Chương 7: Tương quan và hồi quy

Phan Quang Sáng

Học viện Nông nghiệp Việt Nam

Hà Nội, Ngày 7 tháng 5 năm 2018
http://fita.vnua.edu.vn/vi/pqsang

Nội dung chính

1. Tương quan
2. Hồi quy

Tương quan là mối quan hệ giữa 2 (hoặc nhiều) biến ngẫu nhiên, bởi một quan hệ hàm dạng

$$Y = f(X)$$

Tương quan là mối quan hệ giữa 2 (hoặc nhiều) biến ngẫu nhiên, bởi một quan hệ hàm dạng

$$Y = f(X)$$

Làm thế nào để phân tích tương quan giữa hai Bnn X và Y?

$\Rightarrow$ Người ta đưa vào các đặc trưng sau đây

Hiệp phương sai (hay còn gọi là mô men hoặc covariance)

$$\mu(X, Y) = cov(X, Y) = E[(X - E(X))(Y - E(Y))]$$

Tương quan là mối quan hệ giữa 2 (hoặc nhiều) biến ngẫu nhiên, bởi một quan hệ hàm dạng

$$Y = f(X)$$

Làm thế nào để phân tích tương quan giữa hai Bnn X và Y?

$\Rightarrow$ Người ta đưa vào các đặc trưng sau đây

Hiệp phương sai (hay còn gọi là mô men hoặc covariance)

$$\begin{aligned} \mu(X,Y) = cov(X,Y) &= E[(X - E(X))(Y - E(Y))] \\ &= E(XY) - E(X)E(Y) \qquad (1) \end{aligned}$$

Nhận xét: $\mu(X,Y)$ □ 0

Hệ số tương quan: Hệ số này cho biết độ mạnh của mối tương quan tuyến tính giữa hai biến số ngẫu nhiên.

$$\rho = \rho(X, Y) = \frac{cov(X, Y)}{\sqrt{D(X)D(Y)}}$$

Hệ số tương quan: Hệ số này cho biết độ mạnh của mối tương quan tuyến tính giữa hai biến số ngẫu nhiên.

$$\begin{aligned} \rho = \rho(X,Y) &= \frac{cov(X,Y)}{\sqrt{D(X)D(Y)}} \\ &= \frac{E(XY) - E(X)E(Y)}{\sqrt{D(X)D(Y)}} \end{aligned} \tag{2}$$

Một số tính chất:

1. Nếu X và Y độc lập thì $\rho = 0$
2. Luôn có $|\rho| \leq 1$
3. $\rho = \pm 1$ khi và chỉ khi X và Y có quan hệ tuyến tính,

$$Y = A + BX$$

4. $|\rho|$ càng gần 1 thì Qh tuyến tính giữa X và Y càng mạnh

Hệ số tương quan mẫu: Giả sử $(x_i, y_i), i = 1, 2, \ldots, n$ là các cặp mẫu ngẫu nhiên cụ thể của (X, Y)

$$r = r(X, Y) = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\sqrt{(\overline{x^2} - \overline{x}^2)(\overline{y^2} - \overline{y}^2)}} \quad (3)$$

Tính chất: tương tự như hệ số tương quan

Hồi quy là gì?

Tìm mối quan hệ hàm giữa Y và X; hoặc tìm cách thay thế Y bởi một hàm $f(X)$, tốt nhất theo một nghĩa nào đó.

Hàm hồi quy trung bình bình phương (lý thuyết)
Hàm $f(X)$ thỏa mãn

$$E[Y - f(X)]^2 \leq E[Y - g(X)]^2, \text{ với mọi hàm } g$$

Đặc biệt: khi xét $f = A^* + B^*X$, $g = A + BX$ là các hàm tuyến tính thì f được gọi là **hàm hồi quy tuyến tính lý thuyết** của Y theo X.

Tìm A, B làm cực tiểu hàm $\varphi(A, B) = E[Y - (A + BX)]^2$
Kết quả:

$$B^* = \frac{E(XY) - E(X)E(Y)}{D(X)}$$

Hàm hồi quy trung bình bình phương (lý thuyết)
Hàm $f(X)$ thỏa mãn

$$E[Y - f(X)]^2 \leq E[Y - g(X)]^2, \text{ với mọi hàm } g$$

Đặc biệt: khi xét $f = A^* + B^*X$, $g = A + BX$ là các hàm tuyến tính thì f được gọi là **hàm hồi quy tuyến tính lý thuyết** của Y theo X.

Tìm A, B làm cực tiểu hàm $\varphi(A, B) = E[Y - (A + BX)]^2$
Kết quả:

$$B^* = \frac{E(XY) - E(X)E(Y)}{D(X)} = \frac{E(XY) - E(X)E(Y)}{E(X^2) - E(X)^2}$$

$$A^* = E(Y) - B^*E(X)$$

Hàm hồi quy tuyến tính mẫu:

Giả sử $(X_i, Y_i), i = 1, 2, \ldots, n$ là các cặp mẫu ngẫu nhiên của (X, Y).
Khi đó hàm hồi quy tuyến tính mẫu là

$$Y = A^* + B^* X, \text{ với,}$$

$$B^* = \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\overline{X^2} - \overline{X}^2}$$

$$A^* = \overline{Y} - B^* \overline{X}$$

Câu IV (2.5 điểm): Thống kê năng suất một loại cây trồng Y (tạ/sào) và số tiền đầu tư cho sản xuất X(triệu đồng/ha) (phân bón, thuốc trừ sâu bệnh, cải tạo đất…) tại huyện A trong 10 năm, người ta thu được kết quả:

Tiền đầu tư X	20	21	21	23	24	25	26	28	30	30
Năng suất Y	2	2,4	2,8	2,6	2,8	3,2	3,0	3,6	3,7	4

1) Tìm hệ số tương quan mẫu giữa X và Y.
2) Tìm hàm hồi qui tuyến tính mẫu của Y theo X. Dự đoán năng suất cây sẽ tăng thêm bao nhiêu nếu tăng mức đầu tư cho sản xuất thêm 2 triệu/ha.