



Chương 8. Hồi quy tuyến tính

Nguyễn Minh Trí

Trường Đại học Công nghệ Thông tin
Đại học Quốc gia Thành phố Hồ Chí Minh



Mục tiêu

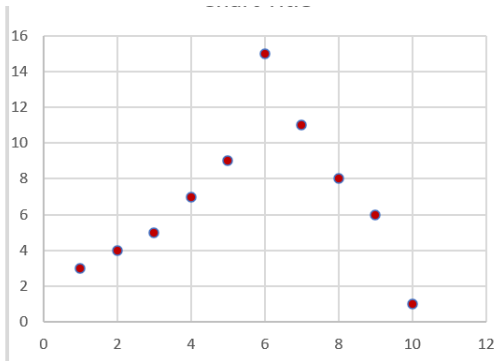
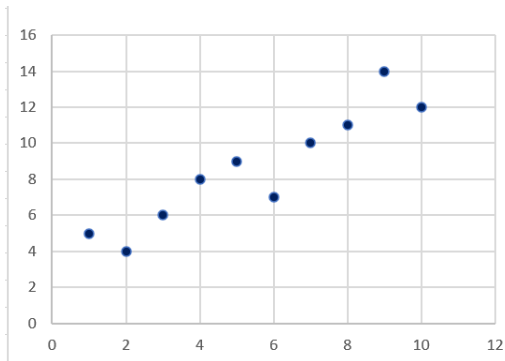
- Tính hệ số tương quan để xác định độ mạnh của quan hệ tuyến tính giữa hai biến
- Xác định đường thẳng hồi quy tuyến tính bằng phương pháp bình phương cực tiểu.
- Dự đoán các giá trị mới của biến

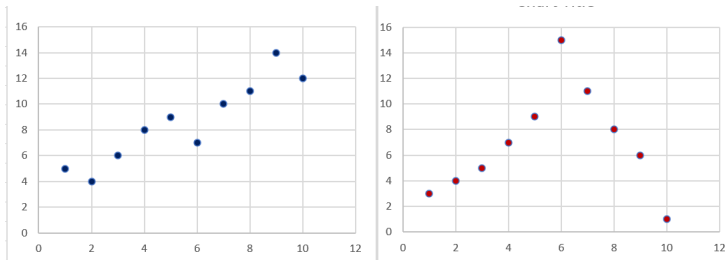


8.1 Hệ số tương quan

8.1 Hệ số tương quan

Cho hai biến (X, Y) lần lượt nhận các giá trị $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. Biểu diễn các điểm giá trị trên mặt phẳng, ta được một biểu đồ mà ta gọi là **biểu đồ phân tán** (scatter diagram). Mỗi điểm (x_i, y_i) được gọi là một điểm dữ liệu.





Nhận xét:

- Hình bên trái: Các điểm dữ liệu được phân bố xung quanh một đường thẳng.
- Hình bên phải: Các điểm dữ liệu không được phân bố xung quanh một đường thẳng.

Ta nói hai biến của hình bên trái có quan hệ tuyến tính mạnh.



8.1 Hệ số tương quan

Định nghĩa 8.1 Số đo xác định độ mạnh của quan hệ tuyến tính giữa hai biến được gọi là hệ số tương quan (correlation coefficient).

- Hệ số tương quan của hai biến X, Y

$$r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{(\overline{x^2} - \bar{x}^2)(\overline{y^2} - \bar{y}^2)}}$$

hay

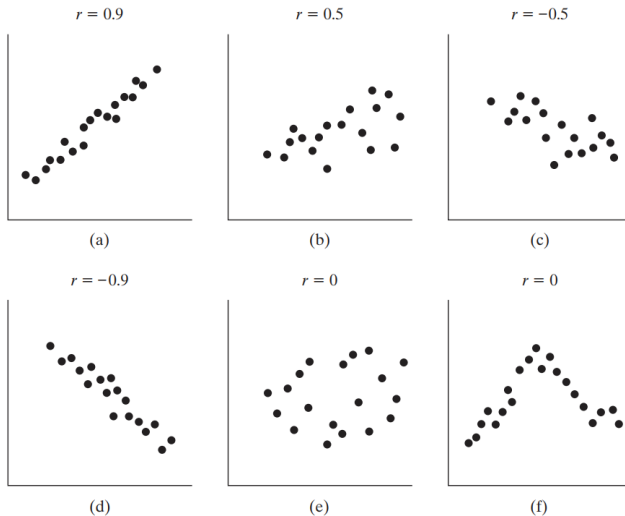
$$r = \frac{n(\sum_{i=1}^n x_i y_i) - (\sum_{i=1}^n x_i) \cdot (\sum_{i=1}^n y_i)}{\sqrt{(n(\sum_{i=1}^n x_i^2) - (\sum_{i=1}^n x_i)^2)(n(\sum_{i=1}^n y_i^2) - (\sum_{i=1}^n y_i)^2)}}$$

trong đó n là số cặp điểm dữ liệu.



8.1 Hệ số tương quan

- Ta có $-1 \leq r \leq 1$.
- Nếu $0,8 \leq |r| \leq 1$ thì ta nói X, Y có tương quan tuyến tính mạnh.
- Nếu $|r| < 0,8$ thì ta nói X, Y có tương quan tuyến tính yếu.
- Nếu r gần bằng 1 thì ta nói có sự tương quan tuyến tính thuận giữa X và Y , tức là nếu X tăng thì Y tăng.
- Nếu r gần bằng -1 thì ta nói có sự tương quan tuyến tính nghịch giữa X và Y , tức là nếu X tăng thì Y giảm.



Ví dụ 8.2 Điểm số môn Xác suất thống kê và số buổi vắng của 7 sinh viên được cho bên dưới

Số buổi vắng (X)	6	2	15	9	12	5	8
Điểm (Y)	8,2	8,6	4,3	7,4	5,8	9,0	7,8

Tìm hệ số tương quan giữa số buổi nghỉ học và điểm môn Xác suất thống kê.

Giải. Ta có

$$\overline{xy} = \frac{6.8,2 + 2.8,6 + 15.4,3 + 9.7,4 + 12.5,8 + 5.9,0 + 8.7,8}{7} = 53,5$$

$$\bar{x} = \frac{6 + 2 + 15 + 9 + 12 + 5 + 8}{7} = 8,14$$

$$\bar{y} = \frac{8,2 + 8,6 + 4,3 + 7,4 + 5,8 + 9,0 + 7,8}{7} = 7,3$$



$$\overline{x^2} = \frac{6^2 + 2^2 + 15^2 + 9^2 + 12^2 + 5^2 + 8^2}{7} = 82,71$$
$$\overline{y^2} = \frac{8,2^2 + 8,6^2 + 4,3^2 + 7,4^2 + 5,8^2 + 9,0^2 + 7,8^2}{7} = 55,7$$

Do đó, hệ số tương quan là

$$r = \frac{53,5 - 8,14 \cdot 7,3}{\sqrt{(82,71 - 8,14^2)(55,7 - 7,3^2)}} = \frac{-5,992}{6,296} = -0,9517.$$

Có một sự tương quan tuyến tính mạnh giữa số buổi vắng và số điểm. Nếu số buổi vắng càng nhiều thì số điểm càng thấp.



Xác định hệ số tương quan bằng R

```
x <- c(73,71,75,72,72,75,67,69,71,69)
y <- c(185,175,200,210,190,195,150,170,180,175)
cor(x,y)
```



Ví dụ 8.3 Trong các giá trị của hệ số tương quan dưới đây, giá trị nào thể hiện sự tương quan tuyến tính mạnh nhất?

- A. $r = 0,8989$
- B. $r = 0,9$
- C. $r = 0,0989$
- D. $r = 0,009$
- E. Tất cả đều sai.

Đáp án: B



Ví dụ 8.4 Trong các giá trị dưới đây, giá trị nào không thể là hệ số tương quan của hai biến?

- A. $r = 0,19$
- B. $r = 0,91$
- C. $r = -0,9$
- D. $r = 1$
- E. $r = 1,9$

Đáp án: E



Ví dụ 8.5 Giả sử hệ số tương quan của hai biến X và Y là $r = -0,97$. Hãy chọn phát biểu đúng

- A. Nếu X tăng thì Y tăng.
- B. Nếu X tăng thì Y giảm.
- C. X và Y không có quan hệ tuyến tính.
- D. Nếu X tăng thì Y không thay đổi.
- E. Không thể xác định mối quan hệ giữa X và Y .

Đáp án: B



8.2 Hồi quy tuyến tính đơn



8.2 Hồi quy tuyến tính đơn

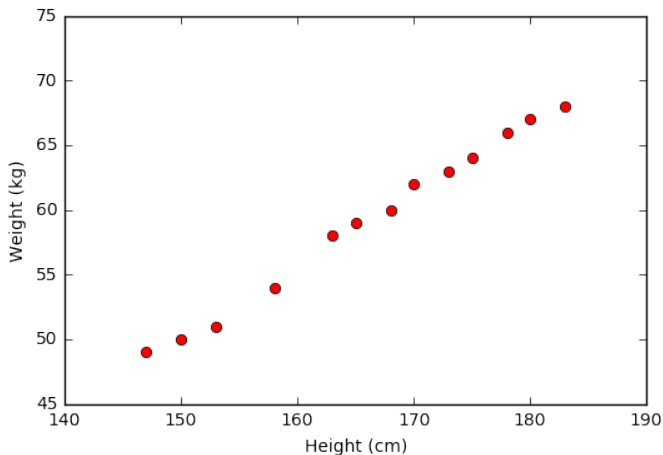
Bài toán. Bảng dữ liệu về chiều cao và cân nặng của 15 người:

Chiều cao (cm)	Cân nặng (kg)	Chiều cao (cm)	Cân nặng (kg)
147	49	168	60
150	50	170	72
153	51	173	63
155	52	175	64
158	54	178	66
160	56	180	67
163	58	183	68
165	59		

Có thể dự đoán cân nặng của một người dựa vào chiều cao của họ không?

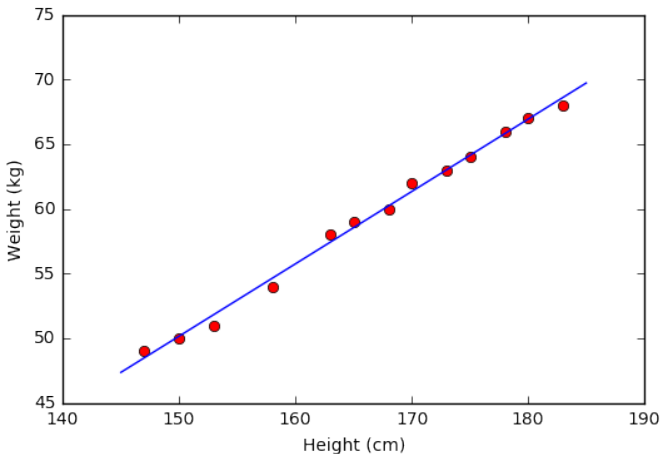
8.2 Hồi quy tuyến tính đơn

Biểu diễn các dữ liệu trên dưới dạng đồ thị như sau



8.2 Hồi quy tuyến tính đơn

Ta thấy rằng các điểm dữ liệu không nằm trên một đường thẳng nhưng chúng có thể được phân bố xung quanh một đường thẳng.





8.2 Hồi quy tuyến tính đơn

- Các điểm dữ liệu nằm khá gần đường thẳng (phương trình $y = a + bx$).
- Ta có thể đưa ra mối liên hệ giữa cân nặng và chiều cao như sau

$$\text{cân nặng} = b \times \text{chiều cao} + a.$$

- Bằng các công cụ tính toán, chúng ta sẽ tính được a, b . Khi đó đường thẳng có phương trình $y = a + bx$ được gọi là **đường thẳng hồi quy** (regression line).
- Sử dụng mô hình này, ta có thể dự đoán cân nặng của một người có chiều cao 155cm, 160 cm hoặc 171cm.
- Mô hình trên là mô hình **hồi quy tuyến tính đơn**.



8.2 Hồi quy tuyến tính đơn

Bài toán. Xây dựng một mô hình toán học hay một hàm số mà có thể dùng để dự đoán giá trị của một biến dựa vào một hay một số biến được gọi là phân tích hồi quy (regression analysis).

- Mô hình hồi quy đơn giản nhất được gọi là hồi quy đơn (simple regression) liên quan đến hai biến trong đó một biến được dự đoán dựa vào một biến khác.
- Trong hồi quy đơn, biến được dự đoán được gọi là biến phụ thuộc (dependent variable) và ký hiệu là y . Biến mà ta dùng để dự đoán biến y được gọi là biến độc lập (independent variable/regressor variable/predictor variable) và ký hiệu là x .
- Giả sử mối quan hệ giữa biến độc lập và biến phụ thuộc được cho bởi một hàm số $y = f(x)$ trong đó f là hàm số mà ta chưa biết.
- Nếu $f(x) = a + bx$ thì đường thẳng được xác định bởi $y = a + bx$ được gọi là đường thẳng hồi quy.



8.2 Hồi quy tuyến tính đơn

- Dùng phương pháp bình phương tối thiểu (method of least squares) để tìm các giá trị a và b .
- Giả sử ta có một mẫu gồm n điểm dữ liệu được xác định bởi các cặp giá trị $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$.
- Tìm một mô hình đường thẳng

$$y = a + bx$$

sao cho các điểm dữ liệu là "gần nhất" với đường thẳng này.

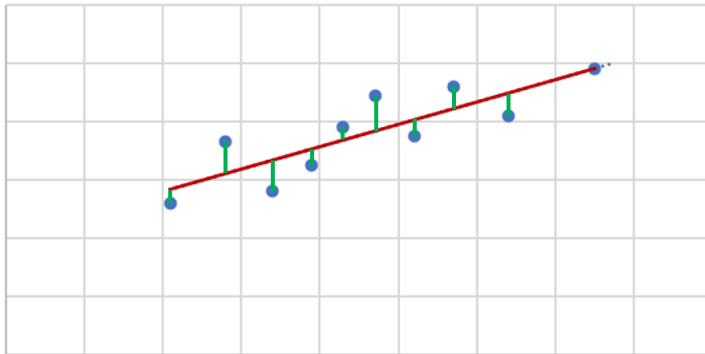
- Tập hợp các điểm (x_i, y_i) được biểu diễn trên mặt phẳng tọa độ được gọi là **biểu đồ phân tán** (Scatter diagram).
- Từ biểu đồ phân tán sẽ cho ta dự đoán được hàm số $f(x)$.

Với mỗi điểm (x_i, y_i) , ta có

$$\hat{y}_i = a + bx_i$$

và

$$y_i - \hat{y}_i = y_i - (a + bx_i)$$





- Khi đó tổng bình phương của các độ lệch giữa giá trị của y_i và \hat{y}_i được kí hiệu

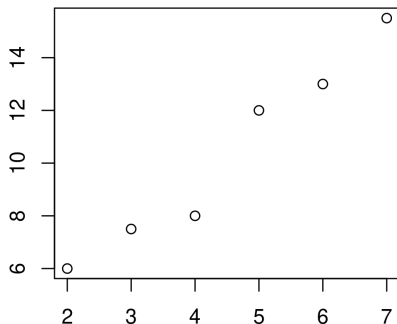
$$L = \sum_{i=1}^n [y_i - (a + bx_i)]^2$$

- Đường thẳng $y = a + bx$ được gọi là "gần nhất" với các điểm dữ liệu đã cho nếu L có giá trị nhỏ nhất.
- Khi đó

$$b = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} \text{ và } a = \bar{y} - b\bar{x}.$$

Ví dụ 8.6 Tìm đường thẳng hồi quy biểu thị mối liên hệ giữa số tiền lương làm theo giờ y (trăm nghìn đồng) và số năm kinh nghiệm x dựa theo bảng dữ liệu sau

x	2	3	4	5	6	7
y	6	7,5	8	12	13	15,5





x	2	3	4	5	6	7
y	6	7,5	8	12	13	15,5

Giải. Dựa vào dữ liệu, ta tính được các giá trị

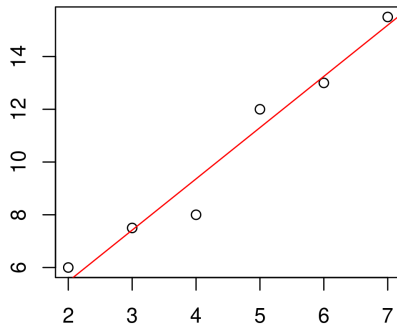
$$\bar{x} = 4,5 \quad \bar{y} = 10,33 \quad b = 1,943$$

và

$$a = \bar{y} - b\bar{x} = 10,33 - 1,943.4,5 = 1,5865$$

Như vậy, phương trình đường thẳng hồi quy tuyến tính là

$$y = 1,5865 + 1,943x$$





Vẽ biểu đồ phân tán và đường thẳng hồi quy bằng R


```
x <- c(2,3,4,5,6,7)
y <- c(6, 7.5, 8, 12, 13, 15.5)
plot(x, y)
abline(lm(y ~ x),col="red")
```

Tìm các hệ số a, b của đường thẳng hồi quy $y = a + bx$ bằng R

```
x <- c(2,3,4,5,6,7)
y <- c(6, 7.5, 8, 12, 13, 15.5)
lm(formula = y ~ x)
```



Dùng Microsoft Excel để tìm đường thẳng hồi quy

- Tạo bảng dữ liệu trong Microsoft Excel
- Tạo biểu đồ phân tán: Chọn bảng dữ liệu → **Insert** → **Charts** → **All Charts** → **X Y (Scatter)** → **OK**
- Tạo đường thẳng hồi quy: Nhấp vào  bên góc phải của Chart vừa hiện ra → **Chart Elements**, chọn **Trendline**
- Hiện phương trình đường thẳng hồi quy: Bên cạnh **Trendline** → ► **More Options**
- Trong bảng **Format Trendline**, chọn , kéo xuống bên dưới và chọn **Display Equation on chart**.

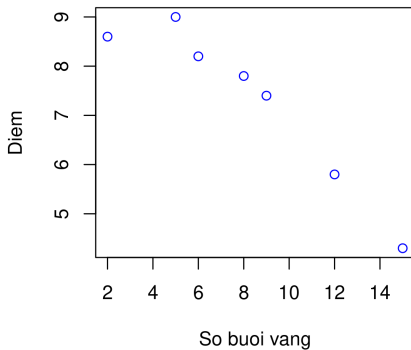
Ví dụ 8.7 Điểm số môn Xác suất thống kê và số buổi vắng của 7 sinh viên được cho bên dưới

Số buổi vắng (x)	6	2	15	9	12	5	8
Điểm (y)	8,2	8,6	4,3	7,4	5,8	9,0	7,8

- Vẽ biểu đồ phân tán thể hiện dữ liệu đã cho.
- Tìm phương trình đường thẳng hồi quy tuyến tính thể hiện mối liên hệ giữa số điểm số và số buổi vắng học.
- Dự đoán số điểm số của sinh viên chỉ vắng 1 buổi học.

Số buổi vắng (x)	6	2	15	9	12	5	8
Điểm (y)	8,2	8,6	4,3	7,4	5,8	9,0	7,8

Giải. a. Biểu đồ phân tán





b. Dựa vào dữ liệu, ta tính được các giá trị

$$\bar{x} = 8,142857 \quad \bar{y} = 7,3 \quad b = -0,3622$$

và

$$a = \bar{y} - b\bar{x} = 7,3 - (-0,3622) \cdot 8,142857 = 10,2493$$

Như vậy, phương trình đường thẳng hồi quy tuyến tính là

$$y = 10,2493 - 0,3622x$$

c. Nếu $x = 1$ thì $y = 9,8871$. Do đó nếu sinh viên vắng một buổi học thì điểm số của sinh viên có thể đạt được là 9,8871 điểm.

Ví dụ 8.8 Bảng khảo sát doanh thu bán hàng online Y và chi phí quảng cáo online X (trong 15 phút) của 7 cửa hàng được cho như sau: Đơn vị tính là trăm nghìn đồng

Doanh số bán hàng	368	340	665	954	331	556	376
Chi phí quảng cáo	1,7	1,5	2,8	5	1,3	2,2	1,3

- Tính hệ số tương quan và nhận xét về tính tuyến tính của X và Y (mạnh hay yếu).
- Viết phương trình hồi quy tuyến tính của Y theo X . Dự đoán doanh số bán hàng khi chi phí quảng cáo online trong 15 phút là 4 trăm nghìn đồng.

Giải. a. Hệ số tương quan

$$r = 0,9804402$$

Từ đó ta thấy doanh số bán hàng và chi phí quảng cáo có tương quan tuyến tính mạnh.



b. Đặt x là chi phí quảng cáo và y doanh thu bán hàng. Từ dữ liệu đã có, ta tính được

$$\bar{y} = 512,8571; \quad \bar{x} = 2,257143; \text{ và } b = 171,5$$

và

$$a = \bar{y} - b\bar{x} = 125,8$$

Phương trình đường thẳng hồi quy tuyến tính

$$y = 125,8 + 171,5x.$$

Như vậy, khi $x = 4$ thì $y = 811,8$. Tức là nếu chi phí quảng cáo trong 15 phút là 4 trăm nghìn đồng thì doanh thu bán hàng đạt được là 81 180 000 đồng.



Bài tập

Bài 8.1 Cho hai biến x, y có các giá trị tương ứng như sau

x	2,4	2,7	5,6	2,6	2,1	3,3	6,6	5,7
y	25,3	14,3	151,6	91,1	80	49	173	95,8

Hai biến x, y có quan hệ tuyến tính không?

Bài 8.2 Lợi nhuận của 7 công ty cho thuê xe là Y (tỉ USD) trong 1 năm và số lượng xe cho thuê X (nghìn chiếc) được cho như sau

Công ty	Số xe (X)	Lợi nhuận (Y) (tỉ USD)
A	630	7
B	290	3,9
C	208	2,1
D	191	2,8
E	134	1,4
F	85	1,5



- a. Tính hệ số tương quan giữa số xe cho thuê và lợi nhuận hàng năm.
- b. Viết phương trình hồi quy tuyến tính của Y theo X . Dự đoán lợi nhuận trong một năm của một công ty có 200 000 xe cho thuê.

Bài 8.3 Thời gian sử dụng liên tục của 8 loại điện thoại Y (giờ) và số mAh X ghi trên pin của điện thoại được khảo sát như sau

Điện thoại	Số mAh (X)	Thời gian sử dụng (Y) (giờ)
A	2800	3,8
B	3000	3,9
C	3700	4,2
D	4000	3,8
E	4300	4,1
F	5000	5
G	5000	4.8
H	6000	4,9

- a. Tính hệ số tương quan giữa số mAh trên pin và thời gian sử dụng.
- b. Viết phương trình hồi quy tuyến tính của Y theo X . Dự đoán thời gian sử dụng của một loại pin điện thoại có 6550 mAh.