Hồi quy và tương quan

Nguyễn Thị Hiên

Ngày 24 tháng 4 năm 2024

1 Giới thiệu Phân tích hồi quy

2 Mô hình hồi quy tuyến tính đơn
- Mô hình hồi quy tuyến tính đơn
- Hệ số tương quan
- Ước lượng phương sai σ^2 của sai số
- Khoảng tin cậy cho mô hình hồi quy
- Kiểm định giả thuyết cho mô hình hồi quy
- Phân tích thặng dư

3 Tương quan

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 quan liên hệ giữa hai hay nhiều biến ngẫu nhiên với nhau. Ví dụ

- Mối liên hệ giữa chiều cao và cỡ giầy của một người, từ đó một cửa hàng bán giầy 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.
- 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)
  - một biến hay nhiều biến độc lập $X_1, X_2, \ldots, X_p$; các biến này 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.
  - Mối liên hệ giữa $X_1, X_2, \ldots, X_p$ và Y được biểu diễn bởi một hàm tuyến tính.
  - Sự thay đổi trong Y được giả sử do những thay đổi trong $X_1, X_2, \ldots, X_p$ 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 giả thích x là phương trình có dạng

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

với

- 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 $Var(\epsilon) = \sigma^2$.

Mô hình hồi quy tuyến tính đơn

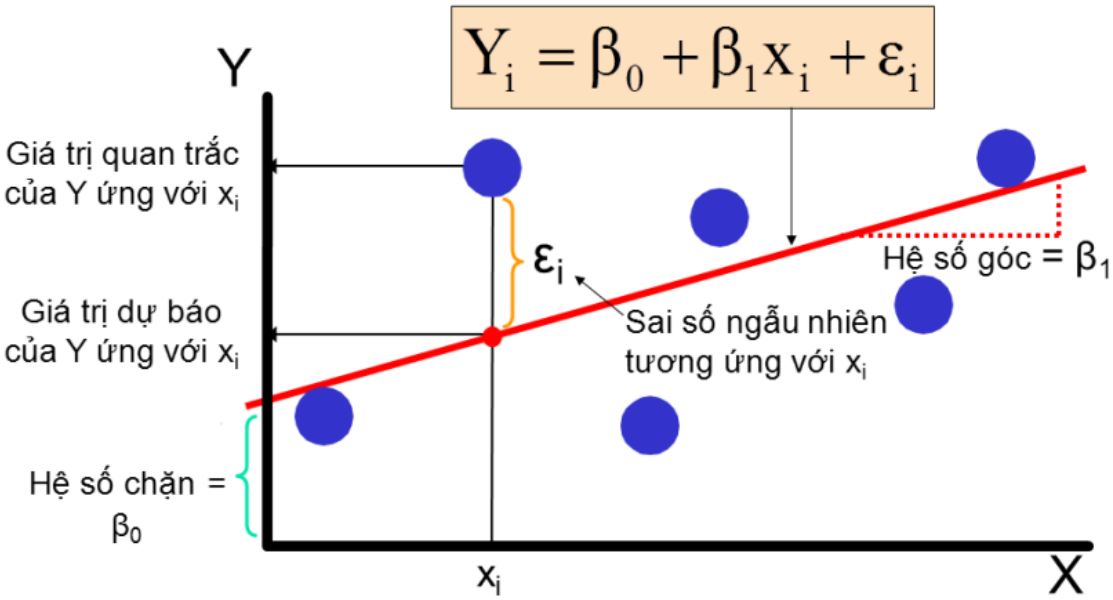
Trong mô hình (1), sự thay đổi của Y được giả sử ảnh hưởng bởi 2 yếu tố:

- Mối liên hệ tuyến tính của X và Y : $\beta_0 + \beta_1 x$. Trong đó, β_0 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

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \; i = 1, 2, \ldots, n \tag{2}$$

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_1, 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) giả sử thỏa các điều kiện sau:
  - các sai số ϵ_i độc lập với nhau,
  - $\mathbb{E}(\epsilon_i) = 0$ và $var(\epsilon_i) = \sigma^2$,
  - Các sai số có phân phối chuẩn: $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.
- Với quan trắc x đã biết,

$$\mathbb{E}(Y|x) = \beta_0 + \beta_1 x$$

- Từ các giả định trên ta có

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

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 hồi quy với các hệ số ước lượng (fitted regression line):

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{3}$$

- 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 (3) ta có

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{4}$$

- 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:

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \tag{5}$$

Phương pháp bình phương bé nhất

Định nghĩa

Tổng bình phương sai số cho n điểm dữ liệu được định nghĩa

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right]^2 \tag{6}$$

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.

Phương pháp bình phương bé nhất

Từ (6), lấy đạo hàm theo β_0 và β_1,

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

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

Ta thu được hệ phương trình

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

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

Ước lượng bình phương bé nhất

Giải hệ phương trình trên, ta tìm được các ước lượng BPBN của β_0 và β_1 là

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2} = \frac{S_{xy}}{S_{xx}} \tag{7}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{8}$$

với S_{xx} và S_{xy} xác định bởi

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n(\bar{x})^2$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

Ước lượng bình phương bé nhất

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

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

đạt giá trị bé nhất và

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

với SE là tổng các thặng dư.

Ví dụ

Một nhà thực vật học khảo sát mối liên hệ giữa tổng diện tích bề mặt (đv:cm^2) của các là cây đậu nành và trọng lượng khố (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 là 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

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

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

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

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

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

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = SST - \hat{\beta}_1 S_{xy}$$

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 liên 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

$$\begin{aligned} \sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ SST &= SSR + SSE \end{aligned}$$

Hệ số xác định

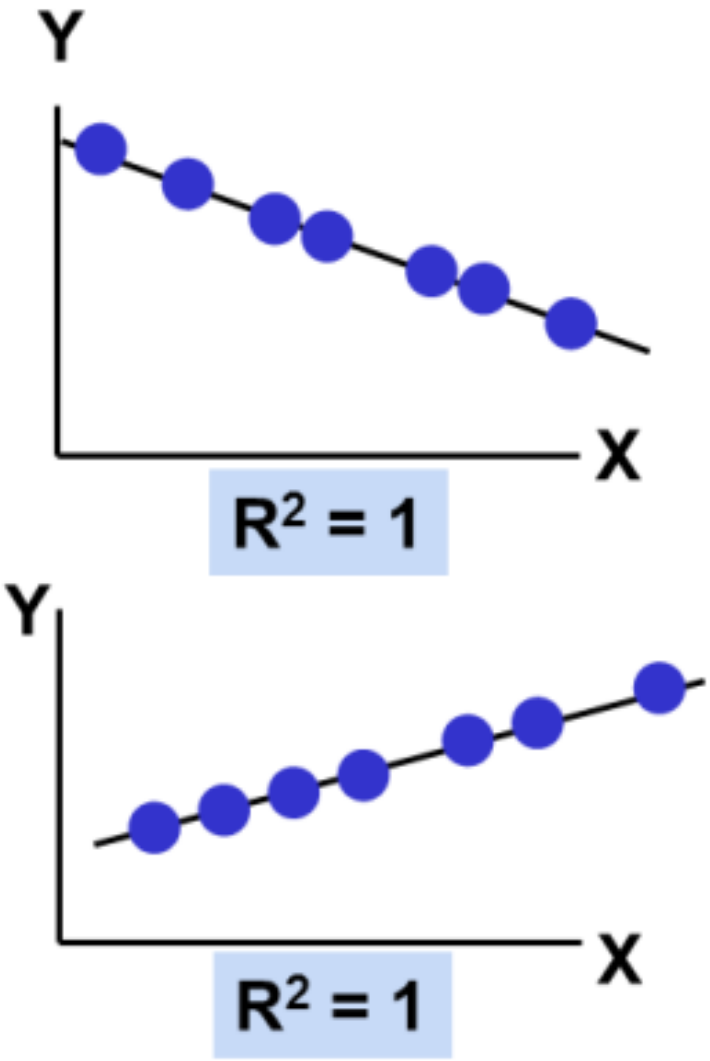
Định nghĩa

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

$$R^2 = \frac{SSR}{SST}$$

Lưu 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 vào biến độc lập X hay không.

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

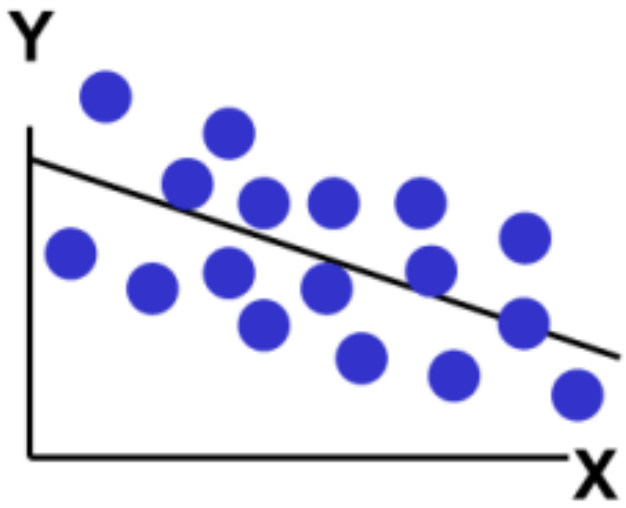


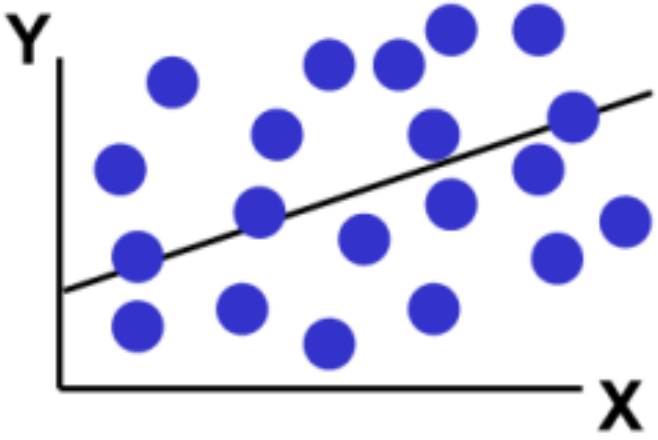
$R^2 = 1$

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



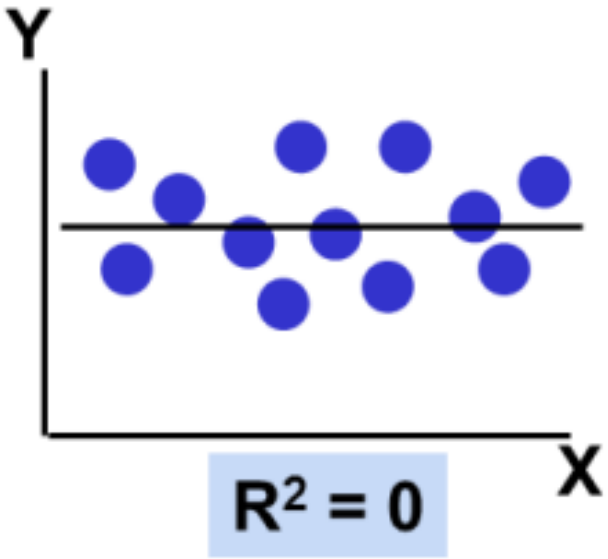


$0 < R^2 < 1$

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 : $\epsilon_i \sim N(0, \sigma^2)$. Ta cần ước lượng phương sai σ^2. Ta có: $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$. Do đó,

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

Ta có

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

Do đó,

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

Ước lượng phương sai σ^2 của sai số

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

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

Định nghĩa 1

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

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

Nói cách khác, trung 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:

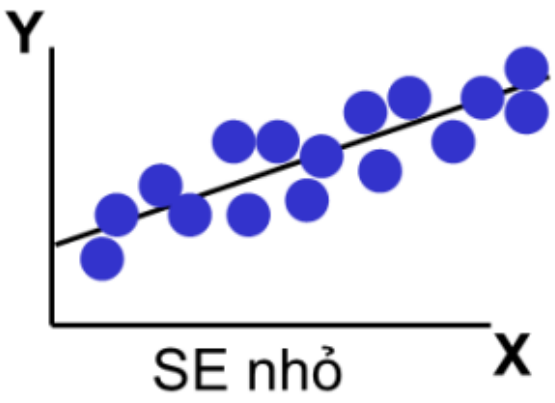
$$SSE = SST - \hat{\beta}_1 S_{xy}$$

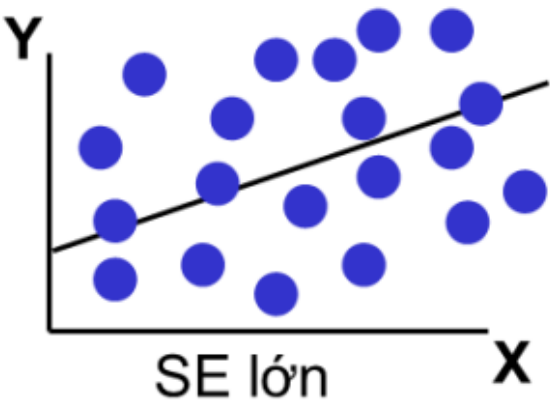
- Sai số chuẩn (Standard Error) của $\hat{\sigma}^2$

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

Sử dụng $SE(\hat{\sigma})$ 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 1

Xét $Y = \beta_0 + \beta_1 x + \epsilon$ *là một mô hình hồi quy tuyến tính đơn với* $\epsilon \sim N(0, \sigma^2)$; *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*

- $\hat{\beta}_0$ *và* $\hat{\beta}_1$ *tuân theo phân phối chuẩn.*
- *Kỳ vọng và phương sai của* $\hat{\beta}_0$ *và* $\hat{\beta}_1$ *lần lượt là*

$$\mathbb{E}(\hat{\beta}_0) = \beta_0, \ \mathbb{V}ar(\hat{\beta}_0) = \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\sigma^2, \tag{9}$$

$$\mathbb{E}(\hat{\beta}_1) = \beta_1, \ \mathbb{V}ar(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}. \tag{10}$$

Tính chất của các ước lượng BPBN

Định nghĩa 2

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

$$SE(\hat{\beta}_0) = \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\hat{\sigma}^2} \tag{11}$$

$$SE(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \tag{12}$$

Định lý 2

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

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

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

$$T_{\beta_1} = \frac{Z_1}{\sqrt{\dfrac{\left(\dfrac{SSE}{\sigma^2}\right)}{n-2}}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\left(\dfrac{SSE}{n-2}\right)/S_{xx}}} = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \sim t(n-2) \quad (13)$$

Khoảng tin cậy cho hệ số hồi quy

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

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

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

$$T_{\beta_0} = \frac{Z_0}{\sqrt{\frac{\left(\frac{SSE}{\sigma^2}\right)}{n-2}}} = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\left(\frac{SSE}{n-2}\right)\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} = \frac{\hat{\beta}_0 - \beta_0}{SE(\hat{\beta}_0)} \sim t(n-2) \tag{14}$$

Khoảng tin cậy cho hệ số hồi quy

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

$$\hat{\beta}_1 - t_{\alpha/2}^{n-2} SE(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2}^{n-2} SE(\hat{\beta}_1) \tag{15}$$

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

$$\hat{\beta}_0 - t_{\alpha/2}^{n-2} SE(\hat{\beta}_0) \leq \beta_0 \leq \hat{\beta}_0 + t_{\alpha/2}^{n-2} SE(\hat{\beta}_0) \tag{16}$$

trong đó,

- $n =$ số cặp giá trị quan trắc (x_i, y_i);
- $t_{\alpha/2}^{n-2}$ là phân vị mức $\alpha/2$ của biến ngẫu nhiên $t(n-2)$;
- $SE(\hat{\beta}_0)$ và $SE(\hat{\beta}_1)$ cho bởi phương trình (9) và (10).

Khoảng tin cậy cho trung bình biến đáp ứng

- Cho trước giá trị x_0, cần tìm khoảng tin cậy cho $\mu_{Y|x_0} = E(Y|x_0) = \beta_0 + \beta_1 x_0$, 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

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

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

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

Khoảng tin cậy cho trung bình biến đáp ứng

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

$$SE(\hat{\mu}_{Y|x_0}) = \sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}}\right)} \tag{17}$$

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

$$\hat{\mu}_{Y|x_0} - t^{n-2}_{\alpha/2} SE(\hat{\mu}_{Y|x_0}) \leq \mu_{Y|x_0} \leq \hat{\mu}_{Y|x_0} + t^{n-2}_{\alpha/2} SE(\hat{\mu}_{Y|x_0}) \tag{18}$$

trong đó

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

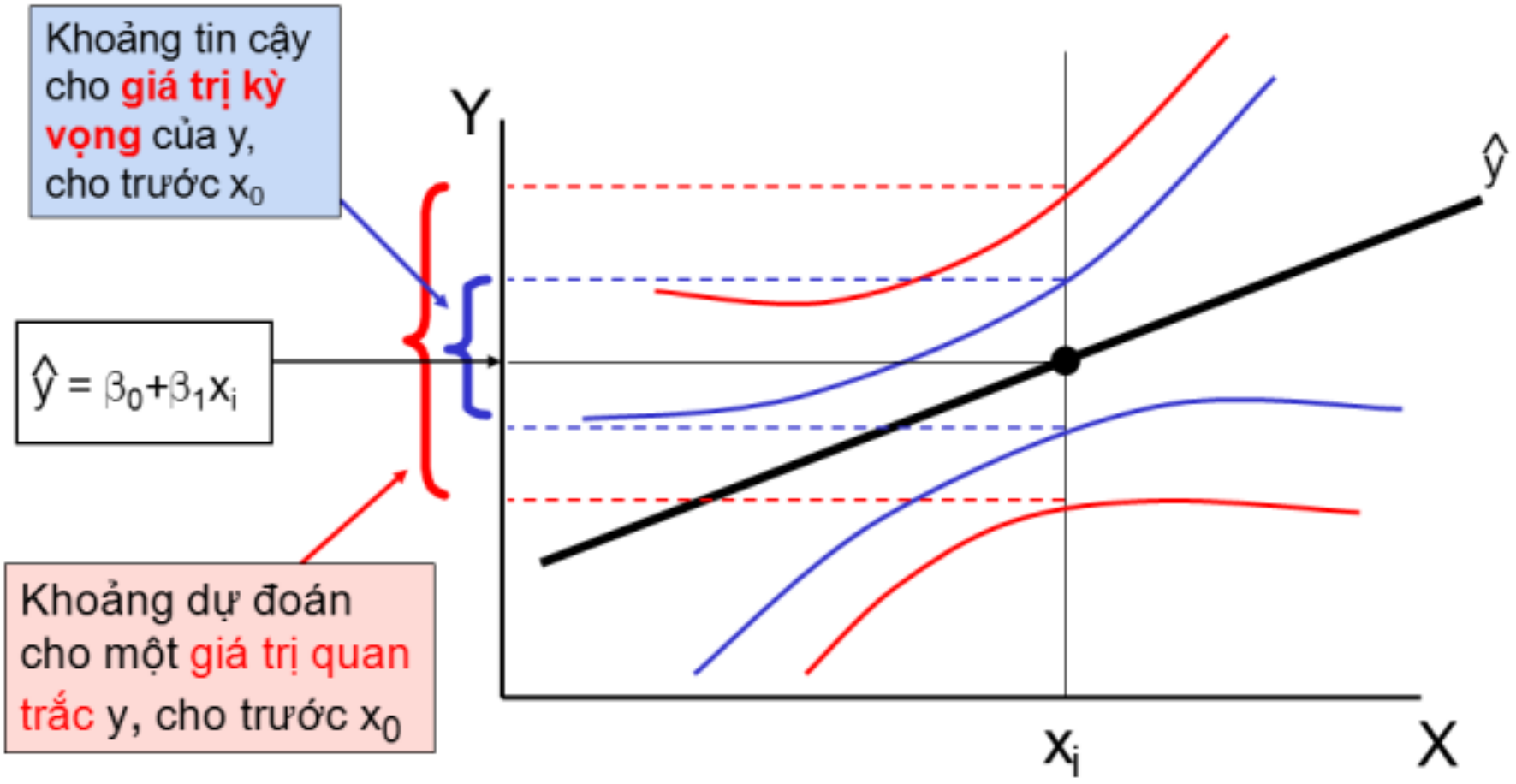
Dự đoán giá trị quan trắc mới

- 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:

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

  $\hat{Y}_0$ 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 ràng 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

$$\begin{aligned}
\mathbb{E}(\eta) &= \mathbb{E}(Y_0) - \mathbb{E}(\hat{Y}_0) = 0 && (19)\\
\mathbb{V}ar(\eta) &= \mathbb{V}ar(Y_0) + \mathbb{V}ar(\hat{Y}_0) = \sigma^2 + \sigma^2\left[\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}}\right] && (20)\\
&= \left[1 + \frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}}\right]\sigma^2 && (21)
\end{aligned}$$

Do đó,

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

Dự đoán giá trị quan trắc mới

♡ Khi đó

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

♡ Nếu ta thay thế σ^2 bởi $\hat{\sigma}^2 = \dfrac{SSE}{n-2}$, ta có

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

Dự đoán giá trị quan trắc mới

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

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

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

$$\hat{Y}_0 - t_{\alpha/2}^{n-2} SE(\hat{Y}_0) \leq Y_0 \leq \hat{Y}_0 + t_{\alpha/2}^{n-2} SE(\hat{Y}_0) \qquad (22)$$

trong đó $t_{\alpha/2}^{n-2}$ là phân vị mức $\alpha/2$ của $t(n-2)$.

Ví dụ

Ví dụ 1

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) *Tìm khoảng tin cậy* 95% *cho các hệ số* β_0 *và* β_1.

d) *Tìm khoảng dự đoán* 95% *tại* $x = 5$.

Kiểm định giả thuyết cho β_0

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

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

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

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 thống kê kiểm định:

$$T_{\beta_0} = \frac{\hat{\beta}_0 - b_0}{SE(\hat{\beta}_0)}$$

T_{β_0} có phân phối Student với $n-2$ bậc tự do và

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

Kiểm định giả thuyết cho β_0

Các bước kiểm định

4 Xác định miền bác bỏ

Đối thuyết	**Miền bác bỏ**	$p-$ **giá trị**
$H_1 : \beta_0 \neq b_0$	$\lvert t_{\beta_0} \rvert > t^{n-2}_{\alpha/2}$	$p = 2\mathbb{P}(T^{n-2} \geq \lvert t_{\beta_0} \rvert)$
$H_1 : \beta_0 < b_0$	$t < -t^{n-2}_{\alpha}$	$p = \mathbb{P}(T^{n-2} \leq t_{\beta_0})$
$H_1 : \beta_0 > b_0$	$t > t^{n-2}_{\alpha}$	$p = \mathbb{P}(T^{n-2} \geq t_{\beta_0})$

5 Kết luận: Bác bỏ H_0 Chưa đủ cơ sở để bác bỏ H_0.

Kiểm định giả thuyết cho β_1

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

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

với giá trị b_1 và mức ý nghĩa α cho trước.

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 thống kê kiểm định:

$$T_{\beta_1} = \frac{\hat{\beta}_1 - b_1}{SE(\hat{\beta}_1)}$$

T_{β_1} có phân phối Student với $n-2$ bậc tự do và

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

Kiểm định giả thuyết cho β_1

Các bước kiểm định

4 Xác định miền bác bỏ

Đối thuyết	**Miền bác bỏ**	$p-$ **giá trị**
$H_1 : \beta_1 \neq b_1$	$\lvert t_{\beta_1} \rvert > t^{n-2}_{\alpha/2}$	$p = 2\mathbb{P}(T^{n-2} \geq \lvert t_{\beta_1} \rvert)$
$H_1 : \beta_1 < b_1$	$t_{\beta_1} < -t^{n-2}_{\alpha}$	$p = \mathbb{P}(T^{n-2} \leq t_{\beta_1})$
$H_1 : \beta_1 > b_1$	$t_{\beta_1} > t^{n-2}_{\alpha}$	$p = \mathbb{P}(T^{n-2} \geq t_{\beta_1})$

5 Kết luận: Bác bỏ H_0/ Chưa đủ cơ sở để bác bỏ H_0.

Phân tích thặng dư

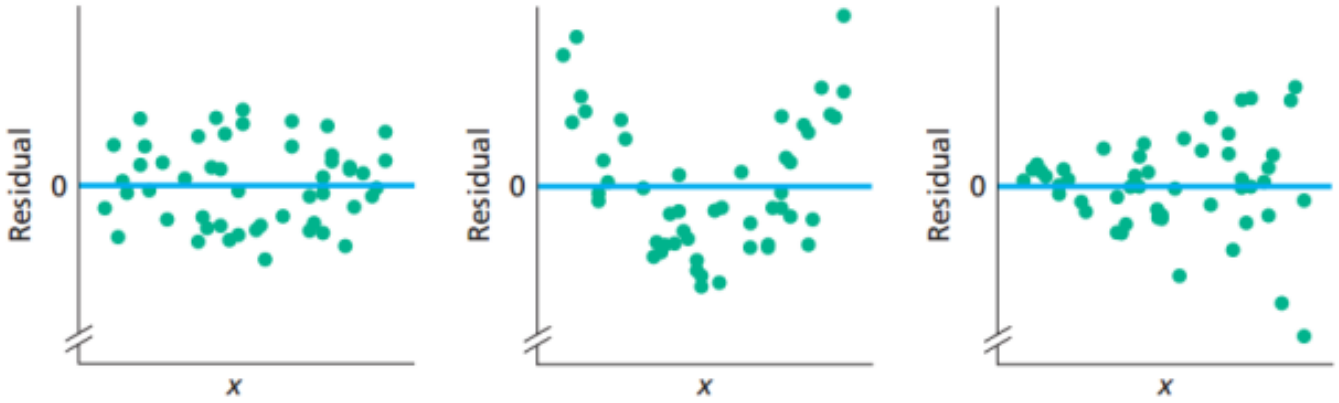
- Phân tích thặng dư (Analysis of Residuals) được sử dụng để kiểm tra các giả định của mô hình hồi quy tuyến tính.
- Các giả định của mô hình:
  1. Đường thẳng hồi quy tổng thể: $Y = \beta_0 + \beta_1 x + \epsilon$ với β_0 và β_1 là các hằng số sao cho với mỗi giá trị x, $\mathbb{E}(Y|x) = \beta_0 + \beta_1 x$.
  2. Phương sai bằng nhau: Phương sai của biến đáp ứng (biến phụ thuộc) Y bằng nhau với mọi giá trị của biến độc lập X, tức là $\mathbb{V}ar(Y|x) = \sigma^2$.
  3. 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, $Y|x \sim N(\beta_0 + \beta_1 x, \sigma^2)$.
  4. Độc lập: các quan trắc của biến đáp ứng Y độc lập với nhau.

Phân tích thặng dư

- Với n cặp quan trắc (x_i, y_i), gọi $e_i = y_i - \hat{y}_i$ là thặng dư thứ i. Để kiểm tra các giả định của mo hình có thoả hay không, nhân xét trên đồ thị của các giá trị thặng sư.

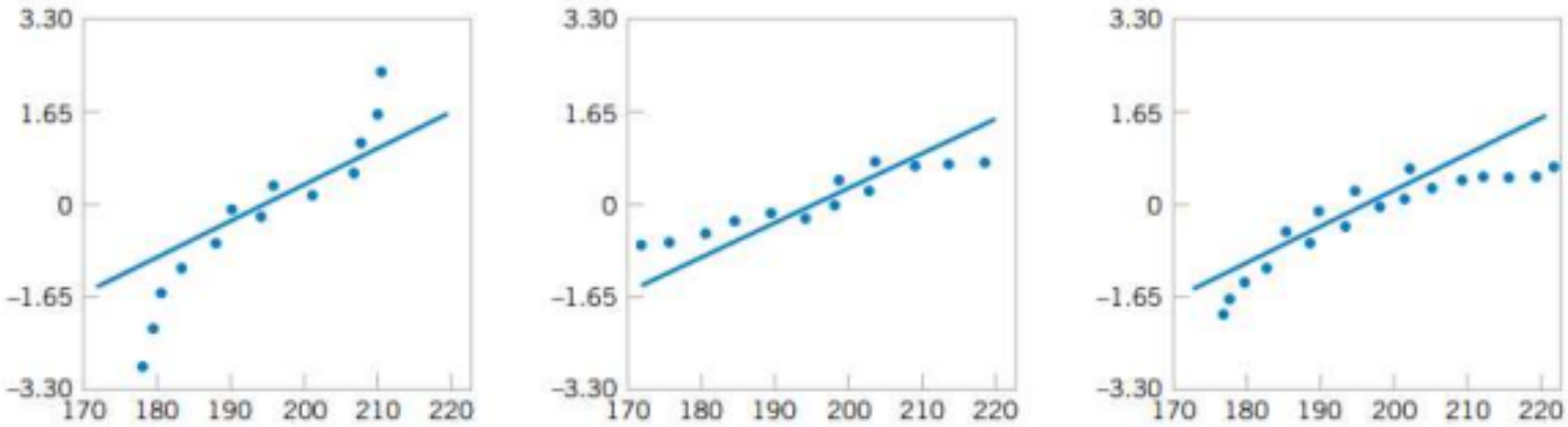
Phân tích thặng dư

- Đồ thị thặng dư: Đồ thị của các giá trị thặng dư tương ứng vưới cá giá trị biến độc lập X phải có hình dạng thô, không đều, trải dọc theo chiều ngang và đối xứng qua trục Ox.



Phân tích thặng dư

- Đồ thị phân vị chuẩn (Normal Q-Q plot): có dạng gần tuyến tính.



Phân tích tương quan

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

Định nghĩa

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

$$Cov(X, Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y) \quad (23)$$

Định nghĩa

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

$$\rho_{XY} = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} \quad (24)$$

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, Y), $i = 1, \ldots, n$. Hệ số tương quan mẫu, ký hiệu r_{XY}, được xác định:

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

Chú ý:

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

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 của mô hình hồi quy tuyến tính đơn bằng với bình phương của hệ số tương quan mẫu

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

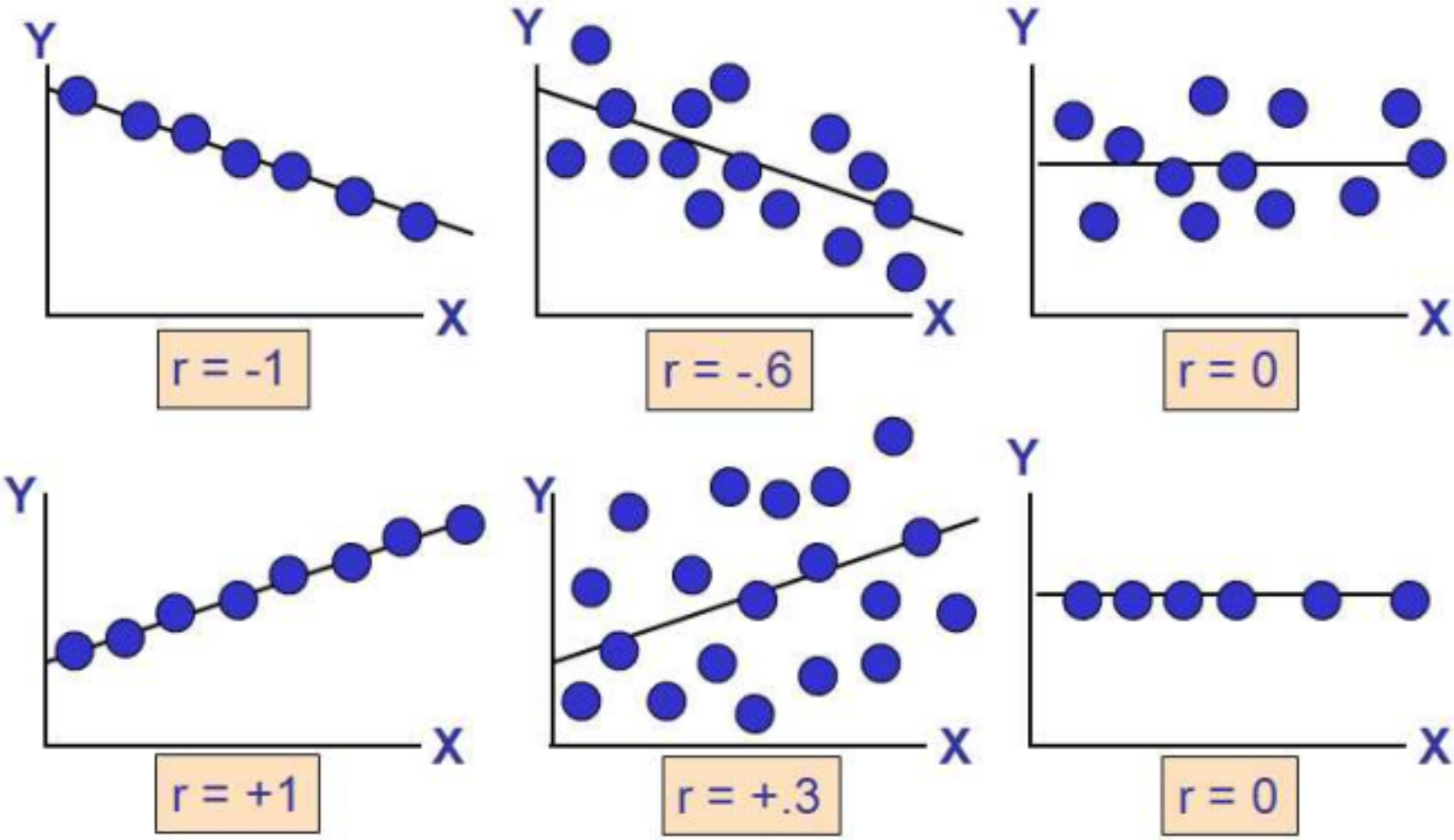
Đánh giá hiệp phương sai

- $Cov(X, Y) > 0$: 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:

$$\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{cases} \qquad \begin{cases} H_0 : \rho = 0 \\ H_1 : \rho < 0 \end{cases} \qquad \begin{cases} H_0 : \rho = 0 \\ H_1 : \rho > 0 \end{cases}$$

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 thống kê kiểm định:

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

T có phân phối Student với $n-2$ bậc tự do.

Kiểm định giả thuyết cho hệ số tương quan

Các bước kiểm định

4. Xác định miền bác bỏ

Đối thuyết	**Miền bác bỏ**	$p-$ **giá trị**
$H_1 : \rho \neq 0$	$\lvert t\rvert > t^{n-2}_{\alpha/2}$	$p = 2\mathbb{P}(T^{n-2} \geq \lvert t\rvert)$
$H_1 : \rho < 0$	$t < -t^{n-2}_{\alpha}$	$p = \mathbb{P}(T^{n-2} \leq t)$
$H_1 : \rho > 0$	$t > t^{n-2}_{\alpha}$	$p = \mathbb{P}(T^{n-2} \geq t)$

5. Kết luận: Bác bỏ H_0 Chưa đủ cơ sở để bác bỏ H_0.

Ví dụ

Ví dụ 3.

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

X	1	1	2	2	3	4	5	5
Y	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) Tìm khoảng tin cậy 95% cho Y nếu liều dùng $x_0 = 4(mg/kg)$.

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 không? $\alpha = 0.05$.