Phân tích tương quan và Hồi quy

Giảng viên: Hoàng Thị Điệp

Khoa CNTT – Đại học Công Nghệ

Nội dung

- Phân tích tương quan tuyến tính
- Kiểm định tính độc lập
- Phân tích hồi quy tuyến tính

Ví dụ 1 (Bài 2, trang 252)

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

X	5	8	10	15	22
Y	6	15	20	30	39

Trong đó X là số tiền chi vào quảng cáo (đơn vị là trăm USD) còn Y là tổng doanh thu (đơn vị là nghìn USD). Hãy xác định hệ số tương quan.

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

Ví dụ 2 (Bài 3, trang 252)

Một trường đại học thu thập các số liệu về số tín chỉ mà một sinh viên theo học và số giờ học ở nhà của anh ta trong một tuần:

X	20	25	30	50	20	23
Y	12	13	12	15	16	16

Ở đó X là số giờ học, Y là số tín chỉ. Tìm hệ số tương quan giữa X và Y. Ở mức ý nghĩa α bằng 5% có sự tương quan tuyến tính giữa hai biến nói trên hay không?

Ví dụ 3 (Bài 4, trang 252)

Một nghiên cứu được tiến hành ở Mỹ để xác định mối quan hệ giữa chiều cao của một người và cỡ giày của họ. Nhà nghiên cứu đã thu được số liệu sau:

X	66	63	67	71	62
Y	9	7	8.5	10	6

Trong đó X là chiều cao (đơn vị là inches) và Y là cỡ giày. Hãy tính hệ số tương quan giữa X và Y.

Ví dụ 4 (Bài 5, trang 253)

Tuổi và huyết áp của 10 bệnh nhân trẻ em (dưới 14 tuổi) chọn ngẫu nhiên được cho trong bảng sau đây:

X	14	1	9	7	9	12	1	3	9	1
Y	100	83	112	152	104	90	92	85	120	130

Trong đó X là tuổi còn Y là huyết áp. Tìm đường hồi quy mẫu của Y đối với X. Tính sai số tiêu chuẩn của đường hồi quy.

Ví dụ 5 (Đề thi cũ)

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. Phân tích tương quan giữa X và Y.

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

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

Hệ số tương quan

- Giả sử X và Y là 2 ĐLNN
- Trong nhiều tình huống thực tế, X và Y không độc lập với nhau
  - X là chiều dài cánh tay, Y là chiều cao của cùng 1 người
  - X là điểm thi tốt nghiệp, Y là điểm thi đại học của cùng 1 người
- **Hệ số tương quan** đo mức độ phụ thuộc tuyến tính giữa X và Y
  - Công thức hệ số tương quan lý thuyết ρ

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

  - $\rho \in [-1;1]$
  - ρ=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
- Xét các bài toán ước lượng và kiểm định ρ căn cứ trên 1 mẫu quan sát

Ước lượng ρ

» Với mẫu quan sát $(x_1, y_1), (x_2, y_2)$,..., (x_n, y_n) của (X,Y)

hệ số tương quan:

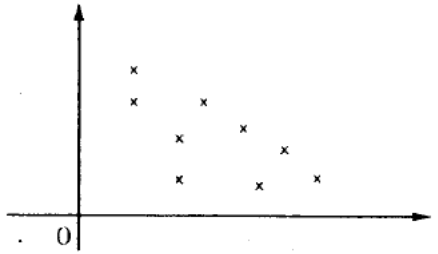
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

» Ví dụ: Một trường đại học thu thập các số liệu về số tín chỉ mà một sinh viên theo học (Y) và số giờ học ở nhà của sinh viên đó trong 1 tuần (X):

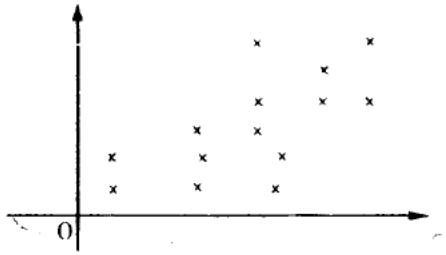
X	20	25	30	50	20	23
Y	12	13	12	15	16	16

Tìm hệ số tương quan giữa X và Y.

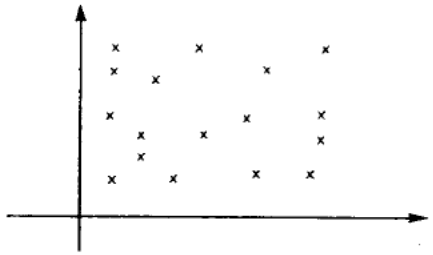
$$r = \frac{n(\sum xy) - (\sum x).(\sum y)}{\sqrt{n\sum x^2 - (\sum x)^2)}\ \sqrt{n\sum y^2 - (\sum y)^2}}$$



$r \approx -1$



$r \approx 1$



$r \approx 0$

Kiểm định xem X với Y có tương quan với nhau không

H_0: $\rho = 0$ (X,Y không tương quan tuyến tính)

H_1: $\rho \neq 0$

- **Định lý**: Nếu X,Y có phân bố chuẩn 2 chiều thì dưới giả thiết H_0, ĐLNN T có phân bố Student với n-2 bậc tự do.

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

- **Ví dụ:** Khảo sát 20 trận đấu bóng đá cho thấy hệ số tương quan giữa số lần sút bóng vào khung thành đối phương và số bàn thắng là 0.21. Với mức ý nghĩa 5%, hãy kiểm định giả thiết "số lần sút bóng và số bàn thắng không tương quan."

Kiểm định $\rho \neq \rho_0$

H_0: $\rho = \rho_0$

H_1: $\rho \neq \rho_0$

Xét $T = \frac{u-m}{\sigma}$ với

» $u = \frac{1}{2}\ln\frac{1+r}{1-r}$

» $m = \frac{1}{2}\ln\frac{1+\rho_0}{1-\rho_0}$

» $\sigma = \frac{1}{\sqrt{n-3}}$

» **Định lý**: Dưới giả thiết H_0, ĐLNN T có phân bố chuẩn tắc.

» Giá trị kiểm định T nói trên cũng cho phép xác định khoảng tin cậy cho hệ số tương quan lý thuyết

- Chi tiết: xem trang 219, 220

Kiểm định tính độc lập (1)

- Phân biệt
  - Dấu hiệu định lượng
    - chiều cao, cân nặng, tuổi
  - Dấu hiệu định tính
    - màu mắt, cảm giác hạnh phúc
- Bài toán: Kiểm định tính độc lập của 2 dấu hiệu định tính A và B
  - Chia dấu hiệu A làm r mức độ
  - Chia dấu hiệu B làm k mức độ
  - Một mẫu ngẫu nhiên gồm n cá thể, mỗi cá thể mang dấu hiệu A ở mức độ A_i nào đó và dấu hiệu B ở mức độ B_j nào đó.
  - n_{ij} là số cá thể mang dấu hiệu A_i và B_j
- Ghi chú: Để mở rộng bài toán này cho dấu hiệu định lượng X, ta chia miền giá trị của X thành m khoảng.

14. Một cuộc thăm dò được tiến hành ở Mỹ bởi viện nghiên cứu xã hội học nổi tiếng Gallup để nghiên cứu mối quan hệ giữa nghề nghiệp của một người với quan niệm của anh ta về tiêu chuẩn đạo đức và tính trung thực. Kết quả của việc khảo sát một mẫu ngẫu nhiên gồm 380 người cho ta số liệu sau đây :

Nghề nghiệp	Quan niệm		
	Cao	Trung bình	Thấp
Bác sĩ	53	35	10
Luật sư	24	43	27
Nhà kinh doanh	18	55	20
Nhà chính trị	14	43	38

Với mức ý nghĩa 10%, hãy xác định xem có mối quan hệ hay không.

$$T = n\left\{\sum \frac{n_{ij}^2}{n_{i0}n_{0j}} - 1\right\}$$

15. Bảng sau đây cho ta số liệu về màu tóc của 422 người:

Màu tóc	Nam	Nữ
Đen	56	32
Hung	37	66
Nâu	84	90
Vàng	19	38

Kiểm định tính độc lập (2)

» Bảng liên hợp các dấu hiệu

	B1	B2	…	Bk	Tổng
A1	n_{11}	n_{12}		n_{1k}	n_{10}
A2	n_{21}	n_{22}		n_{2k}	n_{20}
…					
Ar	n_{r1}	n_{r2}		n_{rk}	n_{r0}
Tổng	n_{01}	n_{02}		n_{0k}	n

» n_{ij} được gọi là tần số quan sát

» Tần số lý thuyết $\widehat{n_{ij}} = \frac{n_{i0}.n_{0j}}{n}$

Kiểm định tính độc lập (3)

$$T = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{\left(n_{ij} - \widehat{n_{ij}}\right)^2}{\widehat{n_{ij}}}$$

- **Định lý:** Nếu n lớn và các tần số lý thuyết >= 5 thì T sẽ có phân bố xấp xỉ phân bố χ^2 với bậc tự do là (k-1)(r-1).
- Công thức cho T trong tính toán thực hành

$$T = n\left\{\sum \frac{n_{ij}^2}{n_{i0}n_{0j}} - 1\right\}$$

Bài tập

$$T = n\left\{\sum \frac{n_{ij}^2}{n_{i0}n_{0j}} - 1\right\}$$

1. Ở các cây ngọc trâm lá có 2 dạng "lá phẳng" hoặc "lá nhăn"; hoa có 2 dạng "hoa bình thường" hoặc "hoa hoàng hậu". Quan sát một mẫu gồm 560 cây ngọc trâm ta thu được kết quả sau:

	Hoa bình thường	Hoa hoàng hậu	Tổng số
Lá phẳng	328	122	450
Lá nhăn	77	33	110
Tổng số	405	155	560

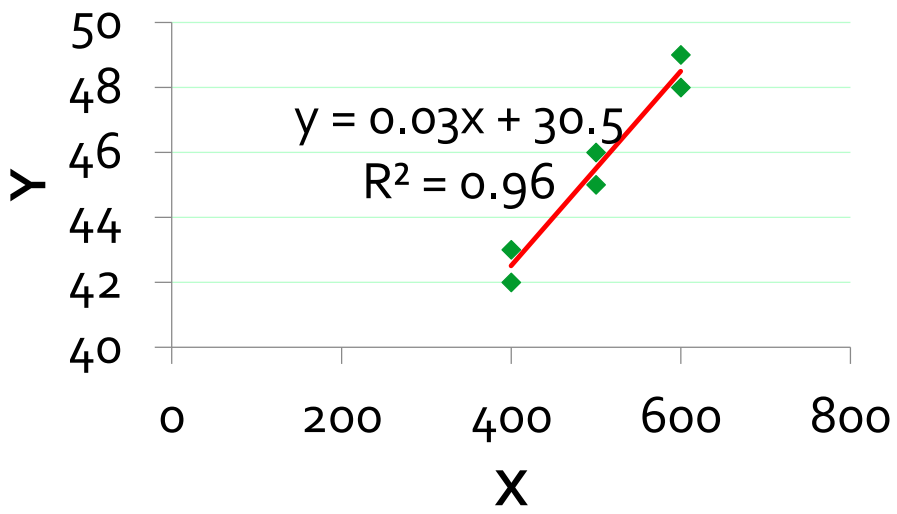
Với mức ý nghĩa 1%, có thể chấp nhận 2 đặc tính về hoa và lá nói trên là độc lập hay không?

2. Bài 16, tr.257

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.

Hồi quy tuyến tính trong thực tế

- Dự đoán việc bán các sản phẩm trong tương lai dựa trên hành vi mua trong quá khứ.
- Dự đoán tăng trưởng kinh tế của một quốc gia hoặc tỉnh thành.
- Dự đoán số bàn thắng mà cầu thủ ghi được trong các trận đấu sắp tới dựa trên thành tích trước đó.
- Ước lượng mức lương công ty sẽ trả cho một người mới dựa trên số năm kinh nghiệm.
- Giúp chủ đầu tư BĐS dự đoán số lượng và mức giá nhà sẽ bán trong những tháng tới.

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 σ^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
- β_0, β_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 β_0, β_1 trên một mẫu quan sát $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- **Bài toán**: Ước lượng σ^2 trên một mẫu quan sát $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

Bài toán ước lượng β_0, β_1

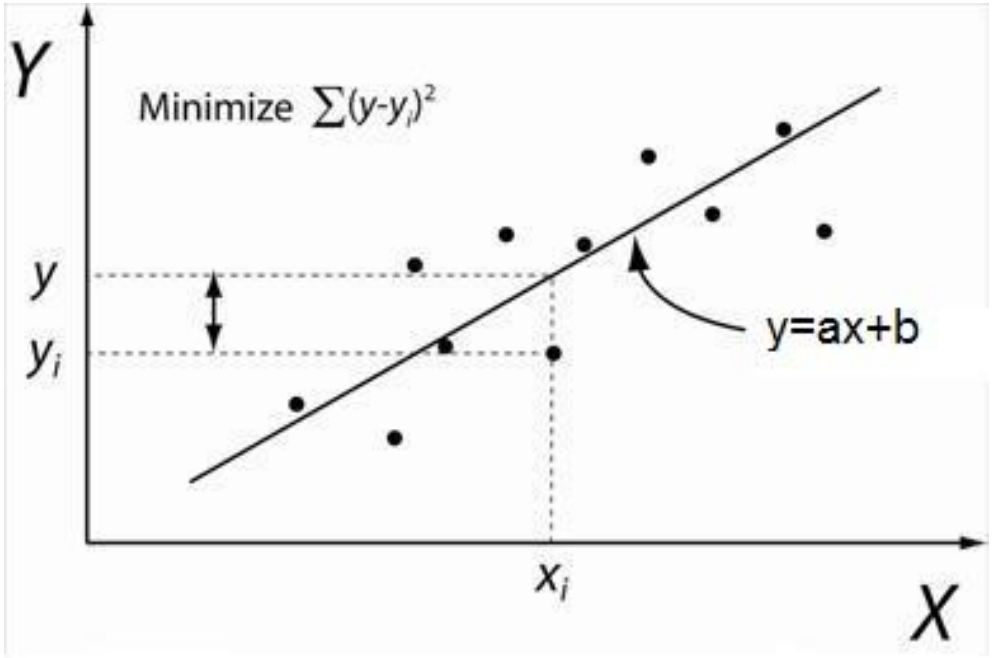
» Dùng phương pháp bình phương tối thiểu

» 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



Ước lượng σ^2

- Kí hiệu $s^2_{Y.X}$
- Công thức 1

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

- Công thức 2

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

- $s_{Y.X}$ được gọi là sai số tiêu chuẩn của đường hồi quy

Bài tập

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

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

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

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

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

b) Hãy tính sai số tiêu chuẩn của đường hồi quy $s_{Y.X}$.

Dự báo

- Dự báo giá trị của Y khi $X = x_0$, kí hiệu $\widehat{y_0}$
- Dự báo kì vọng của Y ứng với $X = x_0$, kí hiệu $\widehat{\mu_{x_0}}$

$$\widehat{y_0} = \widehat{\mu_{x_0}} = ax_0 + b$$

- Khoảng tin cậy cho các giá trị dự báo nói trên: tr.236

Bài tập

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

c) Dự báo giá bán của 1 cuốn sách 450 trang với độ tin cậy 95%.

Kiểm định hệ số hồi quy lý thuyết $\beta_1 \neq 0$

H0: $\beta_1 = 0$

H1: $\beta_1 \neq 0$

» Hệ số hồi quy a có độ lệch tiêu chuẩn s_a

$$s_a = \frac{s_{Y.X}}{s_X\sqrt{n-1}} = \frac{s_{Y.X}}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}}$$

» Kiểm định thống kê $T = \frac{a}{s_a}$ có phân bố Student với n-2 bậc tự do nếu H0 đúng.

Bài tập

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

d) Với mức ý nghĩa 5%, hãy kiểm định giả thiết H0: "Hệ số góc của đường thẳng hồi quy lý thuyết của Y đối với X bằng 0."