_p$ có thể là biến liên tục, rời rạc hoặc phân loại.

Phân tích hồi quy

- Mối liên hệ giữa $X_1, \ldots, X_p$ và Y được biểu diễn bởi một hàm tuyến tính.
- Sự thay đổi trong Y được giả sử do những thay đổi trong $X_1, \ldots, X_p$ gây ra.

- Trên cơ sở xác định mối liên hệ giữa biến phụ thuộc Y và các biến giải thích X_1, X_2, $\ldots$, X_p, ta có thể:
  - dự đoán, dự báo giá trị của Y,
  - giải thích tác động của sự thay đổi trong các biến giải thích lên biến phụ thuộc.

Mô hình hồi quy tuyến tính đơn

Định nghĩa

Một mô hình thống kê tuyến tính đơn (Simple linear regression model) liên quan đến một biến ngẫu nhiên Y và một biến giải thích x là phương trình có dạng

$$Y = \beta_0 + \beta_1 X + \epsilon, \tag{1}$$

trong đó

- β_0, β_1 *là các tham số chưa biết, gọi là các hệ số hồi quy,*
- X *là biến độc lập, giải thích cho* Y*,*
- ϵ *là thành phần sai số,* ϵ *được giả sử có phân phối chuẩn với* $\mathbb{E}(\epsilon) = 0$ *và* $\mathbb{V}ar(\epsilon) = \sigma^2$*.*

Mô hình hồi quy tuyến tính đơn

Trong mô hình (1), sự thay đổi của Y được giả sử ảnh hưởng bởi 2 yếu tố:

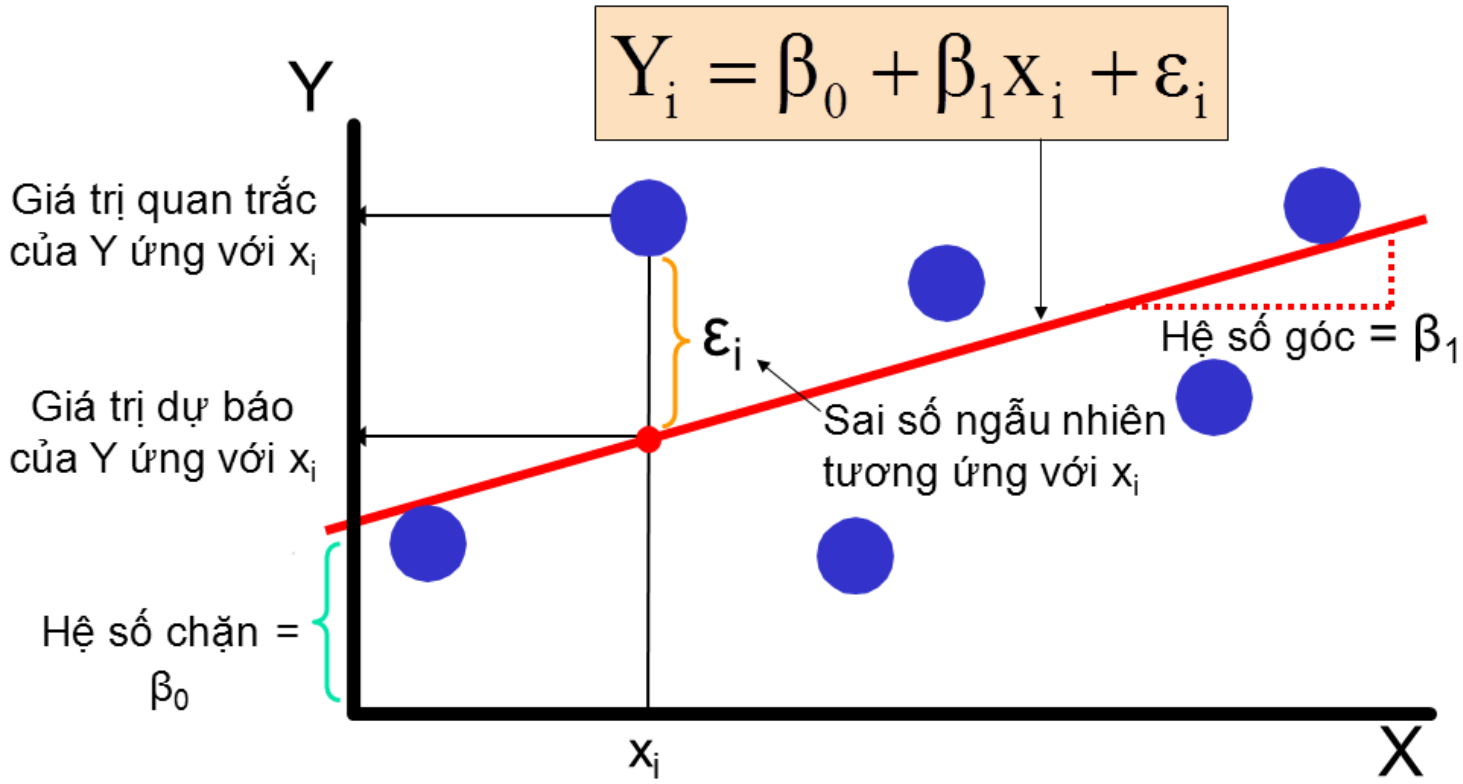
- Mối liên hệ tuyến tính của X và Y: $\beta_0 + \beta_1 x$. Trong đó, β_0 được gọi là hệ số chặn (intercept) và β_1 gọi là hệ số góc (slope).
- Tác động của các yếu tố khác (không phải X): thành phần sai số ϵ.

• Với $(x_1, y_1), \ldots, (x_n, y_n)$ là n cặp giá trị quan trắc của một mẫu ngẫu nhiên cỡ n, từ (1) ta có

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \ldots, n. \tag{2}$$

Mô hình hồi quy tuyến tính đơn

• Sử dụng đồ thị phân tán (scatter plot) để biểu diễn các cặp giá trị quan trắc (x_i, y_i) trên hệ trục tọa độ Oxy.



Các giả định về sai số ngẫu nhiên

- Các sai số ngẫu nhiên $\epsilon_i, i = 1, \ldots, n$ trong mô hình (2) được giả sử thỏa các điều kiện sau
  - Các sai số ϵ_i độc lập với nhau,
  - $\mathbb{E}(\epsilon_i) = 0$ và $\mathbb{V}ar(\epsilon_i) = \sigma^2$,
  - Các sai số có phân phối chuẩn: $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ với phương sai không đổi.
- Với quan trắc x đã biết,

$$\mathbb{E}(Y|X = x) = \beta_0 + \beta_1 x. \tag{3}$$

- Từ (3) ta có

$$Y|x \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2). \tag{4}$$

Ước lượng các hệ số hồi quy

- Gọi $\hat{\beta}_1$ và $\hat{\beta}_0$ là các ước lượng của β_0 và β_1.
- Đường thẳng "ướm" (fitted regression line):

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x. \tag{5}$$

- Một đường thẳng ước lượng tốt phải "gần với các điểm dữ liệu".
- Tìm $\hat{\beta}_0$ và $\hat{\beta}_1$: dùng phương pháp bình phương bé nhất (method of least squares).

Phương pháp bình phương bé nhất

- Với dữ liệu $(x_i, y_i), i = 1, \ldots, n$, từ (5) ta có

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{6}$$

- Độ sai khác giữa giá trị quan trắc y_i và giá trị dự đoán $\hat{y}_i$ gọi là thặng dư (residual) thứ i, xác định như sau

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i). \tag{7}$$

Phương pháp bình phương bé nhất

Định nghĩa

Tổng bình phương sai số (Sum of Squares for Errors - SSE) hay tổng bình phương thặng dư cho n điểm dữ liệu được định nghĩa như sau

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right]^2. \tag{8}$$

Nội dung của PPBPBN là tìm các ước lượng $\hat{\beta}_0$ và $\hat{\beta}_1$ sao cho *SSE* đạt giá trị bé nhất.

Phương pháp bình phương bé nhất

Từ (8), lấy đạo hàm theo β_0 và β_1,

$$\frac{\partial SSE}{\partial \beta_0} = -2\sum_{i=1}^{n}\left[y_i - (\beta_0 + \beta_1 x_i)\right] = 0,$$

$$\frac{\partial SSE}{\partial \beta_1} = -2\sum_{i=1}^{n}\left[y_i - (\beta_0 + \beta_1 x_i)\right] x_i = 0,$$

ta thu được hệ phương trình

$$\begin{aligned} n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i, \\ \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i y_i. \end{aligned} \tag{9}$$

Ước lượng bình phương bé nhất

Giả hệ (9), ta tìm được các ước lượng BPBN của β_0 và β_1 là

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n}}{\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}} = \frac{S_{xy}}{S_{xx}}, \tag{10}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \tag{11}$$

với S_{xx} và S_{xy} xác định bởi

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}, \tag{12}$$

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n}. \tag{13}$$

Ước lượng bình phương bé nhất

- Các ước lượng $\hat{\beta}_0$ và $\hat{\beta}_1$ tìm được gọi là các ước lượng BPBN.
- Đường thẳng $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ gọi là đường thẳng BPBN, thỏa các tính chất sau:

(1)

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

đạt giá trị bé nhất,

(2)

$$\text{SE} = \sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} e_i = 0.$$

với SE là tổng các thặng dư (Sum of Errors).

Ví dụ

Ví dụ

Một nhà thực vật học khảo sát mối liên hệ giữa tổng diện tích bề mặt (đv: cm^2) của các lá cây đậu nành và trọng lượng khô (đv: g) của các cây này. Nhà thực vật học trồng 13 cây trong nhà kính và đo tổng diện tích lá và trọng lượng của các cây này sau 16 ngày trồng, kết quả cho bởi bảng sau

X	411	550	471	393	427	431	492	371	470	419	407	489	439
Y	2.00	2.46	2.11	1.89	2.05	2.30	2.46	2.06	2.25	2.07	2.17	2.32	2.12

(a) *Vẽ biểu đồ phân tán biểu diễn diện tích lá X và trọng lượng khô Y của cây đậu nành với mẫu quan sát đã cho.*

(b) *Tìm đường thẳng hồi quy biểu diễn mối liên hệ giữa trọng lượng cây Y theo diện tích lá X. Vẽ đường thẳng hồi quy tìm được trên đồ thị phân tán.*

Độ đo sự biến thiên của dữ liệu

Gọi

- SST: Tổng bình phương toàn phần (Total Sum of Squares)

$$\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

- SSR: Tổng bình phương hồi quy (Regression Sum of Squares)

$$\text{SSR} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2.$$

- *SSE*: Tổng bình phương sai số (Error Sum of Squares)

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Độ đo sự biến thiên của dữ liệu

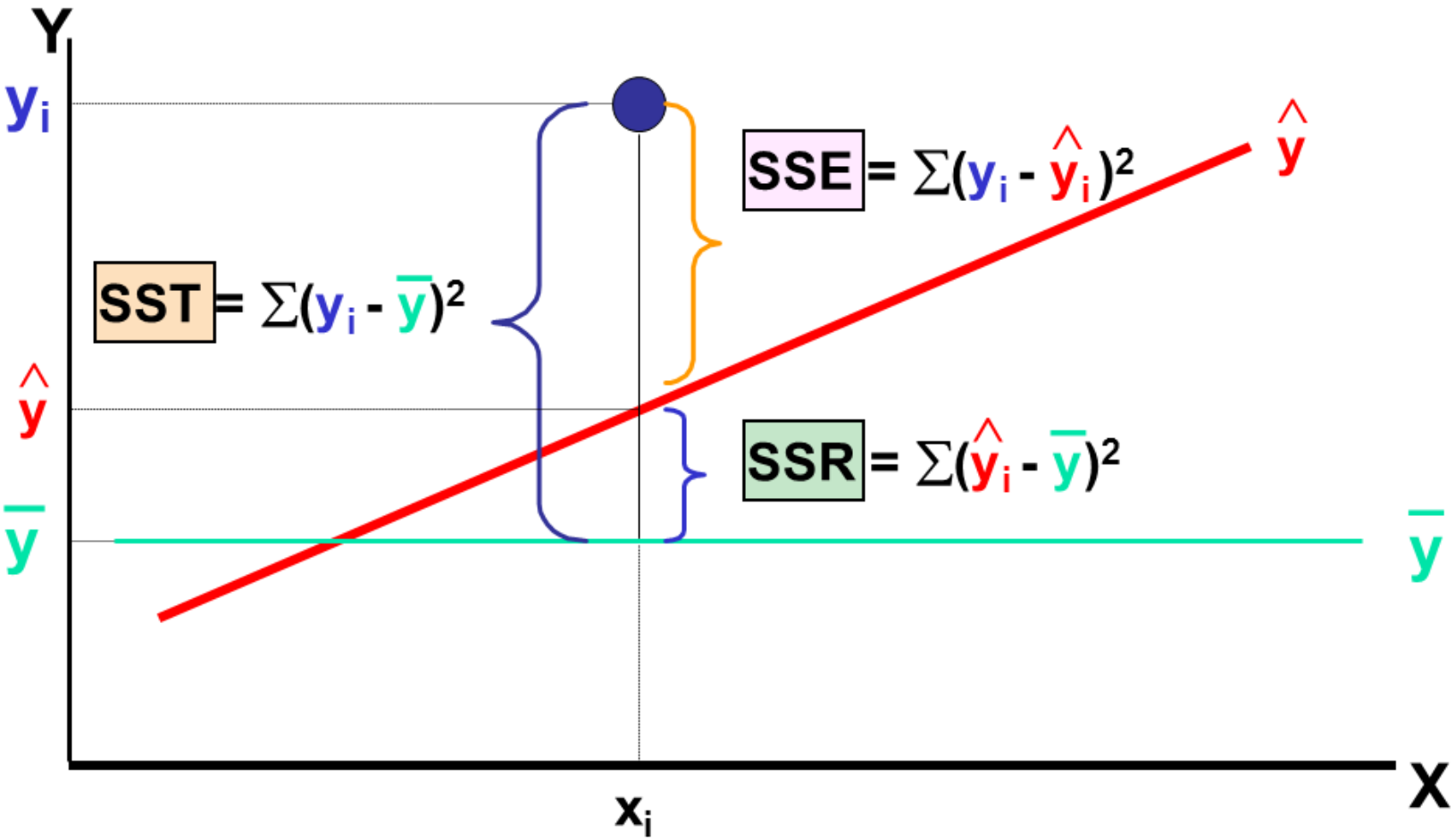
- SST: đo sự biến thiên của các giá trị y_i xung quanh giá trị trung tâm của dữ liệu $\bar{y}$,
- SSR: giải thích sự biến thiên liên quan đến mối quan hệ tuyến tính của X và Y,
- SSE: giải thích sự biến thiên của các yếu tố khác (không liên quan đến mối quan hệ tuyến tính của X và Y).

Ta có:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{14}$$

$$\text{SST} = \text{SSR} + SSE.$$

Độ đo sự biến thiên của dữ liệu



Hệ số xác định

Định nghĩa

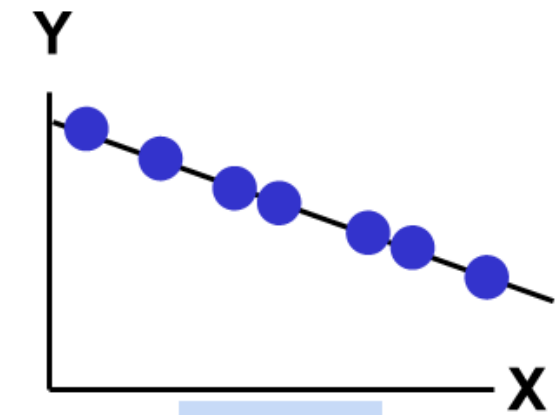
Hệ số xác định (Coefficient of Determination) là tỷ lệ của tổng sự biến thiên trong biến phụ thuộc gây ra bởi sự biến thiên của các biến độc lập (biến giải thích) so với tổng sự biến thiên toàn phần.
Hệ số xác định thường được gọi là R - bình phương (R-squared), ký hiệu là R^2.
Công thức tính:

$$R^2 = \frac{\text{SSR}}{\text{SST}}. \tag{15}$$

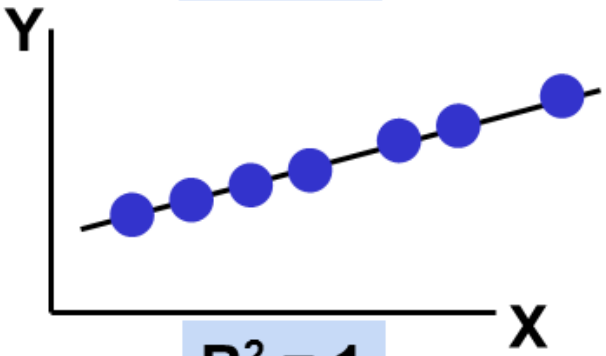
Chú ý: $0 \leq R^2 \leq 1$.

- Hệ số xác định của một mô hình hồi quy cho phép ta đánh giá mô hình tìm được có giải thích tốt cho mối liên hệ giữa biến phụ thuộc Y và biến phụ thuộc X hay không.

Hệ số xác định và mối liên hệ giữa X và Y



$R^2 = 1$



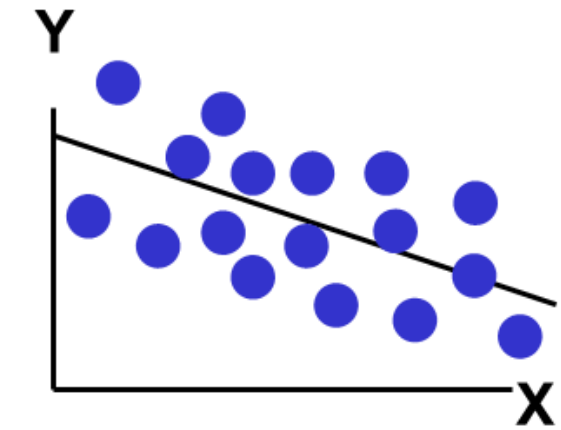
$R^2 = 1$

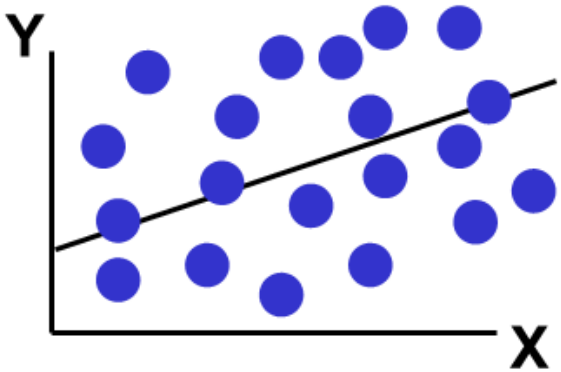
$R^2 = 1$

X và Y có mối liên hệ tuyến tính mạnh:

100% sự biến thiên của Y được giải thích bởi sự biến thiên của X

Hệ số xác định và mối liên hệ giữa X và Y



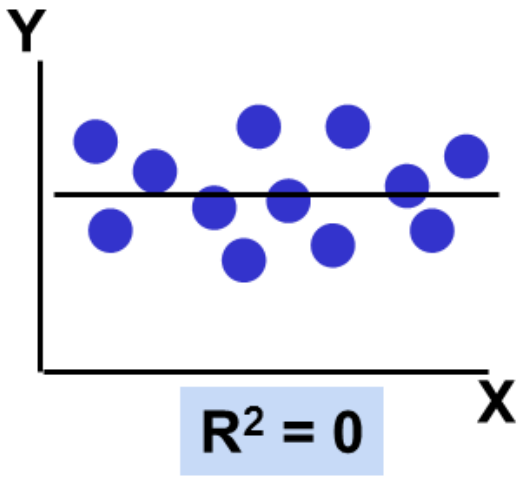


$0 < R^2 < 1$

X và Y có mối liên hệ tuyến tính yếu:

Một vài nhưng không phải tất cả sự biến thiên trong Y được giải thích bởi sự biến thiên trong X

Hệ số xác định và mối liên hệ giữa X và Y



$R^2 = 0$

Không có mối liên hệ tuyến tính giữa X và Y:

Giá trị của Y không phụ thuộc vào X. (Không có sự biến thiên nào của Y được giải thích bởi sự biến thiên của X)

Ước lượng phương sai σ^2 của sai số

Xét mô hình

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n$$

Thành phần sai số thứ i: $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Ta cần ước lượng phương sai σ^2.
Từ (4), ta có: $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$. Do đó,

$$\frac{Y_i - (\beta_0 + \beta_1 x_i)}{\sigma} \sim \mathcal{N}(0, 1).$$

Ta có,

$$\sum_{i=1}^{n} \frac{\left[Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right]^2}{\sigma^2} = \frac{SSE}{\sigma^2} \sim \chi^2(n-2).$$

Nên,

$$\mathbb{E}\left[\frac{SSE}{\sigma^2}\right] = n - 2 \quad \text{hay} \quad \mathbb{E}\left[\frac{SSE}{n-2}\right] = \sigma^2.$$

Ước lượng phương sai σ^2 của sai số

Ta kết luận rằng $\frac{SSE}{n-2}$ là một ước lượng không chệch cho σ^2. Suy ra ước lượng $\hat{\sigma}^2$ của σ^2 được tính bởi

$$\hat{\sigma}^2 = \frac{SSE}{n-2}. \tag{16}$$

Định nghĩa

Trung bình bình phương sai số (Mean Squares Error - MSE) của mô hình hồi quy tính tính đơn được định nghĩa bởi

$$\text{MSE} = \frac{SSE}{n-2}.$$

Nói cách khác, trung bình bình phương sai số chính là ước lượng không chệch cho phương sai của thành phần sai số của mô hình.

Ước lượng phương sai σ^2 của sai số

- Tìm SSE:

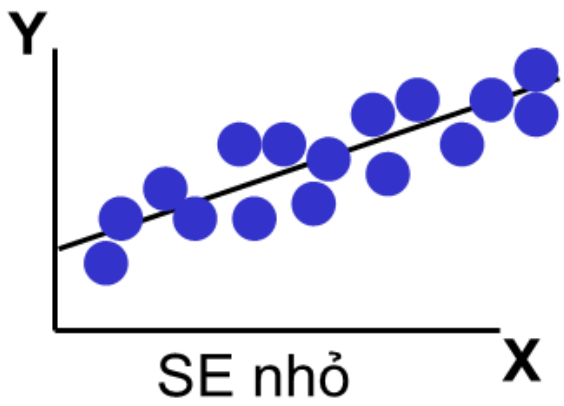
$$SSE = \text{SST} - \hat{\beta}_1 S_{xy}.$$

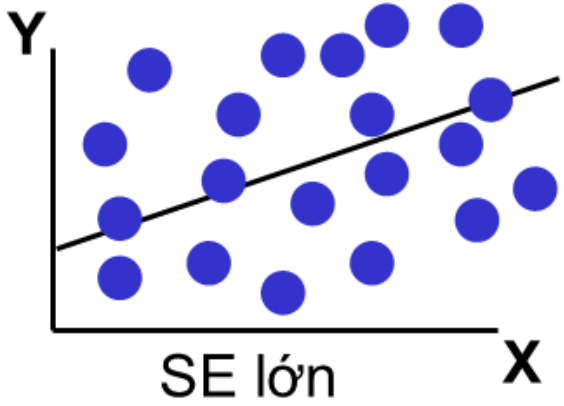
- Sai số chuẩn (Standard Error) của $\hat{\sigma}^2$

$$\text{SE}(\hat{\sigma}) = \sqrt{\frac{SSE}{n-2}}.$$

Sử dụng $\text{SE}(\hat{\sigma})$ để đo sự biến thiên của các giá trị quan trắc y với đường thẳng hồi quy.

So sánh sai số chuẩn





Tính chất của các ước lượng BPBN

Định lý

Xét $Y = \beta_0 + \beta_1 x + \epsilon$ *là một mô hình hồi quy tuyến tính đơn với* $\epsilon \sim \mathcal{N}(0, \sigma^2)$*; với* n *quan trắc độc lập* $y_i, i = 1, \ldots, n$ *ta có tương ứng các sai số* ϵ_i*. Gọi* $\hat{\beta}_0$ *và* $\hat{\beta}_1$ *là các ước lượng của* β_0 *và* β_1 *tìm được từ phương pháp bình phương bé nhất, khi đó*

(a) $\hat{\beta}_0$ *và* $\hat{\beta}_1$ *tuân theo luật phân phối chuẩn.*

(b) *Kỳ vọng và phương sai của* $\hat{\beta}_0$ *và* $\hat{\beta}_1$ *lần lượt là*

$$\mathbb{E}(\hat{\beta}_0) = \beta_0, \ \mathbb{V}ar(\hat{\beta}_0) = \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\sigma^2, \tag{17}$$

$$\mathbb{E}(\hat{\beta}_1) = \beta_1, \ \mathbb{V}ar(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}. \tag{18}$$

Tính chất của các ước lượng BPBN

Định nghĩa

Trong mô hình hồi quy tuyến tính đơn, sai số chuẩn (SE) của các ước lượng $\hat{\beta}_0$ và $\hat{\beta}_1$ là

$$\text{SE}(\hat{\beta}_0) = \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\hat{\sigma}^2}, \tag{19}$$

$$\text{SE}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}. \tag{20}$$

Định lý (Gauss - Markov)

Xét mô hình hồi quy tuyến tính đơn

$$Y = \beta_0 + \beta_1 x + \epsilon,$$

có $\hat{\beta}_0$ và $\hat{\beta}_1$ là các ước lượng BPBN cho β_0 và β_1, khi đó $\hat{\beta}_0$ và $\hat{\beta}_1$ là các ước lượng không chệch tốt nhất.

Khoảng tin cậy cho hệ số hồi quy

- Xét đường thẳng hồi quy ước lượng:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

- Vì $\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$, đặt

$$Z_1 = \frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} \text{ thì } Z_1 \sim \mathcal{N}(0,1).$$

- Do $\frac{SSE}{\sigma^2}$ độc lập với $\hat{\beta}_1$ và $\frac{SSE}{\sigma^2} \sim \chi^2(n-2)$ nên

$$T_{\beta_1} = \frac{Z_1}{\sqrt{\frac{\left(\frac{SSE}{\sigma^2}\right)}{n-2}}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\left(\frac{SSE}{n-2}\right)/S_{xx}}} = \frac{\hat{\beta}_1 - \beta_1}{\text{SE}(\hat{\beta}_1)} \sim t(n-2). \quad (21)$$

Khoảng tin cậy cho hệ số hồi quy

- Tương tự, vì $\hat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\sigma^2\right)$, đặt

$$Z_0 = \frac{\hat{\beta}_0 - \beta_0}{\sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} \sim \mathcal{N}(0,1).$$

- Do $\hat{\beta}_0$ và SSE/σ^2 độc lập và $SSE/\sigma^2 \sim \chi^2(n-2)$ nên ta có

$$T_{\beta_0} = \frac{Z_0}{\sqrt{\frac{\left(\frac{SSE}{\sigma^2}\right)}{n-2}}} = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\left(\frac{SSE}{n-2}\right)\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} = \frac{\hat{\beta}_0 - \beta_0}{\mathrm{SE}(\hat{\beta}_0)} \sim t(n-2). \qquad (22)$$

Khoảng tin cậy cho hệ số hồi quy

- Khoảng tin cậy $100(1-\alpha)\%$ cho β_1:

$$\hat{\beta}_1 - t^{n-2}_{1-\alpha/2}\,\mathrm{SE}(\hat{\beta}_1) \le \beta_1 \le \hat{\beta}_1 + t^{n-2}_{1-\alpha/2}\,\mathrm{SE}(\hat{\beta}_1). \tag{23}$$

- Khoảng tin cậy $100(1-\alpha)\%$ cho β_0:

$$\hat{\beta}_0 - t^{n-2}_{1-\alpha/2}\,\mathrm{SE}(\hat{\beta}_0) \le \beta_0 \le \hat{\beta}_0 + t^{n-2}_{1-\alpha/2}\,\mathrm{SE}(\hat{\beta}_0). \tag{24}$$

với

- $n =$ số cặp giá trị quan trắc (x_i, y_i).
- $t^{n-2}_{1-\alpha/2}$ là phân vị mức $1-\alpha/2$ của biến ngẫu nhiên $t(n-2)$.
- $\mathrm{SE}(\hat{\beta}_0)$ và $\mathrm{SE}(\hat{\beta}_1)$ cho bởi phương trình (19) và (20).

Khoảng tin cậy cho trung bình biến đáp ứng

• Cho trước giá trị x_0, cần tìm khoảng tin cậy cho $\mu_{Y|x_0} = \mathbb{E}(Y|x_0) = \beta_0 + \beta_1 x_0$, gọi là trung bình biến đáp ứng. Ước lượng của $\mu_{Y|x_0}$ từ đường thẳng hồi quy là

$$\hat{\mu}_{Y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

• $\hat{\mu}_{Y|x_0}$ có các tính chất sau

(1) $\hat{\mu}_{Y|x_0}$ tuân theo luật phân phối chuẩn.

(2) Kỳ vọng và phương sai của $\hat{\mu}_{Y|x_0}$ lần lượt là

$$\mathbb{E}(\hat{\mu}_{Y|x_0}) = \beta_0 + \beta_1 x_0.$$

$$\mathbb{V}ar(\hat{\mu}_{Y|x_0}) = \left[\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}}\right]\sigma^2.$$

Khoảng tin cậy cho trung bình biến đáp ứng

- Ta có

$$\frac{\hat{\mu}_{Y|x_0} - \mathbb{E}(\hat{\mu}_{Y|x_0})}{\sqrt{\mathbb{V}ar(\hat{\mu}_{Y|x_0})}} = \frac{\hat{\mu}_{Y|x_0} - (\beta_0 + \beta_1 x_0)}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}}\right)}} \sim \mathcal{N}(0,1).$$

- Vì $\hat{\mu}_{Y|x_0}$ độc lập với $SSE/\sigma^2 \sim \chi^2(n-2)$ nên

$$\frac{\hat{\mu}_{Y|x_0} - (\beta_0 + \beta_1 x_0)}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}}\right)}\sqrt{\frac{SSE/\sigma^2}{n-2}}} = \frac{\hat{\mu}_{Y|x_0} - (\beta_0 + \beta_1 x_0)}{\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}}\right]}} \sim t(n-2). \qquad (25)$$

Khoảng tin cậy cho trung bình biến đáp ứng

- Sai số chuẩn của $\hat{\mu}_{Y|x_0}$ cho bởi

$$\text{SE}(\hat{\mu}_{Y|x_0}) = \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}}\right]}. \tag{26}$$

- Khoảng tin cậy $100(1-\alpha)\%$ cho trung bình biến đáp ứng là

$$\hat{\mu}_{Y|x_0} - t^{n-2}_{1-\alpha/2}\,\text{SE}(\hat{\mu}_{Y|x_0}) \le \mu_{Y|x_0} \le \hat{\mu}_{Y|x_0} + t^{n-2}_{1-\alpha/2}\,\text{SE}(\hat{\mu}_{Y|x_0}). \tag{27}$$

với

- $\hat{\mu}_{Y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$, và
- $t^{n-2}_{1-\alpha/2}$: phân vị mức $1-\alpha/2$ của biến ngẫu nhiên $t(n-2)$.

Dự đoán giá trị quan trắc mới

• Giả sử với giá trị x_0, ta cần dự đoán giá trị quan trắc Y_0 trong tương lai tương ứng với x_0 bằng bao nhiêu. Từ mô hình hồi quy, ta có

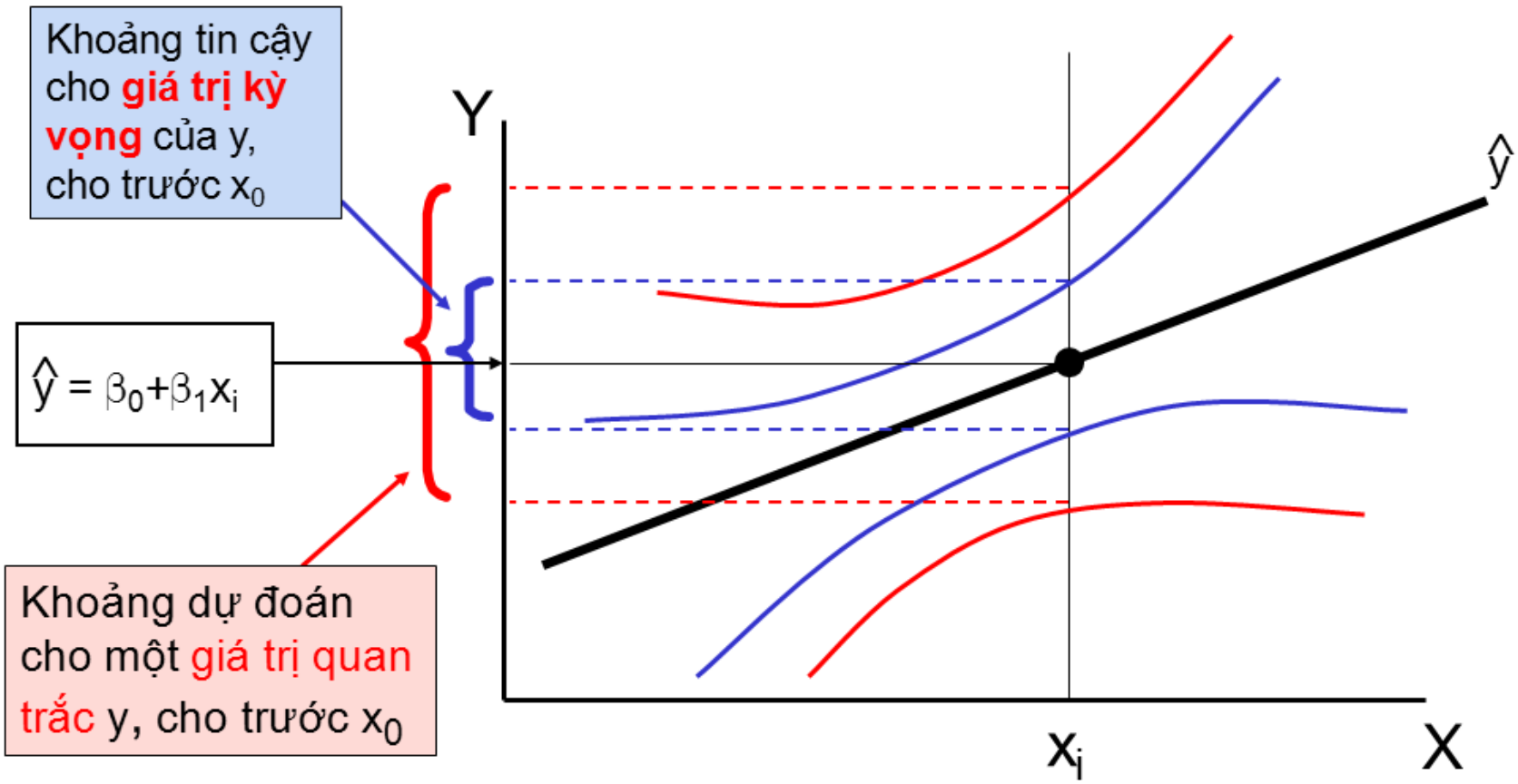
$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 \tag{28}$$

$\hat{Y}_0$ là một ước lượng điểm của giá trị quan trắc mới Y_0.

• Cần tìm khoảng tin cậy cho Y_0.

• Cho trước giá trị x_0, cần phân biệt rõ khoảng tin cậy giữa trung bình của biến ngẫu nhiên Y là $\mu_{Y|x_0}$ và khoảng tin cậy của giá trị quan trắc thực sự của Y tương ứng với x_0.

Dự đoán giá trị quan trắc mới



Dự đoán giá trị quan trắc mới

- Đặt

$$\eta = Y_0 - \hat{Y}_0$$

Vì Y_0 và $\hat{Y}_0$ có phân phối chuẩn nên η có phân phối chuẩn với kỳ vọng và phương sai là

$$\mathbb{E}(\eta) = \mathbb{E}(Y_0) - \mathbb{E}(\hat{Y}_0) = 0$$

$$\mathbb{V}ar(\eta) = \mathbb{V}ar(Y_0) + \mathbb{V}ar(\hat{Y}_0) = \sigma^2 + \sigma^2\left[\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}}\right]$$

$$= \left[1 + \frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}}\right]\sigma^2$$

Do đó,

$$\eta \sim \mathcal{N}\left(0, \left[1 + \frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}}\right]\sigma^2\right).$$

Dự đoán giá trị quan trắc mới

- Và,

$$Z = \frac{\hat{Y}_0 - Y_0}{\sqrt{\hat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}}\right]}} \sim \mathcal{N}(0, 1).$$

- Nếu ta thay thế σ^2 bởi $\hat{\sigma}^2 = \frac{SSE}{n-2}$, thu được

$$T = \frac{\hat{Y}_0 - Y_0}{\sqrt{\hat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}}\right]}} \sim t(n-2). \tag{29}$$

Dự đoán giá trị quan trắc mới

• Sai số chuẩn $\mathrm{SE}(\hat{Y}_0)$ cho bởi

$$\mathrm{SE}(\hat{Y}_0) = \sqrt{\hat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}}\right]}. \tag{30}$$

• Khoảng tin cậy $100(1-\alpha)\%$ cho giá trị dự báo mới Y_0 ứng với một giá trị x_0 cho trước là

$$\hat{Y}_0 - t^{n-2}_{1-\alpha/2}\,\mathrm{SE}(\hat{Y}_0) \le Y_0 \le \hat{Y}_0 + t^{n-2}_{1-\alpha/2}\,\mathrm{SE}(\hat{Y}_0). \tag{31}$$

với $t^{n-2}_{1-\alpha/2}$ là phân vị mức $1-\alpha/2$ của $t(n-2)$.

Ví dụ

Ví dụ

Xét mẫu ngẫu nhiên gồm 10 *cặp giá trị* (x_i, y_i) *cho bởi bảng*

x	-1	0	2	-2	5	6	8	11	12	-3
y	-5	-4	2	-7	6	9	13	21	20	-9

(a) *Vẽ biểu đồ phân tán cho dữ liệu, tìm đường thẳng hồi quy.*

(b) *Tìm ước lượng* $\hat{\sigma}^2$ *cho phương sai* σ^2 *của sai số ngẫu nhiên.*

(c) *Thiết lập khoảng tin cậy* 95% *cho các hệ số* β_0 *và* β_1.

(d) *Thiết lập khoảng dự đoán* 95% *tại* $x = 5$.

Kiểm định giả thuyết cho các hệ số hồi quy

Đặt vấn đề:

- Giả sử ta cần xây dựng một mô hình hồi quy với biến phụ thuộc Y và một tập các biến giải thích $X_1, X_2, \ldots, X_p$.
- Trong tập hợp các biến $X_1, X_2, \ldots, X_p$ này, có những biến giải thích tốt cho Y, cũng có thể có những biến không liên quan hoặc có mối liên hệ rất nhỏ với Y.
- Ta có thể xét mô hình hồi quy tuyến tính tổng quát (hồi quy bội):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon.$$

- Để xác định biến nào có ý nghĩa đối với mô hình, ta có thể thực hiện kiểm định giả thuyết đối với các hệ số hồi quy tương ứng, cụ thể,

$$H_0 : \beta_j = 0 \quad \text{với} \quad H_1 : \beta_j \neq 0,$$

  với $j = 0, \ldots, p$.
- Trong nội dung chương trình học, ta đang khảo sát mô hình hồi quy tuyến tính đơn $Y = \beta_0 + \beta_1 X + \epsilon$, nên ta sẽ xét bài toán kiểm định giả thuyết cho β_0 và β_1.

Kiểm định giả thuyết cho β_0

• Bài toán kiểm định giả thuyết cho hệ số chặn β_0 trong mô hình hồi quy tuyến tính đơn gồm các trường hợp sau:

$$(a)\begin{cases} H_0 : \beta_0 = b_0 \\ H_1 : \beta_0 \neq b_0 \end{cases} \qquad (b)\begin{cases} H_0 : \beta_0 = b_0 \\ H_1 : \beta_0 < b_0 \end{cases} \qquad (c)\begin{cases} H_0 : \beta_0 = b_0 \\ H_1 : \beta_0 > b_0 \end{cases}$$

với giá trị b_0 và mức ý nghĩa α cho trước. Trường hợp mặc định, $b_0 = 0$.

Kiểm định giả thuyết cho β_0

Các bước kiểm định

1. Phát biểu giả thuyết H_0 và đối thuyết H_1,
2. Xác định mức ý nghĩa α,
3. Tính giá trị thống kê kiểm định:

$$t_{\beta_0} = \frac{\hat{\beta}_0 - b_0}{\text{SE}(\hat{\beta}_0)}, \quad \text{với } \text{SE}(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2 \left(1 + \frac{\bar{x}^2}{S_{xx}}\right)}.$$

4. Bác bỏ H_0 khi: $|t_{\beta_0}| > t^{n-2}_{1-\alpha/2}$.
5. Kết luận: Bác bỏ H_0 / Chưa đủ cơ sở để bác bỏ H_0.
6. Hoặc ta có thể sử dụng p-giá trị tính bởi

$$p = 2\mathbb{P}(T_{n-2} \geq |t_{\beta_0}|),$$

và bác bỏ H_0 khi $p \leq \alpha$.

Kiểm định giả thuyết cho β_1

• Bài toán kiểm định giả thuyết cho hệ số góc β_1 trong mô hình hồi quy tuyến tính đơn gồm các trường hợp sau:

$$(a)\begin{cases} H_0 : \beta_1 = b_1 \\ H_1 : \beta_1 \neq b_1 \end{cases} \qquad (b)\begin{cases} H_0 : \beta_1 = b_1 \\ H_1 : \beta_1 < b_1 \end{cases} \qquad (c)\begin{cases} H_0 : \beta_1 = b_1 \\ H_1 : \beta_1 > b_1 \end{cases}$$

với giá trị b_1 và mức ý nghĩa α cho trước. Trường hợp mặc định, $b_1 = 0$.

Kiểm định giả thuyết cho β_1

Các bước kiểm định

1. Phát biểu giả thuyết H_0 và đối thuyết H_1,
2. Xác định mức ý nghĩa α,
3. Tính giá trị thống kê kiểm định:

$$t_{\beta_1} = \frac{\hat{\beta}_1 - b_1}{SE(\hat{\beta}_1)}, \quad \text{với } \mathrm{SE}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}.$$

4. Bác bỏ H_0 khi: $|t_{\beta_1}| > t^{n-2}_{1-\alpha/2}$.
5. Kết luận: Bác bỏ H_0 / Chưa đủ cơ sở để bác bỏ H_0.
6. Hoặc ta có thể sử dụng p-giá trị tính bởi

$$p = 2\mathbb{P}(T_{n-2} \geq |t_{\beta_1}|),$$

và bác bỏ H_0 khi $p \leq \alpha$.

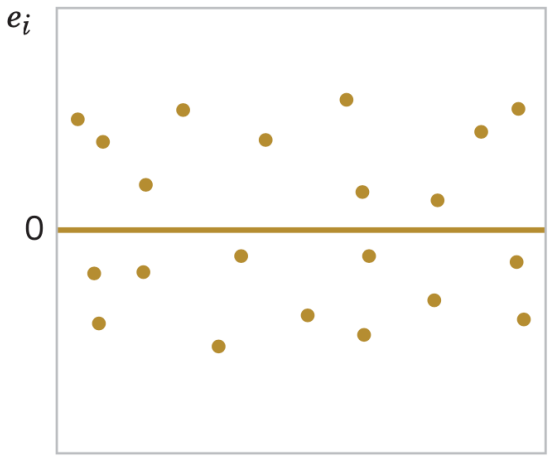
Phân tích thặng dư

- Phân tích thặng dư (Residual Analysis) được sử dụng để kiểm tra các giả định của mô hình hồi quy tuyến tính.
- Các giả định của mô hình:
  1. Tuyến tính: mối quan hệ giữa X và Y là tuyến tính tính, tức là
  $$\mathbb{E}[Y|X = x] = \beta_0 + \beta_1 x.$$
  2. Phương sai bằng nhau: phương sai của biến đáp ứng (biến phụ thuộc) Y là hằng số với mọi giá trị của biến độc lập X, tức là $\mathbb{V}\text{ar}(Y|X = x) = \sigma^2$.
  3. Độc lập: các quan trắc của biến đáp ứng Y độc lập với nhau.
  4. Phân phối chuẩn: với mỗi giá trị của biến độc lập, phân phối có điều kiện (cho trước giá trị x) của biến đáp ứng là phân phối chuẩn, $Y|X = x \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2)$.
- Việc kiểm tra các giả định trên thông thường sẽ được thực hiện thông qua các giá trị thặng dư, cho bởi
$$e_i = y_i - \hat{y}_i, \quad i = 1, \ldots, n,$$
với $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

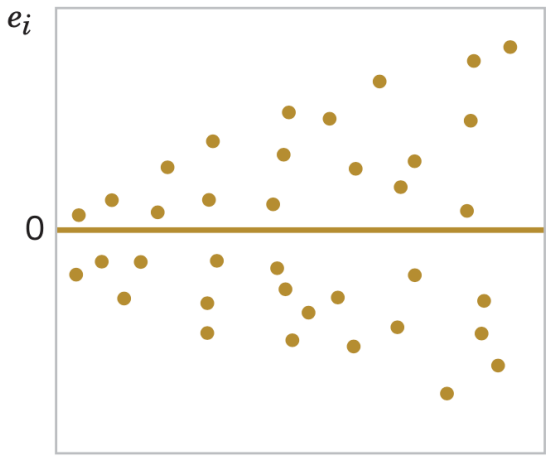
Phân tích thặng dư

- Đồ thị các giá trị thặng dư: các cặp $(\hat{y}_i, e_i)$, $i = 1, \ldots, n$. (Hoặc ta vẽ các giá trị e_i tương ứng với các giá trị của biến độc lập x_i).
- Nếu các giả định về 1, 2 và 3 thỏa thì ta sẽ nhận thấy đồ thị thặng dư gồm các điểm phân tán đều trên mặt phẳng Oxy và phân tán đề xung quanh đường thẳng $y = 0$.
- Trường hợp một trong các giả định trên bị vi phạm, chẳng hạn như phương sai thay đổi, mối quan hệ giữa các biến không tuyến tính, ta sẽ thấy các điểm trên đồ thị thặng dư sẽ phân bố theo một hình dạng cụ thể nào đó.
- Đồ thị thặng dư cũng giúp cho ta xác định được sự tồn tại của các điểm outlier.
- Để kiểm tra giả định về phân phối chuẩn (giả định 4), ta thường dùng đồ thị Normal Q-Q Plot.

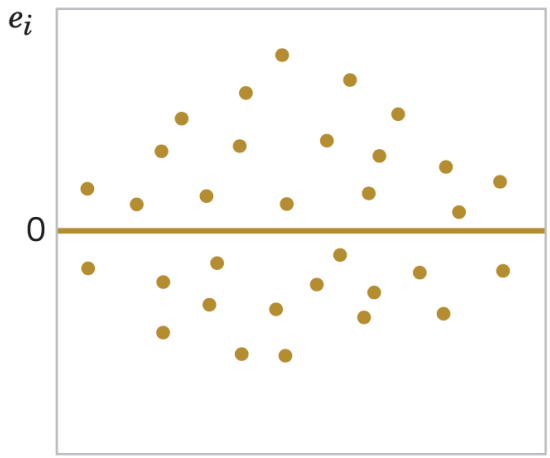
Phân tích thặng dư



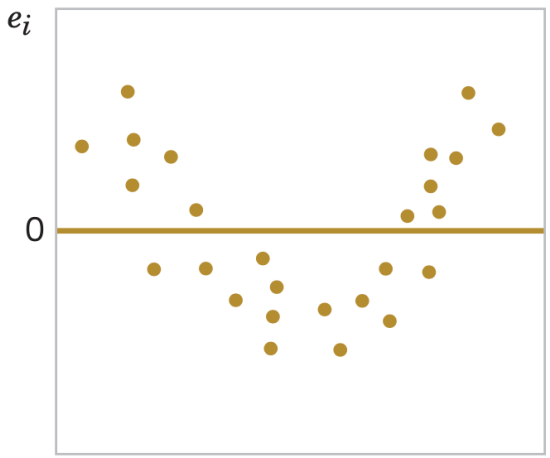
(a)



(b)



(c)

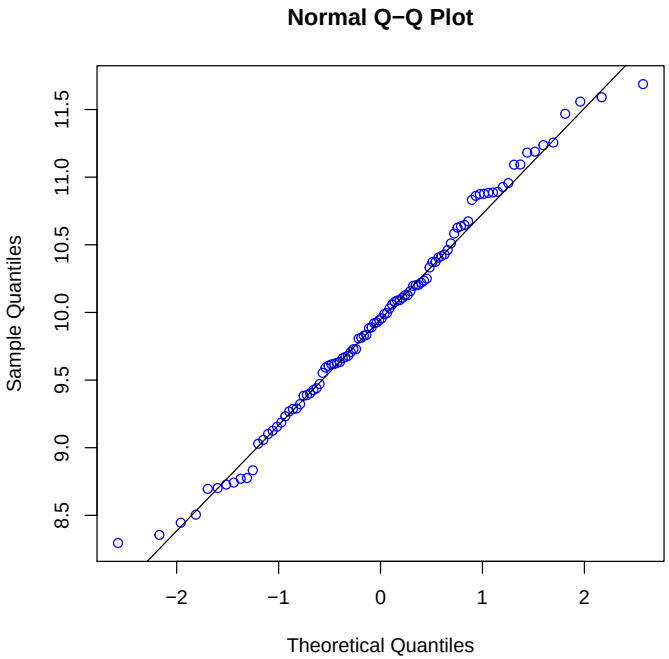


(d)

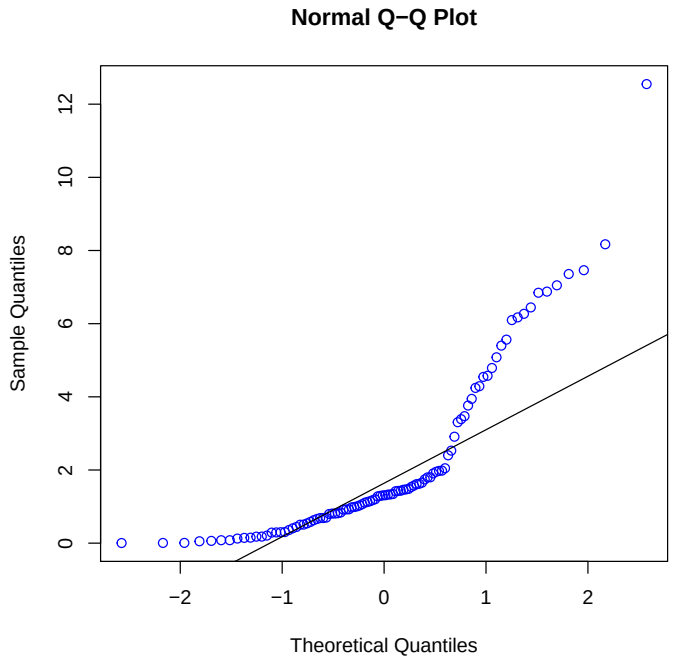
- (a): các giả định của mô hình được thỏa mãn.
- (b): phương sai tăng dần theo thời gian hoặc theo biên độ của x_i hay y_i. (c): phương sai không bằng nhau.
- (d): mối quan hệ giữa X và Y là phi tuyến tính.

Phân tích thặng dư

- Kiểm tra phân phối chuẩn sử dụng đồ thị Normal Q-Q Plot.



Dữ liệu tuân theo phân phối chuẩn



Dữ liệu không tuân theo phân phối chuẩn

Phân tích tương quan

• Phân tích tương quan (Correlation Analysis) dùng để đo độ mạnh của mối liên hệ tuyến tính giữa hai biến ngẫu nhiên.

Định nghĩa

Xét hai biến ngẫu nhiên X, Y. Hiệp phương sai (Covariance) của X và Y, ký hiệu là $Cov(X, Y)$, được định nghĩa như sau

$$Cov(X, Y) = \mathbb{E}\left[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\right] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \tag{32}$$

Định nghĩa

Hệ số tương quan (Correlation coefficient) của hai biến ngẫu nhiên X và Y, ký hiệu ρ_{XY}, được xác định như sau

$$\rho_{XY} = \frac{Cov(X, Y)}{\sqrt{\mathbb{V}ar(X)\mathbb{V}ar(Y)}} \tag{33}$$

Với hai biến ngẫu nhiên X và Y bất kỳ: $-1 \leq \rho_{XY} \leq 1$.

Phân tích tương quan

Định nghĩa

Với mẫu ngẫu nhiên cỡ n: $(X_i, Y_i), i = 1, \ldots, n$. Hệ số tương quan mẫu, ký hiệu r_{XY}, được xác định như sau

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{S_{XY}}{\sqrt{S_{XX} SST}} \tag{34}$$

Chú ý rằng:

$$\hat{\beta}_1 = \sqrt{\frac{SST}{S_{XX}}} r_{XY},$$

suy ra,

$$r_{XY}^2 = \hat{\beta}_1^{\,2} \frac{S_{XX}}{SST} = \hat{\beta}_1 \frac{S_{XY}}{SST} = \frac{SSR}{SST}.$$

• Hệ số xác định , R^2, của mô hình hồi quy tuyến tính đơn bằng với bình phương của hệ số tương quan mẫu

$$R^2 = r_{XY}^2.$$

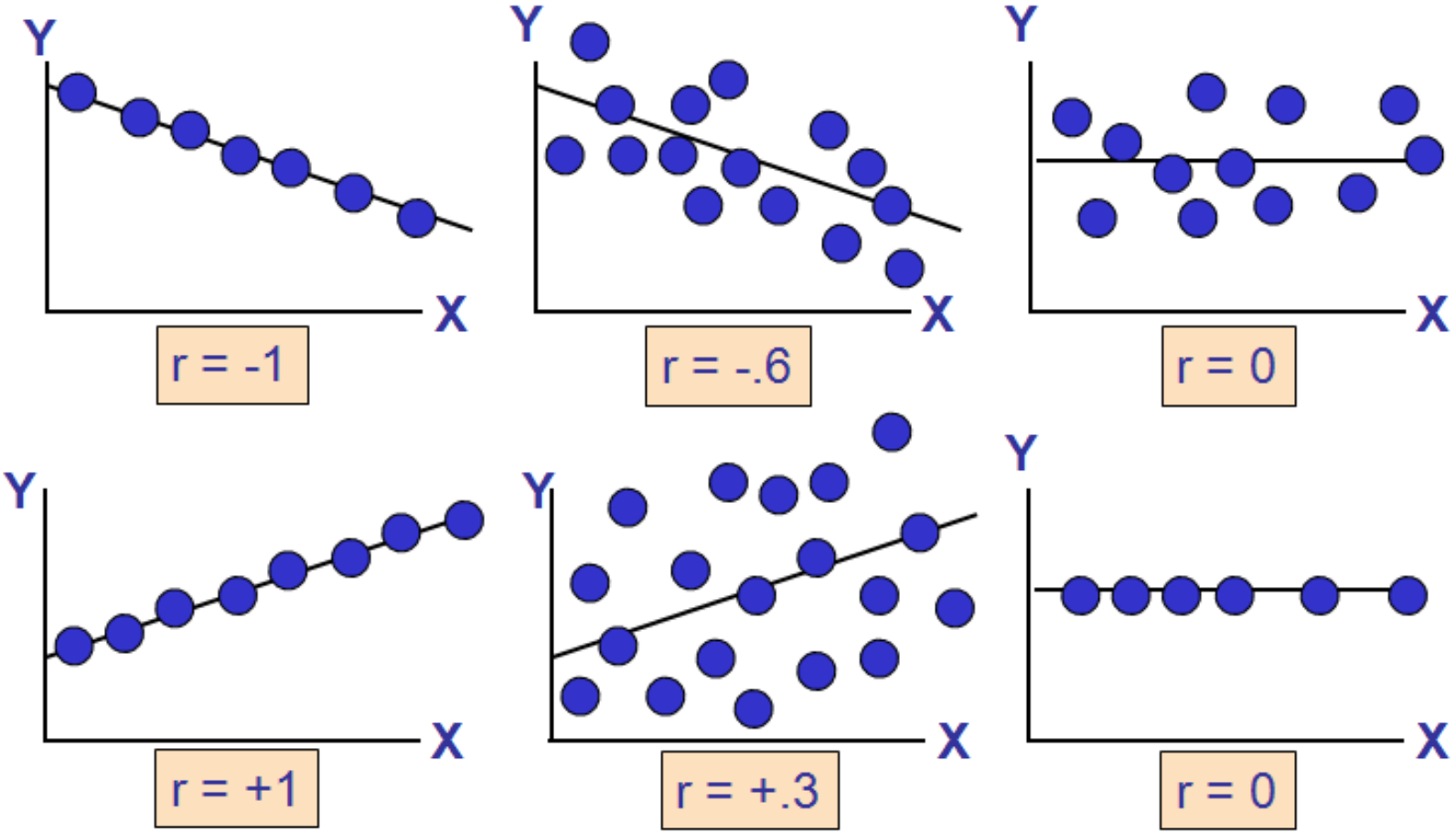
Đánh giá hiệp phương sai

- $Cov(X, Y) > 0$: X và Y có xu hướng thay đổi cùng chiều.
- $Cov(X, Y) < 0$: X và Y có xu hướng thay đổi ngược chiều.
- $Cov(X, Y) = 0$: X và Y độc lập (tuyến tính).

Đánh giá hệ số tương quan

- Miền giá trị: $-1 \leq r_{XY} \leq 1$,
- $-1 \leq r_{XY} < 0$: tương quan âm. r_{XY} càng gần -1 biểu thị mối liên hệ tuyến tính nghịch giữa X và Y càng mạnh.
- $0 < r_{XY} \leq 1$: tương quan dương. r_{XY} càng gần 1 biểu thị mối liên hệ tuyến tính thuận giữa X và Y càng mạnh.
- r_{XY} càng gần 0, biểu thị mối liên hệ tuyến tính yếu. $r_{XY} = 0$: không có mối liên hệ tuyến tính giữa X và Y.

Đánh giá hệ số tương quan



Kiểm định giả thuyết cho hệ số tương quan

• Bài toán kiểm định giả thuyết cho hệ số tương quan ρ_{XY} của mô hình hồi quy tuyến tính đơn gồm các trường hợp sau

$$(a)\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{cases} \qquad (b)\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho < 0 \end{cases} \qquad (c)\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho > 0 \end{cases},$$

với mức ý nghĩa α cho trước.

Kiểm định giả thuyết cho hệ số tương quan

Các bước kiểm định

1. Phát biểu giả thuyết H_0 và đối thuyết H_1,
2. Xác định mức ý nghĩa α,
3. Tính giá trị thống kê kiểm định:

$$t_0 = r_{XY}\sqrt{\frac{n-2}{1-r_{XY}^2}}.$$

4. Xác định miền bác bỏ: bác bỏ H_0 nếu $|t_0| > t_{1-\alpha/2}^{n-2}$.
5. Tra bảng Student tìm $t_{1-\alpha/2}^{n-2}$ hoặc tính p-giá trị:

$$p = 2\mathbb{P}(T_{n-2} \geq |t_0|),$$

   và bác bỏ H_0 nếu $p \leq \alpha$.
6. Kết luận.

Bài tập

Bài tập 1

Trong một bài báo về Nghiên cứu Bê tông: "Đặc tính bề mặt gần bê tông: tính thấm nội tại" trình bày dữ liệu về cường độ nén (X) và độ thấm nội tại (Y) của các hỗn hợp bê tông và phương pháp xử lý khác nhau. Số liệu được tóm tắt như sau:

$$n = 14, \sum_{i=1}^{n} x_i = 43, \sum_{i=1}^{n} x_i^2 = 157.42, \sum_{i=1}^{n} y_i = 572,$$
$$\sum_{i=1}^{n} y_i^2 = 23530, \sum_{i=1}^{n} x_i y_i = 1697.80.$$

(a) Xác định đường thẳng hồi quy ước lượng mô tả mối quan hệ tuyến tính giữa cường độ nén và độ thấm nội tại của bê tông.

(b) Ước lượng phương sai σ^2 của sai số.

(c) Sử dụng đường thẳng hồi quy ước lượng, hãy tiên đoán độ thấm nội tại của bê tông khi cường độ nén $x_0 = 4.3$?

(d) Tính hệ số xác định R^2 và cho nhận xét về mối liên hệ giữa X và Y.

Bài tập

Bài tập 2

Xét mẫu gồm 10 cặp giá trị (x_i, y_i) cho bởi bảng

x_i	-1	0	2	-2	5	6	8	11	12	-3
y_i	-5	-4	2	-7	6	9	13	21	20	-9

(a) Vẽ biểu đồ phân tán cho dữ liệu, tìm đường thẳng hồi quy ước lượng.

(b) Tìm ước lượng $\hat{\sigma}^2$ cho phương sai σ^2 của sai số ngẫu nhiên.

(c) Tính hệ số xác đinh R^2 và hệ số tương quan mẫu r_{XY}.

(d) Thực hiện kiểm định giả thuyết cho hệ số β_1.

Bài tập

Bài tập 3

Một nghiên cứu ảnh hưởng việc gia tăng liều dùng X (mg/kg) của một loại thuốc ngủ trên thời gian ngủ Y (giờ). Kết quả thực nghiệm ghi nhận được như sau:

x_i	1	1	2	2	3	4	5	5
y_i	1	1.2	1.5	1.7	2	2.2	2.5	2.2

(a) Tìm phương trình hồi quy của Y theo X.

(b) Tìm $\hat{\sigma}^2$ và hệ số xác định R^2.

(c) Nếu liều dùng thuốc ngủ là $x_0 = 4$ (mg/kg), thì thời gian ngủ dự đoán bằng bao nhiêu?

(d) Có tài liệu cho biết phương trình hồi quy của Y theo X là $y = 0.29x + 0.93$. Hỏi kết quả quan sát có phù hợp với phương trình cho biết không? $\alpha = 0.05$.