XÁC SUẤT - THỐNG KÊ

CHƯƠNG 8: HỒI QUY TUYẾN TÍNH ĐƠN

TS. Phan Thị Hường

Trường Đại học Bách Khoa TP HCM
Khoa Khoa học ứng dụng, bộ môn Toán ứng dụng
Email: huongphan@hcmut.edu.vn

TP. HCM — 2020.

NỘI DUNG

NỘI DUNG

NỘI DUNG

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,

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,

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,

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,

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,
- . . .

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 phải là biến liên tục,
    - Các biến độc lập $X_1, X_2, \ldots, X_p$ có thể là biến liên tục, 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 2.1

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,* 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:

$$\mathbb{E}[Y|x] = \beta_0 + \beta_1 x,$$

  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.

MÔ HÌNH HỒI QUY TUYẾN TÍNH ĐƠN

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

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

Một mô hình hồi quy tuyến tính đơn cần các giả định:

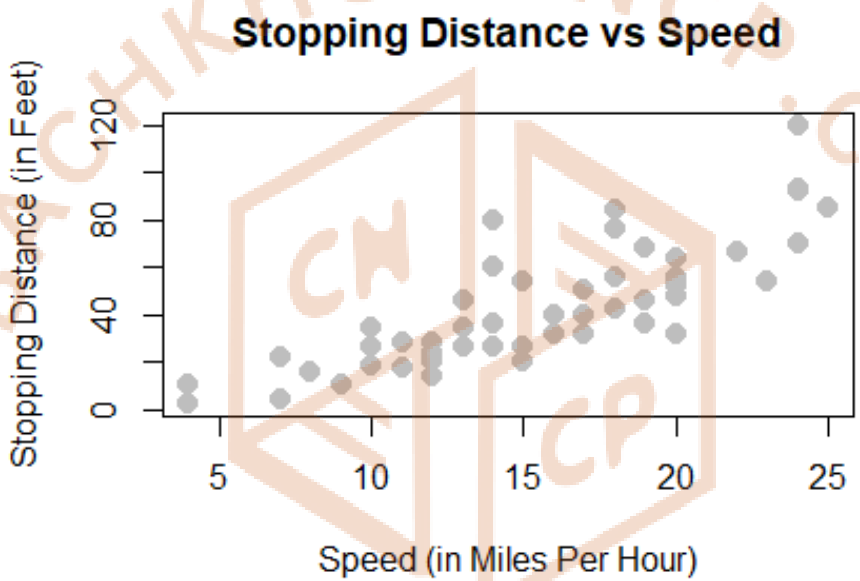
- Các thành phần saia số ε_i là độc lập với nhau.
- $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ or $Y \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2)$.

ĐỒ THỊ PHÂN TÁN CHO DỮ LIỆU CẶP

Dữ liệu `cars` cung cấp 50 giá trị quan trắc của 2 biến `speed`(mph) và `dist` (ft). Biến `speed`(mph) ghi chú tốc độ xe và biến `dist` (ft) mô tả khoảng cách cần thiết trước khi xe dừng.

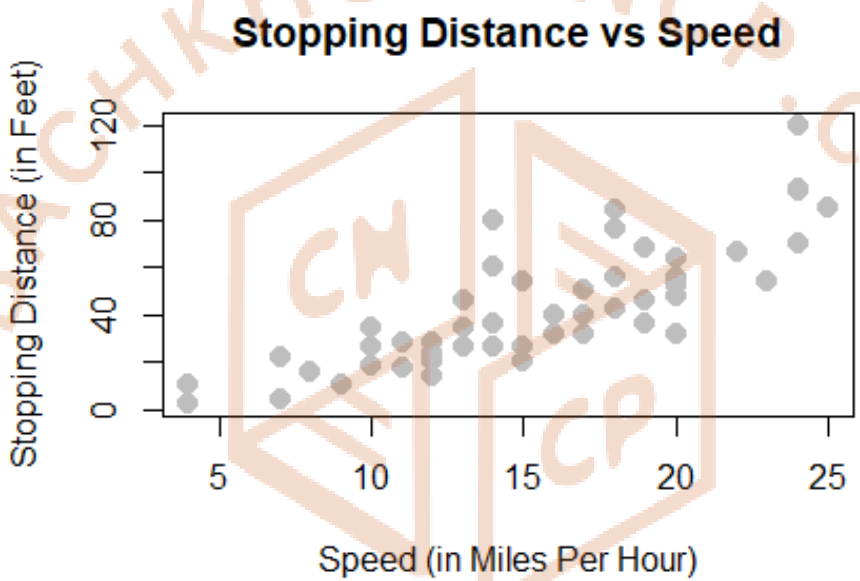
	speed	dist
1	4.00	2.00
2	4.00	10.00
3	7.00	4.00
4	7.00	22.00
5	8.00	16.00
...	...	...
48	24.00	93.00
49	24.00	120.00
50	25.00	85.00

ĐỒ THỊ PHÂN TÁN CHO DỮ LIÊU CẶP



HÌNH: The scatter diagram of the `cars` dataset.

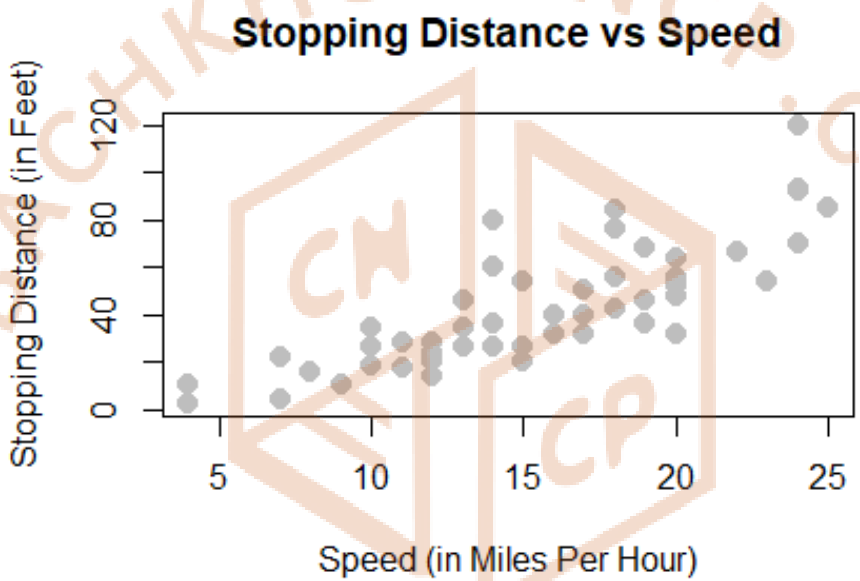
ĐỒ THỊ PHÂN TÁN CHO DỮ LIÊU CẶP



HÌNH: The scatter diagram of the `cars` dataset.

- 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.

ĐỒ THỊ PHÂN TÁN CHO DỮ LIÊU CẶP



HÌNH: The scatter diagram of the `cars` dataset.

• 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. ⇒ Đồ thị phân tán thể cho chúng ta một đánh giá trực quan về sự phù hợp của mô hình hồi quy tuyến tính đơn.

Ước lượng các hệ số hồi quy

ƯỚC LƯỢNG CÁC HỆ SỐ HỒI QUY

- Gọi $\hat{\beta}_0$, $\hat{\beta}_1$ lần lượt là các ước lượng của β_0 và β_1.

ƯỚC LƯỢNG CÁC HỆ SỐ HỒI QUY

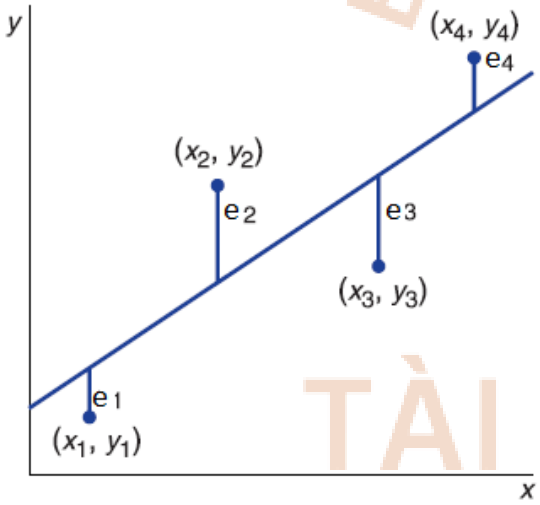
- Gọi $\hat{\beta}_0$, $\hat{\beta}_1$ lần lượt là các ước lượng của β_0 và β_1.
- Đường thẳng hồi quy được định nghĩa bởi

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

ƯỚC LƯỢNG CÁC HỆ SỐ HỒI QUY

- Gọi $\hat{\beta}_0$, $\hat{\beta}_1$ lần lượt là các ước lượng của β_0 và β_1.
- Đường thẳng hồi quy được định nghĩa bởi

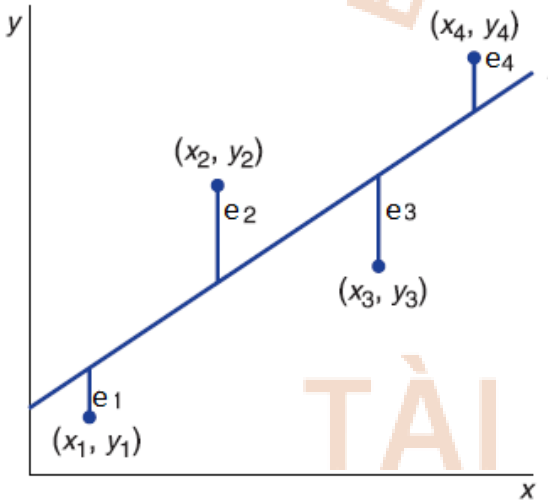
$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$$



ƯỚC LƯỢNG CÁC HỆ SỐ HỒI QUY

- Gọi $\hat{\beta}_0$, $\hat{\beta}_1$ lần lượt là các ước lượng của β_0 và β_1.
- Đường thẳng hồi quy được định nghĩa bởi

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

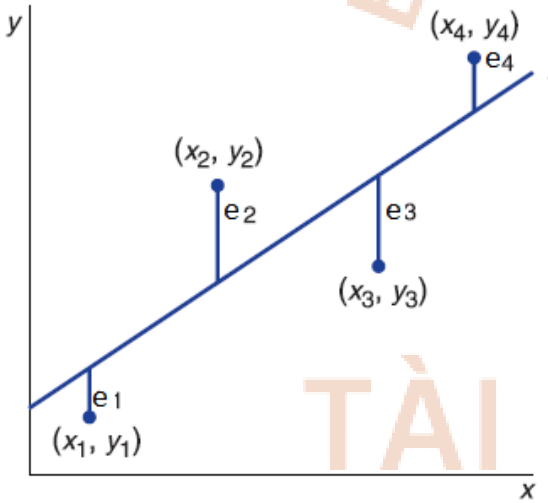


Thành phần sai số của giá trị quan trắc thứ i residual: $e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) = y_i - \hat{y}_i$ mô tả độ sai khác giữa giá trị quan trắc y_i và giá trị dự đoán $\hat{y}_i$.

ƯỚC LƯỢNG CÁC HỆ SỐ HỒI QUY

- Gọi $\hat{\beta}_0$, $\hat{\beta}_1$ lần lượt là các ước lượng của β_0 và β_1.
- Đường thẳng hồi quy được định nghĩa bởi

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



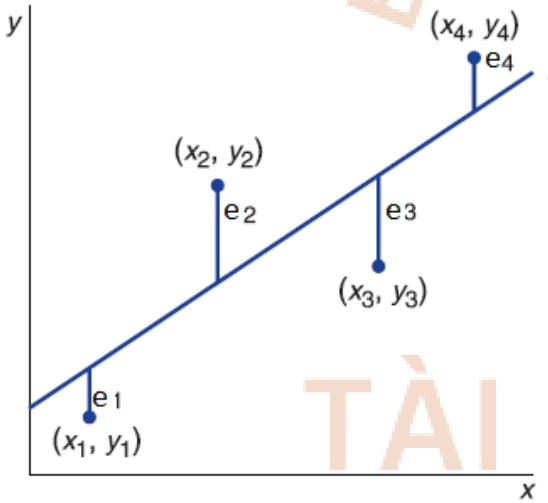
Thành phần sai số của giá trị quan trắc thứ i residual: $e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) = y_i - \hat{y}_i$ mô tả độ sai khác giữa giá trị quan trắc y_i và giá trị dự đoán $\hat{y}_i$.

- Một đường thẳng ước lượng tốt nhất phải "gần các điểm dữ liệu nhất".

ƯỚC LƯỢNG CÁC HỆ SỐ HỒI QUY

- Gọi $\hat{\beta}_0$, $\hat{\beta}_1$ lần lượt là các ước lượng của β_0 và β_1.
- Đường thẳng hồi quy được định nghĩa bởi

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



Thành phần sai số của giá trị quan trắc thứ i residual: $e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) = y_i - \hat{y}_i$ mô tả độ sai khác giữa giá trị quan trắc y_i và giá trị dự đoán $\hat{y}_i$.

- Một đường thẳng ước lượng tốt nhất phải "gần các điểm dữ liệu nhất".
- $\hat{\beta}_0$ và $\hat{\beta}_1$ Được ước lượng từ phương pháp bình phương bé nhất.

PHƯƠNG PHÁP BÌNH PHƯƠNG BÉ NHẤT

ĐỊNH NGHĨA 2.2

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

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right]^2 \tag{3}$$

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 (3), 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

$$\begin{cases} n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases} \quad (4)$$

ƯỚC LƯỢNG BÌNH PHƯƠNG BÉ NHẤT

Giả hệ (4), ta tìm được các ước lượng BPBN của β_0 và β_1 là

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{S_{xy}}{S_{xx}} \tag{5}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{6}$$

với S_{xx} và S_{xy} xác định bởi

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} \tag{7}$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n} \tag{8}$$

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 sai số (Sum of Errors).

Ví dụ

Ví dụ 2.1

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	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) *Tìm đường thẳng hồi quy biểu diễn mối liên hệ giữa trọng lượng cây Y theo diện tích lá X.*

(B) *Tính giá trị sai số tại cặp quan trắc* $(x_i, y_i) = (271, 2.11)$.

(C) *Khi diện tích bề mặt lá là* 400 *thì trọng lượng khô của lá được kỳ vọng là bao nhiêu.*

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

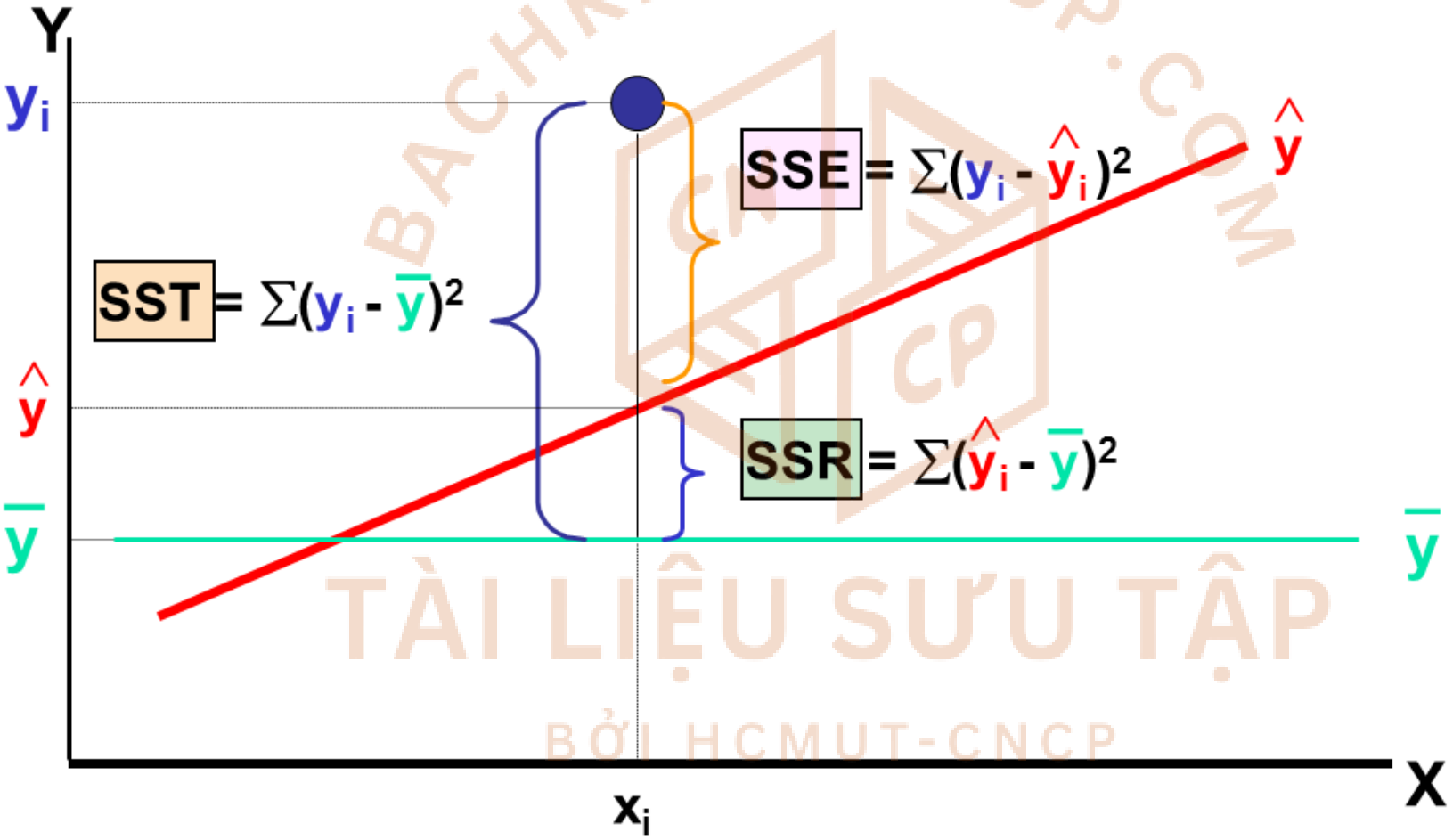
- SST: đo sự biến thiên của các giá trị y_i xung quanh giá trị trung tâm của dữ liệu $\bar{y}$,
- SSR: giải thích sự biến thiên liên quan đến mối quan hệ tuyến tính của X và Y,
- SSE: giải thích sự biến thiên của các yếu tố khác (không liên quan đến mối quan hệ tuyến tính của X và Y).

Ta có:

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{9}$$

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

ĐỘ ĐO SỰ BIẾN THIÊN CỦA DỮ LIỆU



HỆ SỐ XÁC ĐỊNH

ĐỊNH NGHĨA 2.3

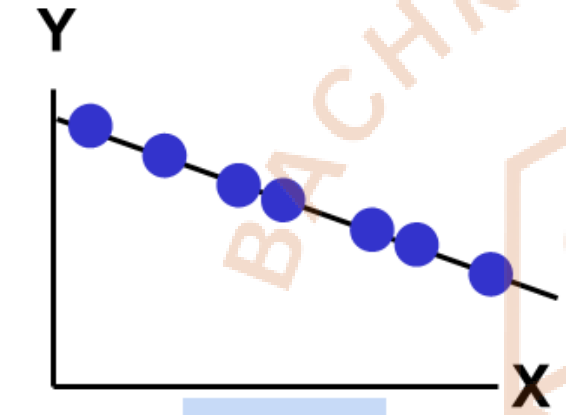
Hệ số xác định (Coefficient of Determination) là tỷ lệ của tổng sự biến thiên trong biến phụ thuộc gây ra bởi sự biến thiên của các biến độc lập (biến giải thích) so với tổng sự biến thiên toàn phần.
Hệ số xác định thường được gọi là R - bình phương (R-squared), ký hiệu là R^2.
Công thức tính:

$$R^2 = \frac{\text{SSR}}{\text{SST}} \tag{10}$$

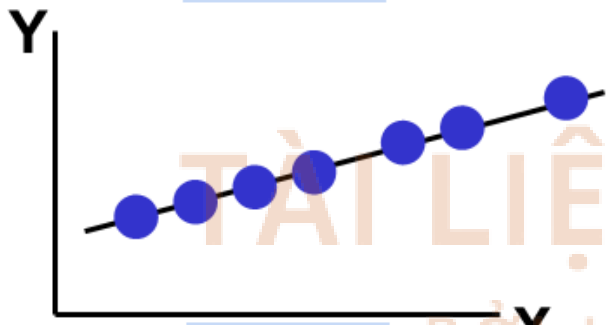
Chú ý: $0 \leq R^2 \leq 1$.

- 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



$R^2 = 1$



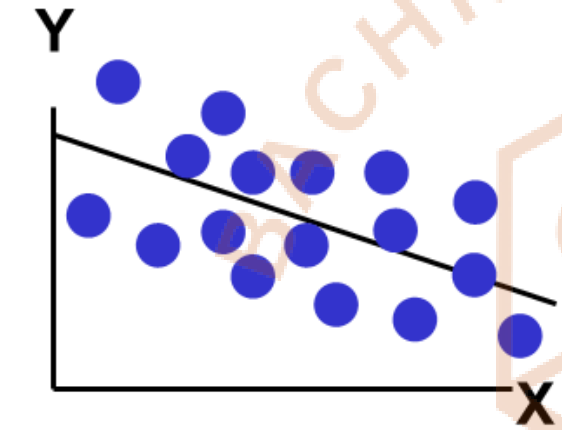
$R^2 = 1$

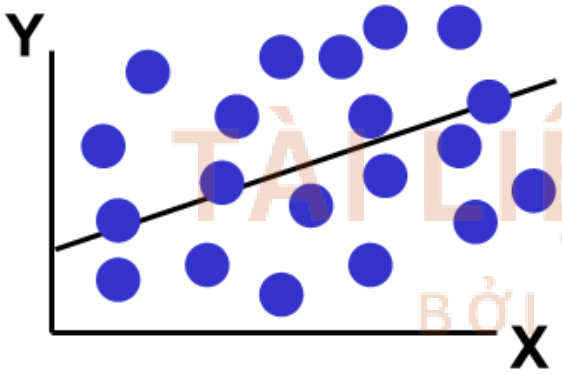
$R^2 = 1$

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



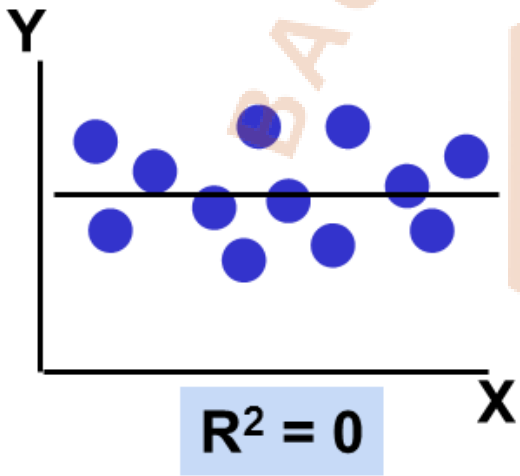


$0 < R^2 < 1$

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 (**??**), 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

$$\hat{\sigma}^2 = \frac{SSE}{n-2} \tag{11}$$

ĐỊNH NGHĨA 2.4

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

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

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:

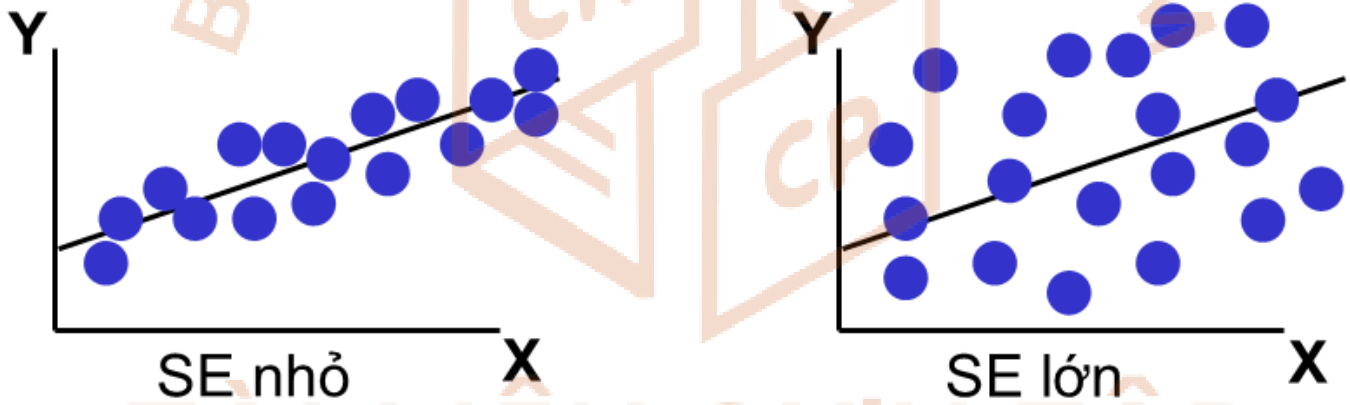
$$SSE = \text{SST} - SSR = \text{SST} - \hat{\beta}_1 S_{xy}$$

- 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

BỔ ĐỀ 2.1

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*

$$\mathbb{E}(\hat{\beta}_0) = \beta_0, \ \mathbb{V}(\hat{\beta}_0) = \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\sigma^2, \tag{12}$$

$$\mathbb{E}(\hat{\beta}_1) = \beta_1, \ \mathbb{V}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}} \tag{13}$$

TÍNH CHẤT CỦA CÁC ƯỚC LƯỢNG BPBN

ĐỊNH NGHĨA 2.5

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

$$\text{SE}(\hat{\beta}_0) = \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\hat{\sigma}^2} \tag{14}$$

$$\text{SE}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \tag{15}$$

KIỂM ĐỊNH CHO CÁC HỆ SỐ HỒI QUY

Xét các bài toán kiểm định $H0: \beta_1 = b1$,

1. Tính giá trị kiểm định thống kê:

$$T_{\beta_1} = \frac{\hat{\beta}_1 - b_1}{SE(\hat{\beta}_1)} \sim t(n-2), \quad SE(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

2.

Giả thuyết đối	**Miền bác bỏ**	p **- value**
$H_1: \beta_1 \neq b_1$	$\lvert t_{\beta_1} \rvert > t^{n-2}_{\alpha/2}$	$p = 2\mathbb{P}(T_{n-2} \geq \lvert t_{\beta_0} \rvert)$
$H_1: \beta_1 < b_1$	$t_{\beta_1} < -t^{n-2}_{\alpha}$	$p = \mathbb{P}(T_{n-2} \leq t_{\beta_0})$
$H_1: \beta_1 > b_1$	$t_{\beta_1} > t^{n-2}_{\alpha}$	$p = \mathbb{P}(T_{n-2} \geq t_{\beta_0})$

$\Rightarrow 100(1-\alpha)\%$ **KTC cho** β_1**:**

$$\hat{\beta}_1 - t^{n-2}_{\alpha/2}\sqrt{\frac{\hat{\sigma}}{S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t^{n-2}_{\alpha/2}\sqrt{\frac{\hat{\sigma}}{S_{xx}}}$$

KIỂM ĐỊNH CHO CÁC HỆ SỐ HỒI QUY

Xét các bài toán kiểm định $H0: \beta_0 = b0$,

1. Tính giá trị kiểm định thống kê:

$$T_{\beta_0} = \frac{\hat{\beta}_0 - b_0}{SE(\hat{\beta}_0)} \sim t(n-2),\ SE(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}$$

2.

Giả thuyết đối	**Miền bác bỏ**	p **- value**
$H_1: \beta_0 \neq b_0$	$\lvert t_{\beta_1}\rvert > t^{n-2}_{\alpha/2}$	$p = 2\mathbb{P}(T_{n-2} \geq \lvert t_{\beta_0}\rvert)$
$H_1: \beta_0 < b_0$	$t_{\beta_1} < -t^{n-2}_{\alpha}$	$p = \mathbb{P}(T_{n-2} \leq t_{\beta_0})$
$H_1: \beta_0 > b_0$	$t_{\beta_1} > t^{n-2}_{\alpha}$	$p = \mathbb{P}(T_{n-2} \geq t_{\beta_0})$

$\Rightarrow 100(1-\alpha)\%$ **KTC cho** β_0**:**

$$\hat{\beta}_0 - t^{n-2}_{\alpha/2}\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\hat{\sigma}^2} \leq \beta_0 \leq \hat{\beta}_0 + t^{n-2}_{\alpha/2}\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\hat{\sigma}^2}$$

Ví dụ

Ví dụ 2.2

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.

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 3.1

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

$$Cov(X,Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \qquad (16)$$

PHÂN TÍCH TƯƠNG QUAN

ĐỊNH NGHĨA 3.2

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

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{\mathbb{V}ar(X)\mathbb{V}ar(Y)}} \tag{17}$$

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

PHÂN TÍCH TƯƠNG QUAN

ĐỊNH NGHĨA 3.3

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

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx} SST}} \tag{18}$$

PHÂN TÍCH TƯƠNG QUAN

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

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

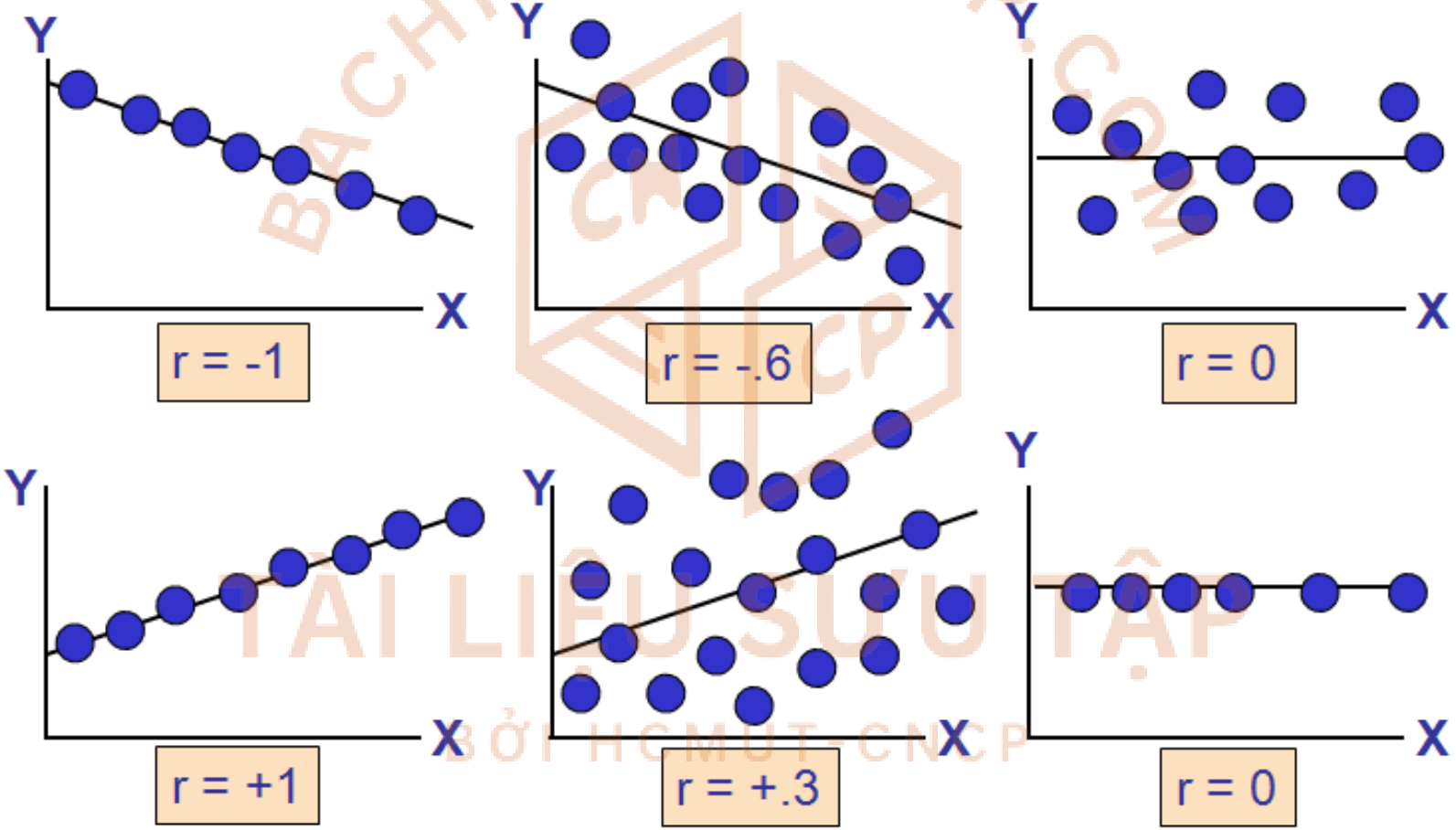
ĐÁNH GIÁ HIỆP PHƯƠNG SAI

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



VÍ DỤ

VÍ DỤ 3.1

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	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.

(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?* $\alpha = 0.05$.