KHÓA TẬP HUẤN GIẢNG VIÊN KHU VỰC MIỀN BẮC 2023

Mô hình hồi quy tuyến tính

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

Đào Thị Thanh Bình
Hanoi University
binhdtt@hanu.edu.vn

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 được sử dụng để xác định mối liên hệ giữa:
  - một biến phụ thuộc Y, 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 phải là biến liên tục (trong bối cảnh ta đang xét là hồi quy tuyến tính),
    - Các biến độc lập $X_1, X_2, \ldots, X_p$ có thể là biến liên tục, rời rạc hoặc phân loại.
  - Mối liên hệ giữa $X_1, \ldots, X_p$ và Y được biểu diễn bởi một hàm tuyến tính, tức là
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \text{ sai số }.$$
  - 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.

Định nghĩa 1

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.*

- Các sai số ngẫu nhiên $\epsilon_i, i = 1, \ldots, n$ trong mô hình (6) 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.

- Cho trước $X = x$, ta có:

$$\mathbb{E}(Y|X = x) = \beta_0 + \beta_1 x. \tag{2}$$

Suy ra phân phối có điều kiện của Y cho trước $X = x$ là

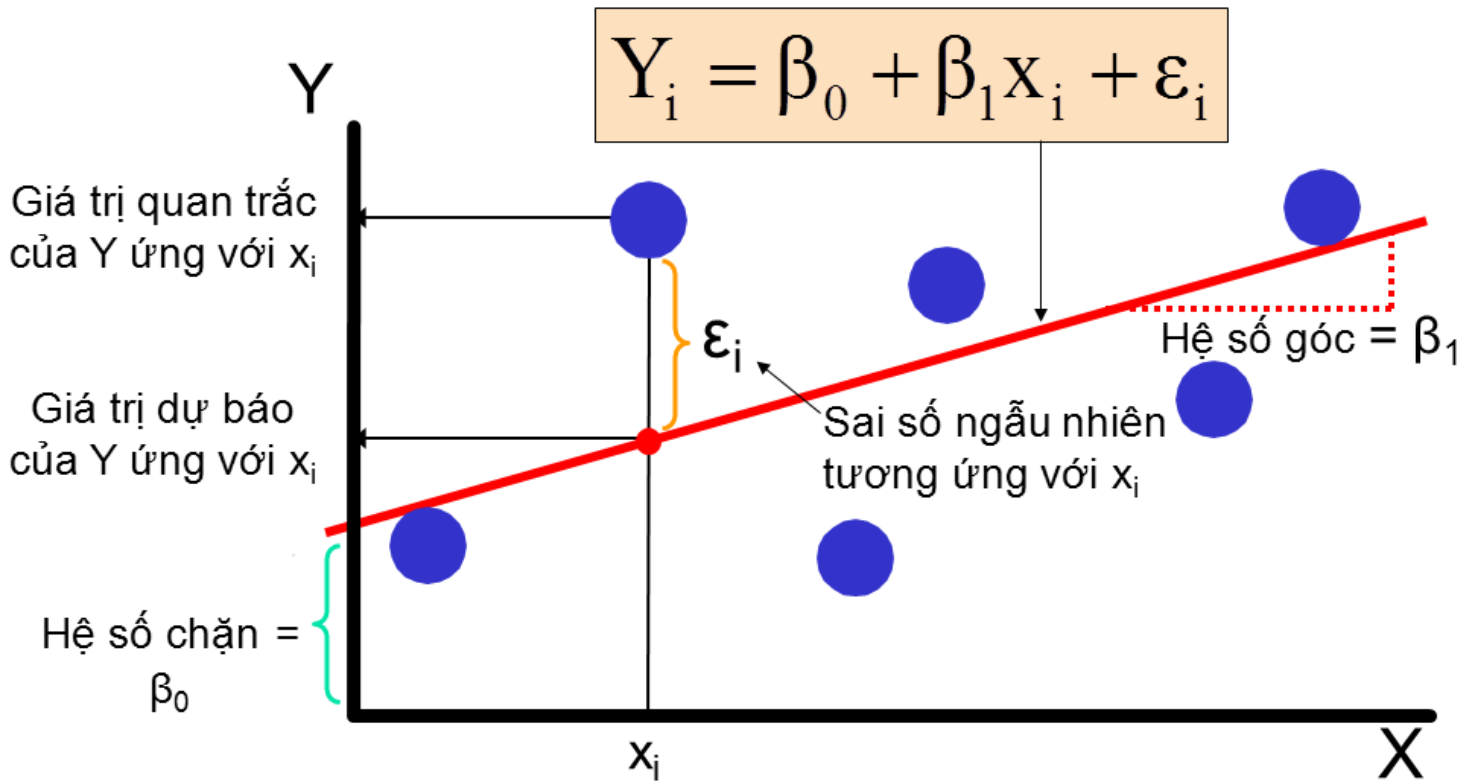
$$Y|X = x \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2) \tag{3}$$

- Trong mô hình (5), 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 (5) ta có

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \ldots, n \tag{4}$$

- 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.



- Gọi $\hat{\beta}_1$ và $\hat{\beta}_0$ là các ước lượng của β_0 và β_1.
- Đường thẳng hồi quy với các hệ số ước lượng (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).

- 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

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

Định nghĩa 2

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

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

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

Từ (8), lấy đạo hàm theo β_0 và β_1,

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

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

ta thu được hệ phương trình

$$n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i, \tag{9}$$

$$\beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i.$$

Giả hệ (9), ta tìm được các ước lượng bình phương bé nhất 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

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

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

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

(1)

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

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).

Gọi:

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

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

SST còn được ký hiệu là S_{yy}.

- 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),

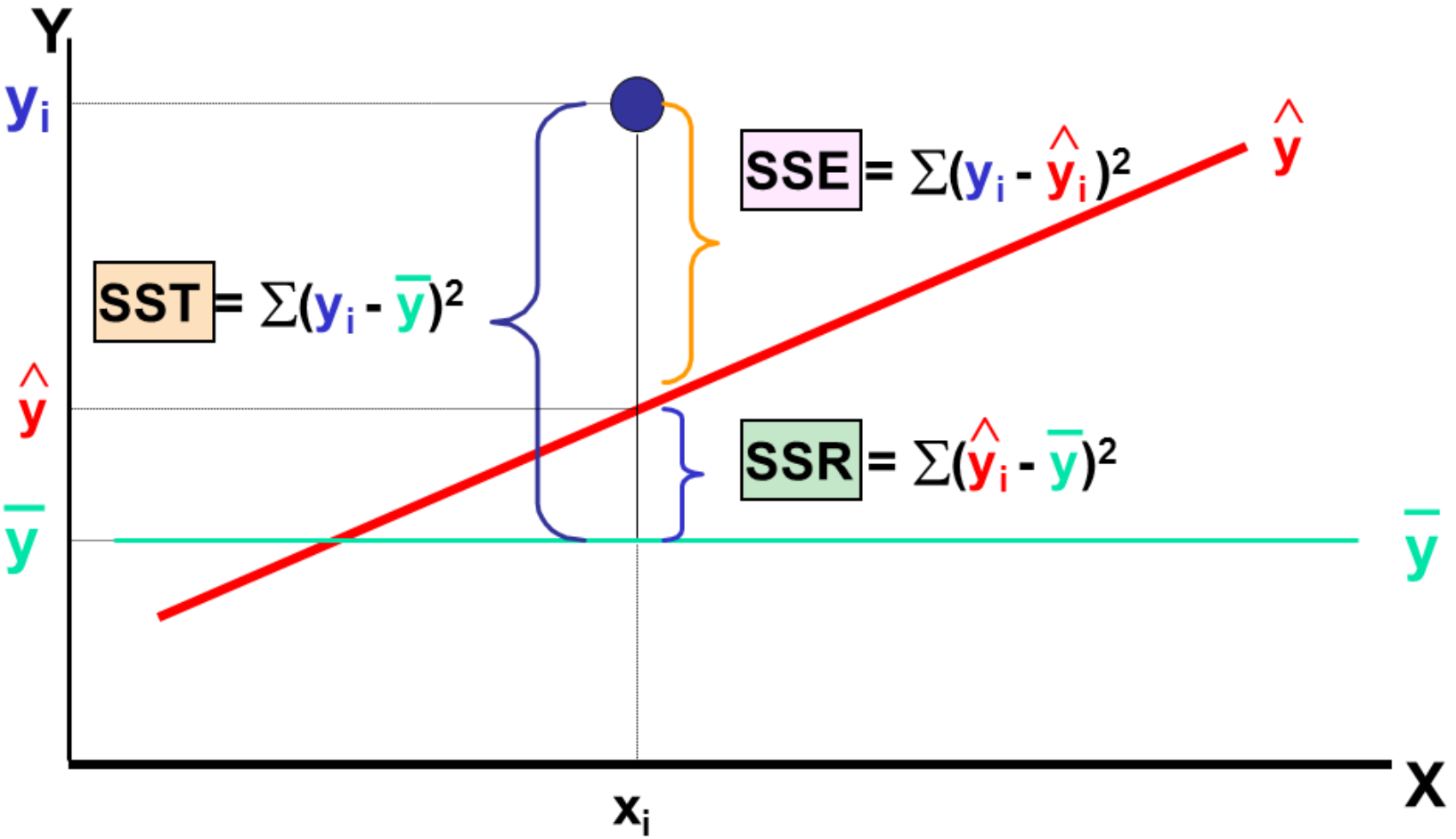
$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

- 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 chứng tỏ được:

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

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



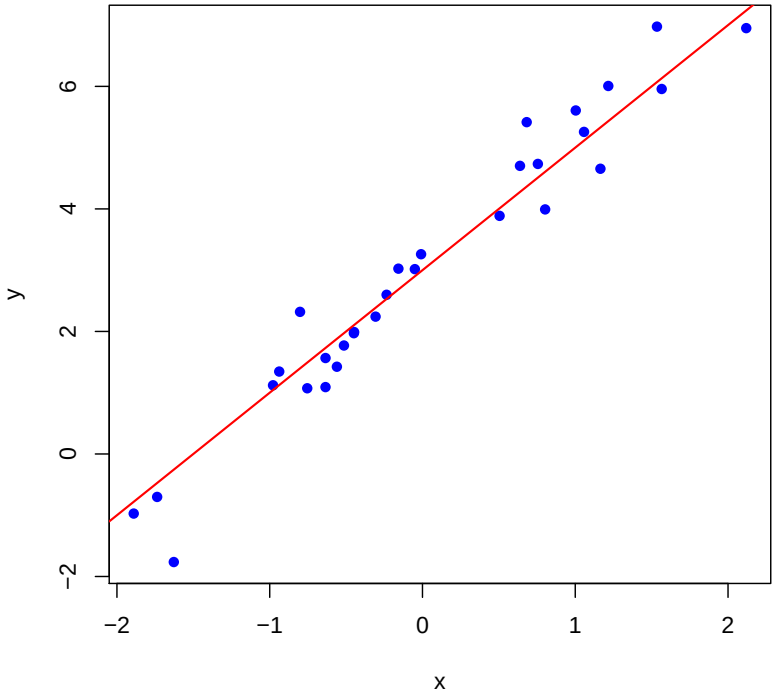
Định nghĩa 3

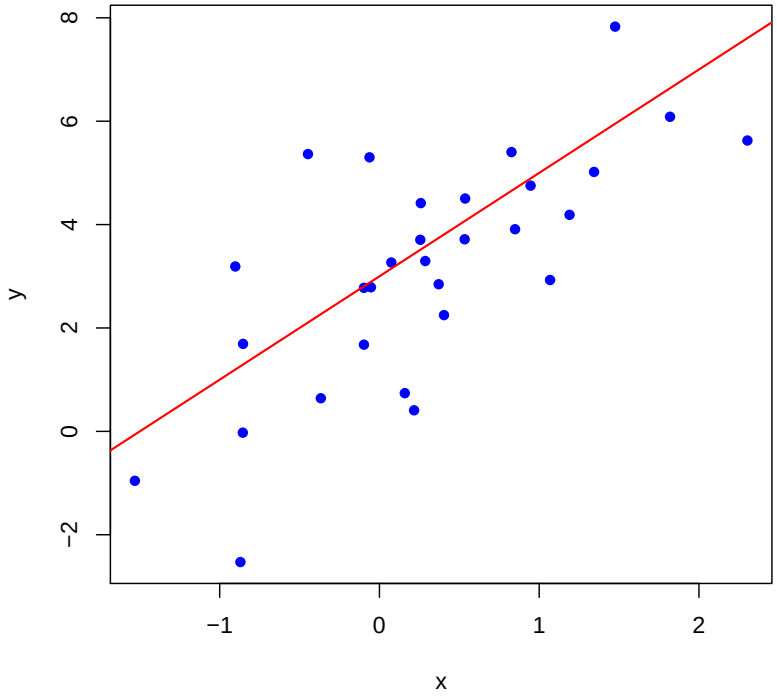
Hệ số xác định (Coefficient of Determination) *là tỷ lệ của tổng sự biến thiên trong biến phụ thuộc gây ra bởi sự biến thiên của các biến độc lập (biến giải thích) so với tổng sự biến thiên toàn phần.*
Hệ số xác định thường được gọi là R - bình phương (R-squared), ký hiệu là R^2.
Công thức tính:

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

Chú ý: $0 \leq R^2 \leq 1$.

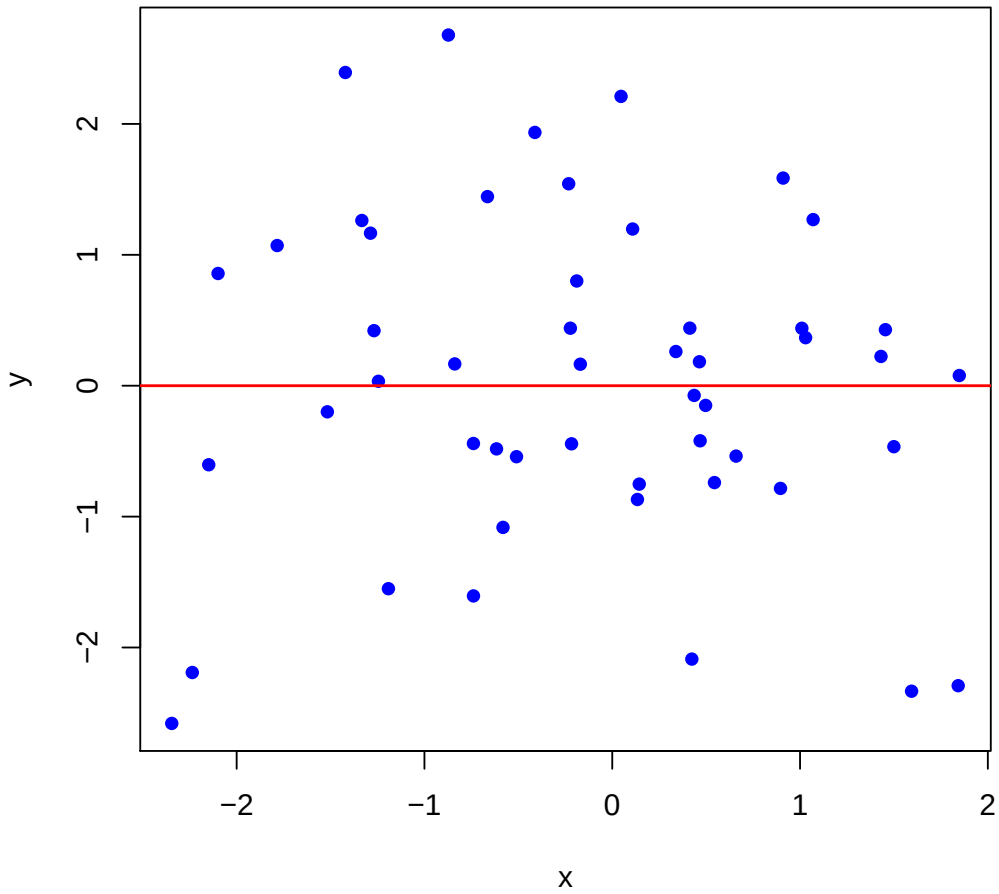
Hệ số xác định của một mô hình hồi quy cho phép ta đánh giá mô hình tìm được có giải thích tốt cho mối liên hệ giữa biến phụ thuộc Y và biến phụ thuộc X hay không.





- R^2 càng gần 1 thì mối quan hệ tuyến tính giữa X và Y càng mạnh. Đa số sự biến thiên của Y được giải thích bởi X.

- R^2 càng gần 0 thì mối quan hệ tuyến tính giữa X và Y càng yếu. Sự biến thiên của Y càng ít được giải thích bởi X.



- $R^2 = 0$: không có mối liên hệ tuyến tính giữa X và Y. Không có sự biến thiên nào của Y được giải thích bởi X.

Ví dụ 1

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_i	411	550	471	393	427	431	492	371	470	419	407	489
y_i	2.00	2.46	2.11	1.89	2.05	2.30	2.46	2.06	2.25	2.07	2.17	2.32

(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.*

(c) *Tính hệ số R^2 và nhận xét về mô hình.*

- Sai số chuẩn (Standard Error - SE) của $\hat{\beta}_0$ và $\hat{\beta}_1$ lần lượt cho bởi

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

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

với

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

Khoảng tin cậy $100(1-\alpha)\%$ cho các hệ số hồi quy β_1 và β_0 lần lượt được cho bởi:

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

$$\hat{\beta}_1 - t_{\alpha/2}^{n-2}\,\mathrm{SE}(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2}^{n-2}\,\mathrm{SE}(\hat{\beta}_1) \tag{16}$$

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

$$\hat{\beta}_0 - t_{\alpha/2}^{n-2}\,\mathrm{SE}(\hat{\beta}_0) \leq \beta_0 \leq \hat{\beta}_0 + t_{\alpha/2}^{n-2}\,\mathrm{SE}(\hat{\beta}_0) \tag{17}$$

Bài toán:

- 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.

- Bài toán kiểm định giả thuyết cho hệ số chặn β_0 trong mô hình hồi quy tuyến tính đơn như sau:

$$\begin{cases} H_0 : \beta_0 = b_0 \\ H_1 : \beta_0 \neq b_0 \end{cases}$$

với giá trị b_0 và mức ý nghĩa α cho trước. Thông thường $b_0 = 0$.

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:

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

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 tri tính bởi

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

và bác bỏ H_0 khi $p \leq \alpha$.

- Bài toán kiểm định giả thuyết cho hệ số góc β_1 trong mô hình hồi quy tuyến tính đơn như sau:

$$\begin{cases} H_0 : \beta_1 = b_1 \\ H_1 : \beta_1 \neq b_1 \end{cases}$$

với giá trị b_1 và mức ý nghĩa α cho trước. Thông thường $b_1 = 0$.

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:

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

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 $p \leq \alpha$.

Định nghĩa 4

Xét hai biến ngẫu nhiên X, Y. Hiệp phương sai (Covariance) của X và Y, ký hiệu là $\mathbb{C}\text{ov}(X, Y)$, được định nghĩa như sau

$$Cov(X, Y) = \mathbb{E}\left[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\right] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y). \quad (18)$$

Định nghĩa 5

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

$$\rho_{XY} = \frac{Cov(X, Y)}{\sqrt{\mathbb{V}ar(X)\mathbb{V}ar(Y)}}. \quad (19)$$

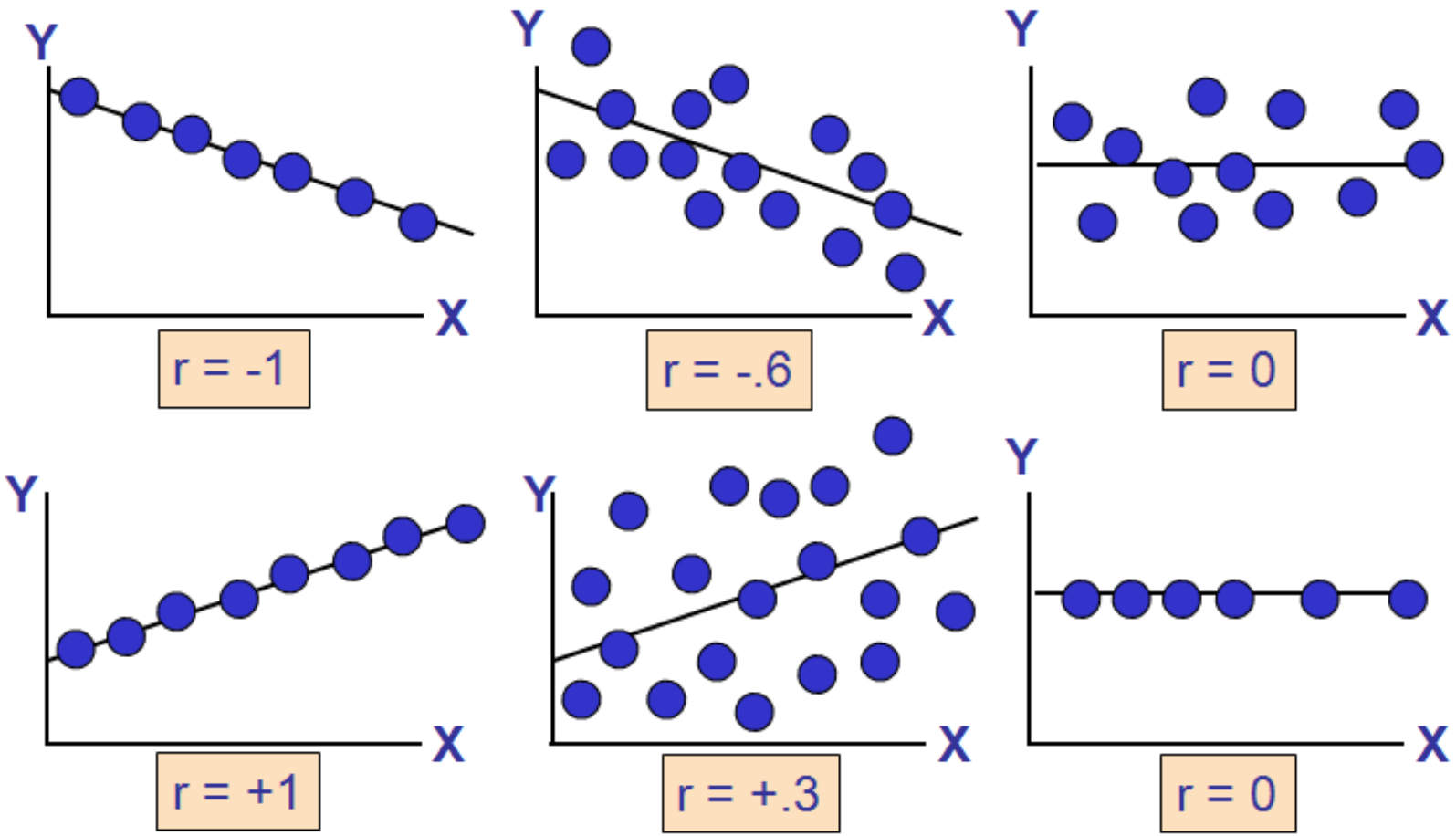
Với hai biến ngẫu nhiên X và Y bất kỳ: $-1 \leq \rho_{XY} \leq 1$.

Định nghĩa 6

Với mẫu cỡ n: $(x_i, y_i), i = 1, \ldots, n$, *hệ số tương quan mẫu, ký hiệu* r_{XY}, *được xác định như sau*

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}. \quad (20)$$

- Ta có: $-1 \leq r_{XY} \leq 1$.
- $-1 \leq r_{XY} < 0$: 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.



- Chú ý rằng,

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \sqrt{\frac{S_{yy}}{S_{xx}}}\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \sqrt{\frac{S_{yy}}{S_{xx}}}r_{XY} = \sqrt{\frac{SST}{S_{xx}}}r_{XY},$$

  vì S_{yy} chính là SST.

- Suy ra,

$$r_{XY}^2 = \hat{\beta}_1{}^2\frac{S_{xx}}{SST} = \hat{\beta}_1\frac{S_{xy}}{S_{xx}}\frac{S_{xx}}{SST} = \frac{\hat{\beta}_1 S_{xy}}{SST} = \frac{SSR}{SST} = R^2.$$

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

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

Chú ý: R^2 ở đây là một ký hiệu, không phải là bình phương của R.

- Bài toán kiểm định giả thuyết cho hệ số tương quan ρ_{XY} của mô hình hồi quy tuyến tính đơn như sau:

$$\begin{cases} H_0 : \rho = 0 & \text{(không có tương quan giữa } X \text{ và } Y) \\ H_1 : \rho \neq 0 & \text{(tồn tại tương quan giữa } X \text{ và } Y) \end{cases}$$

với mức ý nghĩa α cho trước.

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:

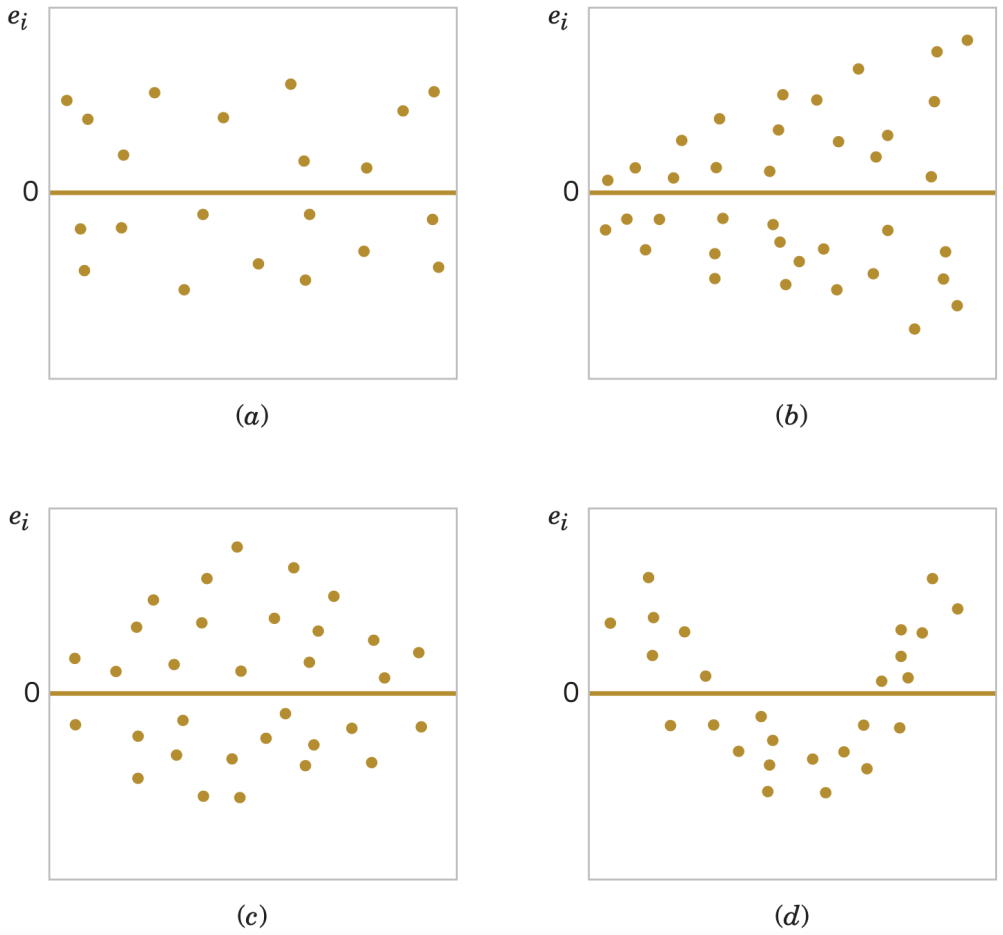
$$p = 2\mathbb{P}(T_{n-2} \geq |t_0|),$$

và bác bỏ H_0 nếu $p \leq \alpha$.

6. Kết luận.

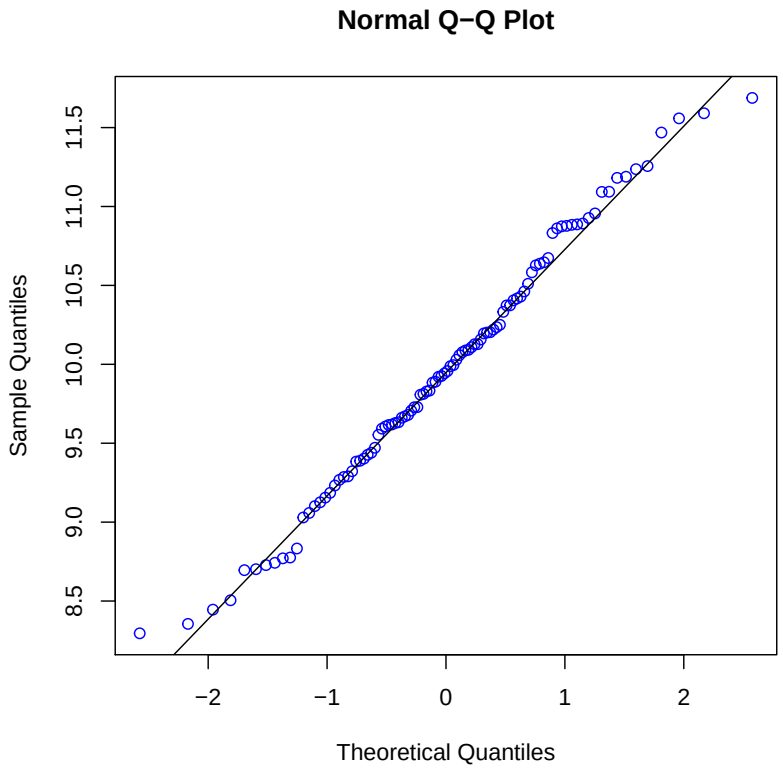
- Phân tích thặng dư (Residual Analysis) được sử dụng để kiểm tra các giả định của mô hình hồi quy tuyến tính.
- Các giả định của mô hình:
  1. Tuyến tính: mối quan hệ giữa X và Y là tuyến tính tính, tức là
  $$\mathbb{E}[Y|X = x] = \beta_0 + \beta_1 x.$$
  2. Phương sai bằng nhau: phương sai của biến 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 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

- 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.

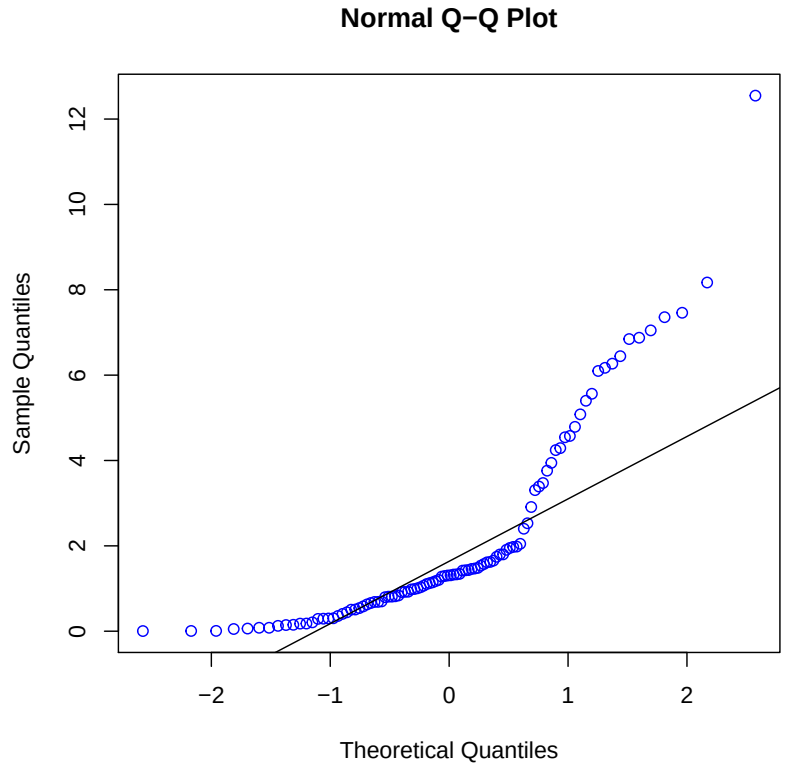


- (a): các giả định của mô hình được thỏa mãn.
- (b): phương sai tăng dần theo thời gian hoặc theo biên độ của x_i hay y_i. (c): phương sai không bằng nhau.
- (d): mối quan hệ giữa X và Y là phi tuyến tính.

- 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