Chương 7

HỒI QUI VÀ TƯƠNG QUAN TUYẾN TÍNH

I. Tương quan tuyến tính :

Xét hai biến ngẫu nhiên Y và X có quan hệ phụ thuộc tuyến tính. Giả sử biến X – biến độc lập, biến Y – biến phụ thuộc vào X và từ tổng thể $\boldsymbol{M}$ ta lấy mẫu quan sát X và Y.

Có hai cách chọn mẫu:

Cách thứ nhất: Cố định X, chẳng hạn . Ứng với ta có một tổng thể con $\boldsymbol{M}_i$ của $\boldsymbol{M}$, $i = 1, ..., n$. Từ $\boldsymbol{M}_i$ ta lấy ngẫu nhiên các thể và xác định . Ở đây Y là biến ngẫu nhiên và mẫu lý thuyết có dạng, còn mẫu thực nghiệm được viết.

Cách thứ hai: Chọn ngẫu nhiên n cá thể từ M và trên mỗi các thể quan sát X và Y. Ở đây X và Y đều là biến ngẫu nhiên và ta có thể dùng hệ số tương quan giữa X và Y để đưa ra các kết luận thống kê, trong khi đó cách thứ nhất không thể làm như vậy được. Mẫu lý thuyết có dạng $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ và mẫu thực nghiệm: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

Không phụ thuộc vào cách chọn mẫu, có hai bước sơ khởi xác định mức độ quan hệ tuyến tính giữa X và Y.

Bước thứ nhất: Vẽ các điểm trên hệ tọa độ xOy. Dựa vào đồ thị ta đưa ra phỏng đoán về sự phụ thuộc tuyến tính giữ X và Y.

Bước thứ hai: Tính hệ số tương quan mẫu

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

trong đó $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$; $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Nếu lớn thì ta phỏng đoán giữa X và Y có quan hệ tuyến tính chặt chẽ.

Nếu $|r|$ lớn thì ta phỏng đoán giữa X và Y có quan hệ tuyến tính chặt chẽ.

II. Phương trình hồi qui tuyến tính :

Ta xét trường hợp X không ngẫu nhiên, với X ngẫu nhiên kết quả cũng tương tự. Xét mẫu lý thuyết $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$.

Giả sử, $$Y_i = ax_i + b + e_i, \quad i = 1, \ldots, n$$

1) Y và X có quan hệ tuyến tính và được biểu diễn bởi phương trình được gọi là mô hình hồi qui tuyến tính đơn của Y theo X, trong đó a và b là các hệ số chưa biết.

2) $e_1, \ldots, e_n$ là các sai số ngẫu nhiên độc lập.

Ta cần dựa vào mẫu để ước lượng a và b bằng phương pháp bình phương nhỏ nhất. Tức là tìm ước lượng a và b của a và b sao cho tổng bình phương sai lệch

$$f(a,b) = \sum_{i=1}^{n} (Y_i - ax_i - b)^2$$

đạt cực tiểu: $\sum_{i=1}^{n} (Y_i - ax_i - \hat{b})^2 = \min_{a,b} f(a,b)$.

Giải hệ phương trình

$$\frac{\partial f(a,b)}{\partial a} = 0$$

$$\frac{\partial f(a,b)}{\partial b} = 0$$

ta tìm được

$$\hat{b} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$a = \bar{Y} - \hat{b}\,\bar{x}$$

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

$$\bar{x} = \sum_{i=1}^{n} x_i$$

Như vậy, ta có phương trình đường thẳng hồi qui thực nghiệm: $y = ax + \hat{b}$. Nghĩa là ước lượng của Y tại giá trị $X = x_i$ là $y_i = ax_i + \hat{b}$.

Nhận xét:

- Có hai cách dự báo giá trị y .

Cách thứ nhất: Dự báo giá trị Y cho một cá thể, mà trên đó có X nhận giá trị x. Trong trường hợp này y là ước lượng tốt nhất của duy nhất giá trị Y ứng với $X = x$.

Cách thứ hai: Dự báo giá trị trung bình của Y đối với tổng thể con ứng với $X = x$. Và ở đây y cũng là ước lượng tốt nhất của giá trị trung bình của Y khi $X = x$.

Sự khác biệt giữa hai cách trên sẽ quan trọng khi xây dựng khoảng tin cậy.

- Ta có thể dự báo X theo Y bằng phương trình:

$$x = (y - \hat{b}) / a \quad .$$

III. Khoảng tin cậy:

Ngoài 2 giả định 1) và 2) trong phần II ở trên, trong phần này giả sử rằng thỏa điều kiện thứ ba sau đây:

3) Các biến ngẫu nhiên $e_1, \ldots, e_n$ có phân phối chuẩn $N(0, \sigma^2)$.

Như vậy với mỗi giá trị $X = x_i$ ta có biến ngẫu nhiên Y_i có luật phân phối chuẩn $N(ax_i + b, \sigma^2)$. Với giả định trên ta xét các khoảng tin cậy sau:

1. Khoảng tin cậy cho $E(Y/x) = ax + b$, kỳ vọng của Y tại $X = x$, có dạng $(y - w, y + w)$, trong đó

$$w = t_{\frac{1+\gamma}{2}}^{n-2} s \sqrt{\frac{1}{n} + \frac{(x - \overline{x})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}}$$

$$s^2 = \frac{\sum_{i=1}^{n} (y_i - y_i)^2}{n - 2}$$

$t_{\frac{1+\gamma}{2}}^{n-2}$ là phân vị $\frac{1+\gamma}{2}$ mức n-2 bậc tự do.

2. Khoảng tin cậy cho Y tại $X = x$, có dạng $(y - w, y + w)$, trong đó

$$w = t_{\frac{1+\gamma}{2}}^{n-2} s \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

<u>Nhận xét:</u> s^2 được dùng để ước lượng σ^2.