

# Phân tích tương quan và Hồi quy

PGS.TS. Lê Sỹ Vinh  
Khoa CNTT – Đại học Công Nghệ

# Phân tích tương quan

Một công ty quan tâm tới việc phân tích hiệu quả của việc quảng cáo. Trong thời gian 5 tháng công ty thu được kết quả sau.

Tiền quảng cáo (\$M)	1	2	3	4	5
Doanh thu (\$M)	6	15	20	30	39

Có mối liên hệ giữa tổng số tiền quảng cáo và doanh thu hay không?

# Phân tích tương quan

Thống kê về số buổi đi học (X) và điểm thi cuối kì môn XSTK (Y) từ 20 sinh viên được cho ở bảng dưới.

X	15	14	10	14	15	7	11	9	14	12
Y	10	9	4	8	9	2	6	8	7	8

X	15	13	5	7	11	14	15	10	12	14
Y	10	8	0	4	6	7	8	5	7	9

Có mối liên hệ giữa số buổi đi học và điểm thi cuối kì hay không?

# Hệ số tương quan

Giả sử  $X$  và  $Y$  là 2 ĐLNN, **Hệ số tương quan** đo mức độ phụ thuộc tuyến tính giữa  $X$  và  $Y$

- Công thức hệ số tương quan lý thuyết  $\rho$

$$\rho = \frac{E(X - \mu_X)(Y - \mu_Y)}{\sigma_X \sigma_Y}$$

- $\rho \in [-1; 1]$
- $\rho=0$  thì không có tương quan tuyến tính giữa  $X$  và  $Y$
- $|\rho|$  càng gần 1 thì sự phụ thuộc tuyến tính giữa  $X$  và  $Y$  càng mạnh
- $|\rho| = 1$  thì  $Y$  là một hàm tuyến tính của  $X$

# Ước lượng $\rho$

Với mẫu quan sát  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  của  $(X, Y)$   
hệ số tương quan:

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

# Ví dụ 1

Một công ty quan tâm tới việc phân tích hiệu quả của việc quảng cáo. Trong thời gian 5 tháng công ty thu được kết quả sau. Tính hệ số tương quan giữa tiền quảng cáo và doanh thu.

Tiền quảng cáo (\$M)	1	2	3	4	5
Doanh thu (\$M)	6	15	20	30	39

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

## Ví dụ 2

Thống kê về số buổi đi học (X) và điểm thi cuối kì môn XSTK (Y) từ 20 sinh viên được cho ở bảng dưới. Tính hệ số tương quan giữa số buổi đi học và điểm thi cuối kì môn XSTK.

X	15	14	10	14	15	7	11	9	14	12
Y	10	9	4	8	9	2	6	8	7	8

X	15	13	5	7	11	14	15	10	12	14
Y	10	8	0	4	6	7	8	5	7	9

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

## Ví dụ 3

Thời gian chơi điện tử của sinh viên một ngày (X) và chỉ số IQ (Y) được cho ở bảng dưới. Tính hệ số tương quan giữa X và Y.

Thời gian chơi điện tử	1	2	3	4	5	4	6	3	1
IQ	90	85	92	85	90	82	95	80	85

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



## Ví dụ 4

Số năm hút thuốc lá (X) và tuổi thọ (Y) từ 20 người được cho ở bảng dưới. Tính hệ số tương quan giữa việc hút thuốc lá và tuổi thọ.

X	10	15	10	15	20	5	10	15	20	15
Y	70	65	66	60	50	72	67	60	55	60

X	15	10	5	12	22	14	16	18	30	14
Y	70	72	75	70	52	54	52	50	45	60

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

## Ví dụ 5

Thời gian chơi điện tử của sinh viên một ngày (X) và mức lương ra trường (Y) từ 9 người được cho ở bảng dưới. Tính hệ số tương quan giữa X và Y.

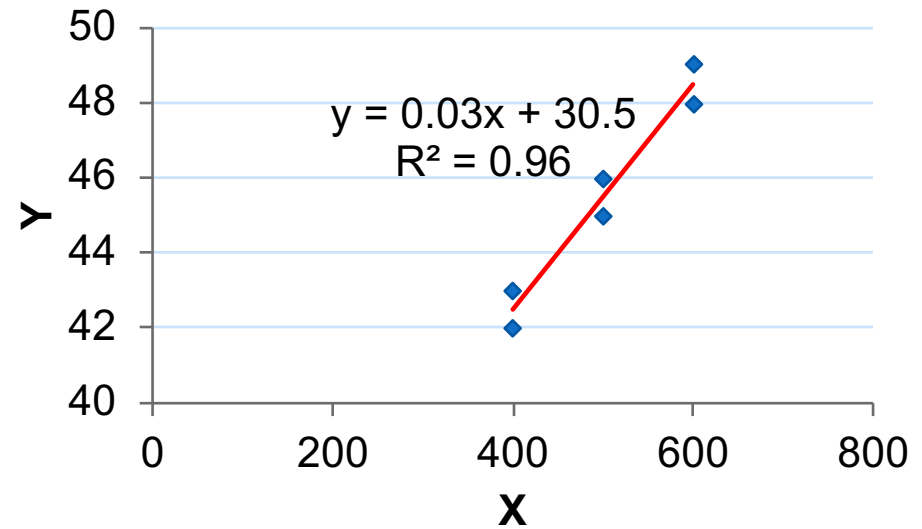
Thời gian chơi điện tử	1	2	3	4	5	4	6	3	1
Mức lương ra trường	12	10	8	6	5	6	4	7	11

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

# Phân tích hồi quy tuyến tính

**Ví dụ:** Các số liệu về số trang của cuốn sách (X) và giá bán của nó (Y) được cho trong bảng dưới đây

Tên sách	X	Y (nghìn)
A	400	43
B	600	48
C	500	45
D	600	49
E	400	42
F	500	46



Hãy tìm đường thẳng hồi quy của Y theo X căn cứ trên số liệu nói trên.

# Phân tích hồi quy tuyến tính

- Giả sử  $X$  là 1 biến nào đó (ngẫu nhiên hay không ngẫu nhiên);  $Y$  là 1 biến ngẫu nhiên phụ thuộc vào  $X$ 
  - Nếu  $X = x$  thì  $Y$  sẽ có **kì vọng** là  $\beta_1 x + \beta_0$  và **phương sai** là  $\sigma^2$
- Ta nói:  $Y$  có hồi quy tuyến tính theo  $X$
- Đường thẳng  $y = \beta_1 x + \beta_0$  là đường thẳng hồi quy lý thuyết của  $Y$  đối với  $X$
- $\beta_0, \beta_1$  gọi là hệ số hồi quy lý thuyết
- $X$  gọi là biến độc lập;  $Y$  gọi là biến phụ thuộc
- **Bài toán:** Ước lượng  $\beta_0, \beta_1$  trên một mẫu quan sát  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$
- **Bài toán:** Ước lượng  $\sigma^2$  trên một mẫu quan sát  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$

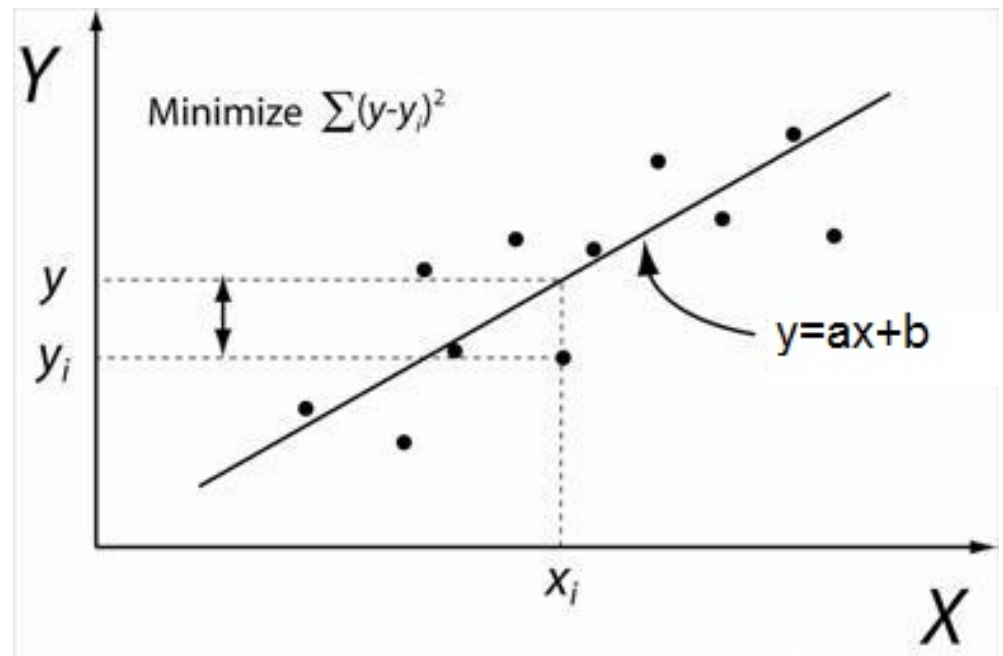
# Bài toán ước lượng $\beta_0, \beta_1$

- Dùng phương pháp bình phương tối thiểu
- a, b làm cực tiểu tổng  $Q(A, B) = \sum_{i=1}^n (y_i - ax_i - b)^2$

$$a = \frac{n \sum xy - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

$$b = \bar{y} - a\bar{x} = \frac{\sum y - a \sum x}{n}$$

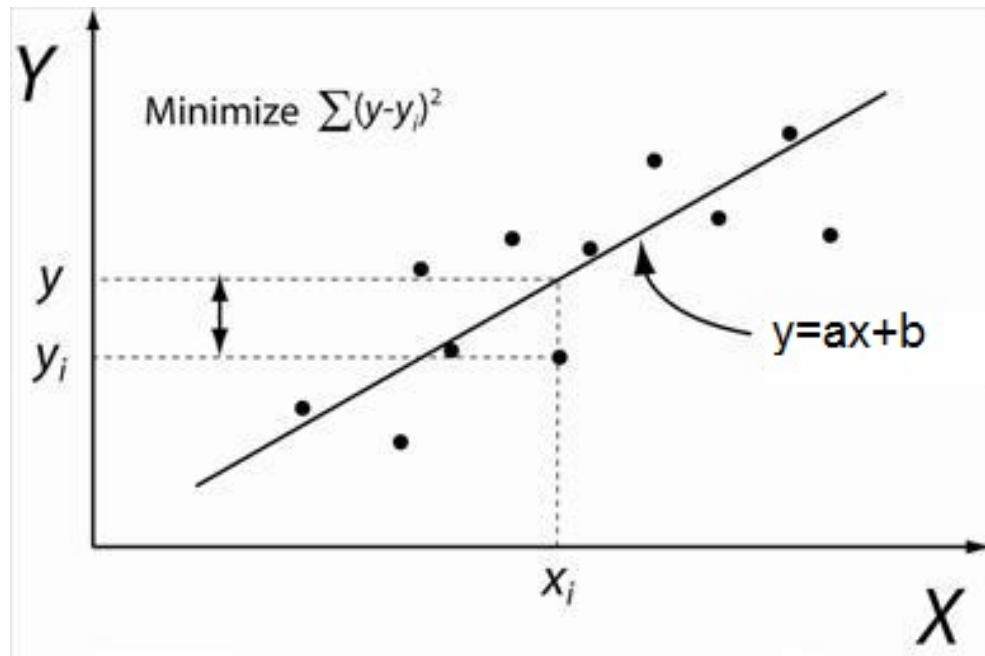
- a, b được gọi là các hệ số hồi quy
- đường thẳng  $y=ax+b$  gọi là đường thẳng hồi quy



# Sai số của đường hồi quy

Kí hiệu  $s_{Y.X}$  sai số tiêu chuẩn của đường hồi quy

$$s_{Y.X}^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - ax_i - b)^2$$



# Bài tập hồi quy 1

Các số liệu về số trang của cuốn sách (X) và giá bán của nó (Y) được cho trong bảng dưới đây

Tên sách	X	Y (nghìn)
A	400	43
B	600	48
C	500	45
D	600	49
E	400	42
F	500	46

$$a = \frac{n \sum xy - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

$$b = \bar{y} - a\bar{x} = \frac{\sum y - a \sum x}{n}$$

$$s_{Y.X}^2 = \frac{\sum y^2 - a \sum xy - b \sum y}{n - 2}$$

- Hãy tìm đường thẳng hồi quy của Y theo X căn cứ trên số liệu nói trên.
- Hãy tính sai số tiêu chuẩn của đường hồi quy.

## Bài tập hồi quy 2

Một công ty quan tâm tới việc phân tích hiệu quả của việc quảng cáo (X) và doanh thu (Y). Trong thời gian 5 tháng công ty thu được kết quả sau.

Tiền quảng cáo (\$M)	1	2	3	4	5
Doanh thu (\$M)	6	15	20	30	39

- Hãy tìm đường thẳng hồi quy của Y theo X căn cứ trên số liệu nói trên.
- Hãy tính sai số tiêu chuẩn của đường hồi quy.



## Bài tập hồi quy 3

Thống kê về số buổi đi học (X) và điểm thi cuối kì môn XSTK (Y) từ 20 sinh viên được cho ở bảng dưới.

X	15	14	10	14	15	7	11	9	14	12
Y	10	9	4	8	9	2	6	8	7	8

X	15	13	5	7	11	14	15	10	12	14
Y	10	8	0	4	6	7	8	5	7	9

- Hãy tìm đường thẳng hồi quy của Y theo X căn cứ trên số liệu nói trên.
- Hãy tính sai số tiêu chuẩn của đường hồi quy.

## Bài tập hồi quy 4

Số năm hút thuốc lá (X) và tuổi thọ (Y) từ 20 người được cho ở bảng dưới.

X	10	15	10	15	20	5	10	15	20	15
Y	70	65	66	60	50	72	67	60	55	60

X	15	10	5	12	22	14	16	18	30	14
Y	70	72	75	70	52	54	52	50	45	60

- Hãy tìm đường thẳng hồi quy của Y theo X căn cứ trên số liệu nói trên.
- Hãy tính sai số tiêu chuẩn của đường hồi quy.

## Bài tập hồi quy 5

Thời gian chơi điện tử của sinh viên một ngày (X) và mức lương ra trường (Y) từ 9 người được cho ở bảng dưới.

Thời gian chơi điện tử	1	2	3	4	5	4	6	3	1
Mức lương ra trường	12	10	8	6	5	6	4	7	11

- Hãy tìm đường thẳng hồi quy của Y theo X căn cứ trên số liệu nói trên.
- Hãy tính sai số tiêu chuẩn của đường hồi quy.