

Lý thuyết Thống kê (tiếp theo)



Tương quan và hồi quy tuyến tính



Trong Machine Learning, một trong những thuật toán quan trọng nhất là Thuật toán Hồi quy tuyến tính (Linear Regression) thuộc nhóm *Học có giám sát* (Supervised Learning).

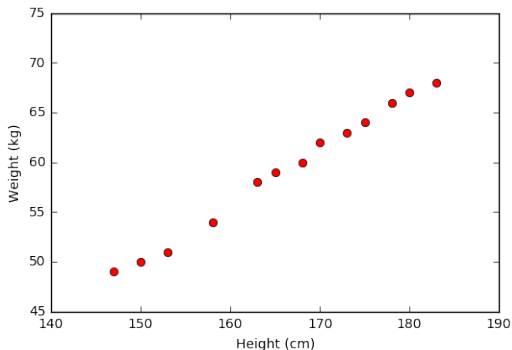
Bài toán

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?

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



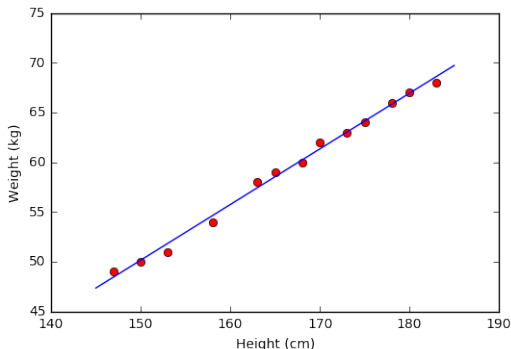
Ta thấy rằng dữ liệu được sắp xếp gần như theo một đường thẳng. Do đó mô hình Hồi quy tuyến tính (Linear Regression) nhiều khả năng sẽ cho kết quả tốt. Ta có thể đưa ra mối liên hệ giữa cân nặng và chiều cao như sau

$$\text{cân nặng} = B \times \text{chiều cao} + A.$$

Bằng các công cụ tính toán, chúng ta sẽ tính được A, B .



Khi đó, các điểm dữ liệu nằm khá gần đường thẳng mà ta dự đoán.



Sử dụng mô hình này, ta có thể dự đoán cân nặng của một người có chiều cao 155cm, 160 cm hoặc 171cm.



Hệ số tương quan

Xét vectơ ngẫu nhiên (X, Y) và tập n giá trị cụ thể

$$(x_1, y_1), \dots, (x_n, y_n).$$

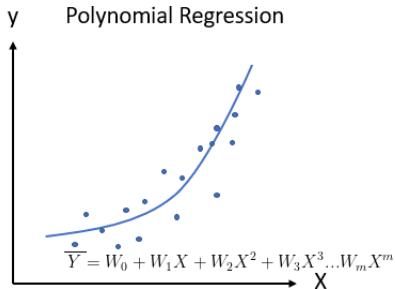
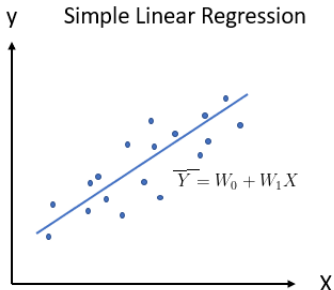
Các cặp giá trị này được gọi là **dữ liệu thực nghiệm**.

Có nhiều kiểu phụ thuộc giữa hai biến ngẫu nhiên X và Y nhưng phổ biến nhất là dạng phụ thuộc hàm số $Y = f(X)$. Một trong những hàm đơn giản nhất là hàm số bậc nhất $Y = aX + b$ hay dạng tuyến tính.

Đường cong phù hợp là một đường cong xấp xỉ tốt nhất (ít sai lệch nhất) với các điểm dữ liệu đã cho.

- Nếu đường cong phù hợp là một đường thẳng thì ta có một **quan hệ tuyến tính** (linear relation) giữa hai biến ngẫu nhiên.
- Nếu đường cong phù hợp **không** là một đường thẳng thì ta có một **quan hệ phi tuyến tính** giữa hai biến ngẫu nhiên.





Bài toán

- ❶ Có một quan hệ tuyến tính hoặc phi tuyến tính giữa hai biến ngẫu nhiên không?
- ❷ Nếu có một quan hệ tuyến tính (phi tuyến tính) giữa hai biến ngẫu nhiên thì có thể biểu diễn mối quan hệ này dưới dạng một hàm số không?

Ta cần một số đo để đo mức độ chặt chẽ trong quan hệ tuyến tính giữa hai biến ngẫu nhiên.

Định nghĩa (Hệ số tương quan - Correlation coefficient)

Hệ số tương quan mẫu của hai biến ngẫu nhiên X, Y

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

Hay

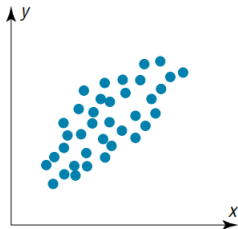
$$r = \frac{n(\sum_{i=1}^n x_i y_i) - (\sum_{i=1}^n x_i) \cdot (\sum_{i=1}^n y_i)}{\sqrt{(n(\sum_{i=1}^n x_i^2) - (\sum_{i=1}^n x_i)^2)(n(\sum_{i=1}^n y_i^2) - (\sum_{i=1}^n y_i)^2)}}$$

trong đó n là số cặp điểm dữ liệu thực nghiệm.

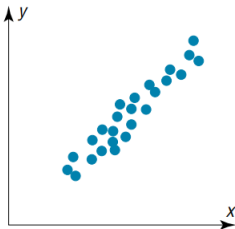


- Ta có $-1 \leq r \leq 1$.
- Hệ số tương quan là một con số đo mức độ phụ thuộc tuyến tính giữa hai biến ngẫu nhiên.
- Nếu $0,8 \leq |r| \leq 1$ thì ta nói X, Y có tương quan tuyến tính mạnh.
- Nếu $|r| < 0,8$ thì ta nói X, Y có tương quan tuyến tính yếu.
- Nếu r gần bằng 1 thì ta nói có sự tương quan tuyến tính thuận giữa X và Y .
- Nếu r gần bằng -1 thì ta nói có sự tương quan tuyến tính nghịch giữa X và Y .

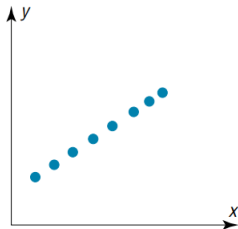




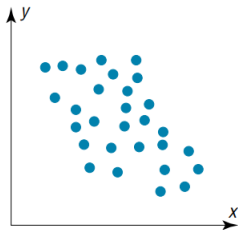
(a) $r = 0.50$



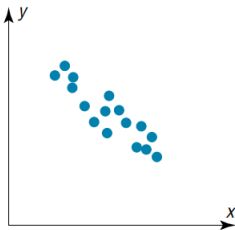
(b) $r = 0.90$



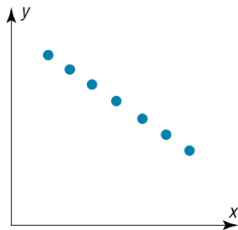
(c) $r = 1.00$



(d) $r = -0.50$



(e) $r = -0.90$



(f) $r = -1.00$



Ví dụ Điểm số môn Đại số tuyến tính 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 Đại số tuyến tính.



Ví dụ Điểm số môn Đại số tuyến tính 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 Đại số tuyến tính.

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

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

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

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

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



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.



Bài toán

Ta muốn khảo sát xem số buổi nghỉ học có ảnh hưởng đến điểm thi cuối kỳ của môn xác suất thống kê. Nếu biết số buổi nghỉ học thì ta có thể dự đoán điểm thi cuối kỳ được không?

- Mục đích của hồi quy là dự đoán một đại lượng này từ các đại lượng khác.
- Nếu biến Y được ước lượng từ biến X bằng một biểu thức $Y = f(X)$ thì biểu thức này được gọi là **phương trình hồi quy** của Y theo X .
- Đường cong biểu diễn đường $Y = f(X)$ được gọi là **đường cong hồi quy** của Y theo X .
- Đường thẳng biểu diễn đường $Y = A + BX$ (phương trình hồi quy tuyến tính) được gọi là **đường thẳng hồi quy** của Y theo X .



Trong việc nghiên cứu mối liên hệ giữa hai biến:

- 1 Thu thập dữ liệu và xây dựng biểu đồ phân tán
- 2 Tính hệ số tương quan r
- 3 Kiểm tra sự tương quan tuyến tính giữa hai biến
- 4 Nếu $|r|$ gần bằng 1 thì ta sẽ xác định đường thẳng hồi quy (regression line) (đường thẳng phù hợp nhất).
- 5 Đường thẳng hồi quy giúp các nhà nghiên cứu có thể nhìn thấy xu hướng và đưa ra các dự báo.



Trong việc nghiên cứu mối liên hệ giữa hai biến:

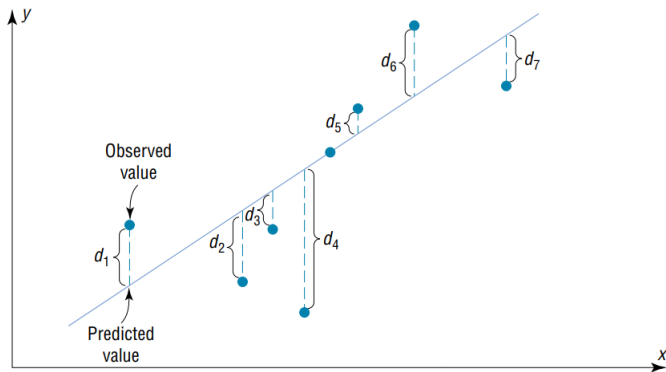
- 1 Thu thập dữ liệu và xây dựng biểu đồ phân tán
- 2 Tính hệ số tương quan r
- 3 Kiểm tra sự tương quan tuyến tính giữa hai biến
- 4 Nếu $|r|$ gần bằng 1 thì ta sẽ xác định đường thẳng hồi quy (regression line) (đường thẳng phù hợp nhất).
- 5 Đường thẳng hồi quy giúp các nhà nghiên cứu có thể nhìn thấy xu hướng và đưa ra các dự báo.

Cho các điểm $(x_1, y_1), \dots, (x_n, y_n)$, ta sẽ tìm phương trình đường thẳng $Y = A + BX$, sao cho

$$\sum_{i=1}^n (y_i - (A + Bx_i))^2$$

là nhỏ nhất. Phương pháp trên được gọi là phương pháp *bình phương cực tiểu* (method of least squares).





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

Điểm số môn Đại số tuyến tính 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 phương trình đường thẳng hồi quy tuyến tính và dự đoán điểm của sinh viên chỉ vắng 1 buổi học.

Giải.

- Phương trình hồi quy tuyến tính $Y = A + BX$.

- $B = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} = \dots -0,362$

- $A = \bar{y} - B\bar{x} = \dots 10,25$

- Phương trình đường thẳng hồi quy tuyến tính cần tìm là $Y = 10,25 - 0,362X$

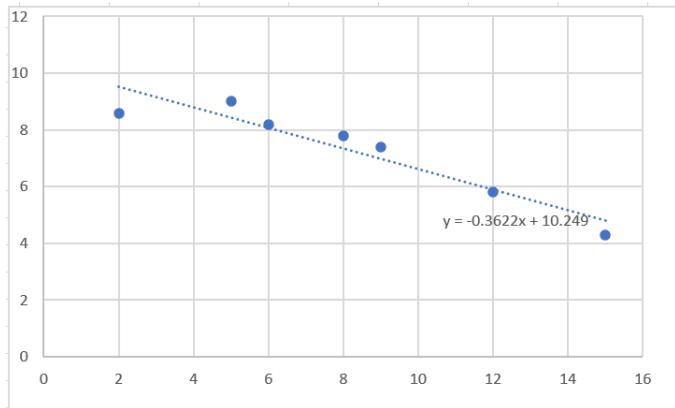
- Khi $X = 1$ thì $Y = 9,888$

Giải. Dùng máy tính CASIO fx-570VN-PLUS

- $\boxed{SHIFT} \rightarrow \boxed{MODE} \rightarrow \nabla \rightarrow$ chọn STAT (trên màn hình - phím 4)
- Màn hình xuất hiện **Frequency**, chọn **OFF**
- $\boxed{SHIFT} \rightarrow \boxed{MODE} \rightarrow$ **Data** (phím $\boxed{2}$)
- Nhập dữ liệu cột X : $\boxed{6} \boxed{=} \boxed{2} \boxed{=} \dots$
- Nhập dữ liệu cột Y : $\boxed{8.2} \boxed{=} \boxed{8.6} \boxed{=} \dots$
- \boxed{ON}
- $\boxed{SHIFT} \rightarrow \boxed{1} \rightarrow$ **Reg** (phím $\boxed{5}$)
- Chọn **A** (phím $\boxed{1}$) $\boxed{=}$
- \boxed{ON}
- $\boxed{SHIFT} \rightarrow \boxed{1} \rightarrow$ **Reg** (phím $\boxed{5}$)
- Chọn **B** (phím $\boxed{2}$) $\boxed{=}$

Khi đó $A = 10,2493$ và $B = -0,3722$. Đường thẳng hồi quy tuyến tính là $Y = 10,2493 - 0,3622X$.







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.



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**.



Một vài lưu ý

- Đường thẳng hồi quy tuyến tính theo phương pháp bình phương tối thiểu luôn đi qua điểm (\bar{x}, \bar{y})
- Khi tính toán cần xác định rõ biến độc lập và biến phụ thuộc
 - Phương trình hồi quy tuyến tính của Y theo X

$$Y = A + BX$$

- Phương trình hồi quy tuyến tính của X theo Y

$$X = A + BY$$



Bài tập. Bảng khảo sát doanh thu bán hàng online Y và chi phí quảng cáo online X (trong thời gian 15 phút) của 7 cửa hàng được cho như sau: Đơn vị tính là đô la

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 (trong 15 phút) khi chi phí quảng cáo online trong 15 phút là 4 đô la.





